Gunilla Kreiss
Per Lötstedt
Axel Målqvist
Maya Neytcheva *Editors*

Numerical Mathematics and Advanced Applications

ENUMATH 2009

Numerical Mathematics and Advanced Applications 2009

Gunilla Kreiss • Per Lötstedt • Axel Målqvist
Maya Neytcheva
Editors

Numerical Mathematics and Advanced Applications 2009

Proceedings of ENUMATH 2009,
the 8th European Conference on Numerical Mathematics and Advanced Applications,
Uppsala, July 2009

Springer

Editors
Gunilla Kreiss
Per Lötstedt
Axel Målqvist
Maya Neytcheva
Uppsala University
Department of Information Technology
751 05 Uppsala
Sweden
gunilla.kreiss@it.uu.se
perl@it.uu.se
axel.malqvist@it.uu.se
maya.neytcheva@it.uu.se

ISBN 978-3-642-11794-7 e-ISBN 978-3-642-11795-4
DOI 10.1007/978-3-642-11795-4
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2010937319

Mathematics Subject Classification (2010): 65-06

Cover design: WMX Design, GmbH

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

The European Conference on Numerical Mathematics and Advanced Applications (ENUMATH) was held from June 29–July 3, 2010, in Uppsala, Sweden. This was the eighth conference in a series of biannual meetings starting in Paris (1995). Subsequent conferences were organized in Heidelberg (1997), Jyväskylä (1999), Ischia (2001), Prague (2003), Santiago de Compostela (2005), and Graz (2007). ENUMATH 2009 attracted over 330 attendees to the scientific programme, with ten invited speakers, one public lecture, 32 minisymposia, and more than 280 presentations. This volume contains a selection of papers by the invited speakers and from the minisymposia and the contributed sessions.

The purpose of the conference was to create a forum for discussion and dissemination of recent results in numerical mathematics and new applications of computational methods. Many subjects were covered in the talks and a few of the topics represented in these proceedings were discontinuous Galerkin methods, finite element methods in different applications, methods for fluid flow, electromagnetism, financial engineering, structural mechanics, optimal control, and biomechanics. The minisymposia listed below with their organizers also give an impression of how broad the scope of the conference was:

- *Adaptivity for non-linear and non-smooth problems*, part I & II, Ralf Kornhuber, Andreas Veeser
- *Advanced techniques in radial basis function approximation for PDEs*, part I & II, Natasha Flyer, Elisabeth Larsson
- *Advances in numerical methods for non-Newtonian flows*, part I & II, Erik Burman, Maxim Olshanskii, Stefan Turek
- *Anisotropic adaptive meshes: error analysis and applications*, part I & II, Thierry Coupez, Simona Perotto
- *Asymptotic linear algebra, numerical methods, and applications*, part I & II, Marco Donatelli, Stefano Serra-Capizzano
- *Biomechanics*, part I & II, Gerhard A. Holzapfel, Axel Klawonn
- *Embedded boundary methods for time-dependent problems*, Daniel Appelö
- *Finite element software development*, Anders Logg
- *Finite elements for convection-diffusion problems*, part I, II & III, Miloslav Feistauer, Petr Knobloch

- *Finite element methods for flow problems*, Johan Hoffman
- *Geometric aspects of the finite element modeling*, part I & II, Sergey Korotov, Tomas Vejchodsky
- *High frequency wave propagation*, Olof Runborg
- *High order methods in CFD*, Bernhard Müller
- *HPC-driven numerical methods and applications*, part I & II, Svetozar Margenov, Maya Neytcheva
- *Multiscale methods for differential equations*, part I & II, Mats Larson, Axel Målqvist
- *Numerical methods for multi-dimensional Lagrangian schemes*, Pierre-Henri Maire, Raphaël Loubere
- *Numerical methods for option pricing*, Cornelis W. Oosterlee, Jari Toivanen
- *Numerical methods for stochastic partial differential equations*, part I & II, Fabio Nobile, Raul Tempone
- *Tensor numerical methods*, Eugene Tyrtyshnikov, Boris Khoromskij
- *Theory and applications of non-conforming finite element methods*, Emmanuil Georgoulis, Max Jensen

The conference was organized by the Division of Scientific Computing of the Department of Information Technology at Uppsala University in collaboration with Akademikonferens in Uppsala. Uppsala University is not as old as the universities in Paris, Heidelberg, and Prague, but it is the oldest university in the Nordic countries. It was founded in 1477 and the first professor in mathematics was appointed in 1593. The first professor in numerical analysis, Heinz-Otto Kreiss, started his work in 1965.

The success of the conference was in a large part due to the invited speakers Martin Berggren, Daniele Boffi, Carsten Carstensen, Vit Dolejsi, Charlie Elliott, Claude Le Bris, Christian Lubich, Marco Picasso, Rob Stevenson, and Anna-Karin Tornberg, as well as to Björn Engquist, who delivered the public lecture. The members of the program committee were Franco Brezzi, Miloslav Feistauer, Roland Glowinski, Rolf Jeltsch, Yuri Kuznetsov, Jacques Périaux, Rolf Rannacher, and Endre Süli. They selected the invited speakers and helped by sharing their knowledge of how are organized these conferences.

The scientific committee consisted of Christine Bernardi, Alfredo Bermudez de Castro, Albert Cohen, Claudio Canuto, Michael Griebel, Peter Hansbo, Jaroslav Haslinger, Thomas Huckle, Karl Kunisch, Ulrich Langer, Stig Larsson, Olivier Pironneau, Sergey Repin, Miro Rozloznik, J. J. Sanz-Serna, Stefan Sauter, Stefano Serra Capizzano, Valeria Simoncini, Olaf Steinbach, Rolf Stenberg, Anders Szepessy, Stefan Turek, Kees Vuik, Ragnar Winther, and Barbara Wohlmuth. Members of the committee, Martin Berggren and Bernhard Müller have served as referees for this volume.

The local committee was assisted by PhD students at our Division: Qaisar Abbas, Kenneth Duru, Magnus Gustafsson, Andreas Hellander, Stefan Hellander, Katharina Kormann, Martin Kronbichler, Erik Lehto, Anna Nissen, Elena Sundkvist, Martin Tillenius, Salman Toor, and He Xin. The change between speakers in the sessions would not have been so smooth without their presence in the lecture rooms. Special thanks to Kenneth Duru for helping with the preparations of the proceedings.

The conference received financial support from Centre for Interdisciplinary Mathematics at Uppsala University, City of Uppsala, Comsol, Swedish Foundation for Strategic Research, Swedish Research Council, Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX), Wenner-Gren Foundations, and John Wiley & Sons. Their generous contributions helped to lower the fees for the participants.

Last but not least, many thanks to Karin Hornay and Maria Bäckström from Akademikonferens for sharing their invaluable experience in organizing conferences.

Uppsala
March 2010

Gunilla Kreiss
Per Lötstedt
Axel Målqvist
Maya Neytcheva

Contents

Part I
Invited Papers

Discrete Differential Forms, Approximation of Eigenvalue Problems, and Application to the p Version of Edge Finite Elements

Daniele Boffi

Abstract We are interested in the approximation of the eigenvalues of Hodge–Laplace operator in the framework of de Rham complex by using exterior calculus and suitable equivalent formulations in mixed form. We discuss the role of discrete compactness property and show how it is related to the classic conditions for the convergence of eigenvalues in mixed form. In this context, we review a recent result concerning the discrete compactness for the p version of discrete differential forms. One of the applications of the presented theory is the convergence analysis of the p version of edge finite elements for the approximation of Maxwell's eigenvalues.

1 Introduction

The use of homological techniques for the analysis of finite element approximations of partial differential equations has become a very popular and effective tool (see [3, 4]). In the framework of de Rham complex it is natural to consider the eigenvalue problem associated with Hodge–Laplace operator. There are several eigenvalue problems of interest for the applications, which can be related to the Hodge–Laplace eigenvalue problem: for instance the standard Laplace eigenvalue problem fits within this framework (0-forms), as well as the Maxwell eigenvalue problem (1-forms in two or three space dimensions), or the eigenvalue problem associated with **grad** div operator (2-forms in three dimensions).

The main object of this paper is to extend the results of [11] to differential forms.

First of all, we consider two mixed variational formulations which give the same solutions as the standard formulation originally designed for the Hodge–Laplace eigenvalue problem. The theory developed in [13] can be used for the analysis of the mixed formulations in order to show the convergence of the eigenpairs; classic results ([6, 32]) give the order of convergence for eigenvalues/eigenfunctions.

D. Boffi
Dipartimento di Matematica "F. Casorati", Università di Pavia, via Ferrata 1 Pavia, Italy
e-mail: daniele.boffi@unipv.it

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_1,

Then, we recall the discrete compactness property that can be naturally written in the context of differential forms. The notion of discrete compactness has been used since long time in the literature: we recall, in particular, the works by Stummel [37], Vaĭnikko [38], Anselone [1], and the more recent book by Chatelin [19]. In the approximation of Maxwell's eigenvalues, it has been used firstly by Kikuchi [30] and then reinterpreted and rephrased by several authors [7, 8, 10, 11, 18, 21, 31, 33]. We refer the interested reader also to [28, 35] and to the references therein for a review on this topic.

Following [11], we show that the discrete compactness property and standard approximation properties are equivalent to the natural convergence conditions for the two equivalent mixed formulation we have introduced.

One of the consequences of the presented theory is that we can show that a discretization that satisfies the discrete compactness property provides convergent eigenpairs and that such convergence can be analyzed by means of the standard Babuška–Osborn theory. A similar result has been obtained in [9] as a consequence of the theory developed by [18] which makes use of the results of [26].

In [5] it has been introduced a comprehensive theory for the convergence of the eigenmodes of the Hodge–Laplace operator. The theory has been used (together with a suitably defined projection operator) for the analysis of the convergence of the h version of finite elements applied to k-forms in any space dimensions (when suitable discrete differential forms are used). The abstract hypotheses of [5] imply, in particular, our discrete compactness property.

When discussing the p version of finite elements for the approximation of the eigenmodes of the Hodge–Laplace operator, it is still an open problem to see whether the assumptions of [5] are satisfied for discrete differential forms. On the other hand, in [9] it has been shown that the discrete compactness in p is valid as a consequence of a recent result on the Poincaré operator (see [20]). In particular, this implies that two- and three-dimensional edge elements provide a good convergence in p for the eigenvalues/eigenfunctions of the Maxwell cavity problem.

The results of the present paper and, in particular, the relationships between the eingevalue problem associated with the Hodge–Laplace operator for differential forms and suitable mixed formulations are discussed in more detail in [12].

2 Short Introduction to de Rham Complex and Differential Forms

Given a domain $\Omega \subset \mathbb{R}^n$ and k with $0 \leq k \leq n$, we denote by $C^\infty(\Omega, \Lambda^k)$ the space of smooth differential forms of order k on Ω. For the sake of simplicity, we assume that Ω is simply connected, but the results of this paper might be generalized to non-trivial cohomologies with natural modifications.

We suppose that we are given an exterior derivative

$$d_k : C^\infty(\Omega, \Lambda^k) \to C^\infty(\Omega, \Lambda^{k+1})$$

for any k. The space $L^2(\Omega, \Lambda^k)$ denotes the space of differential k-forms on Ω with square integrable coefficients in their canonical basis representation; its inner product is given by

$$(\mathbf{u}, \mathbf{v}) = \int_\Omega \mathbf{u} \wedge \star \mathbf{v},$$

where $\star$ denotes the Hodge star operator mapping k-forms to $(n-k)$-forms.

We shall make use of the Hilbert spaces

$$H(d_k, \Omega) = \{\mathbf{v} \in L^2(\Omega, \Lambda^k) : d_k \mathbf{v} \in L^2(\Omega, \Lambda^{k+1})\}$$

and

$$H_0(d_k, \Omega) = \{\mathbf{v} \in H(d_k, \Omega) : \mathbf{tr}_{\partial\Omega} \mathbf{v} = 0\}.$$

We refer the interested reader to [4] for a canonical definition of the trace operator $\mathbf{tr}_{\partial\Omega}$. In particular, we are interested in the following complex

$$\mathbb{R} \longrightarrow H_0(d_0, \Omega) \xrightarrow{d_0} H_0(d_1, \Omega) \xrightarrow{d_1} \cdots \xrightarrow{d_{n-1}} H_0(d_n, \Omega) \longrightarrow 0.$$

We recall in particular that $d_{k+1} \circ d_k = 0$ and that the range of d_k coincides with the kernel of $d_k + 1$.

Given spaces of discrete differential forms $V_p^k \subset H_0(d_k, \Omega)$, a typical setting involves appropriate projection operators $\pi_p^k : H_0(d_k, \Omega) \to V_p^k$ such that the following full de Rham complex commutes

$$\begin{array}{ccccccccccc}
\mathbb{R} & \longrightarrow & H_0(d_0, \Omega) & \xrightarrow{d_0} & H_0(d_1, \Omega) & \xrightarrow{d_1} & \cdots & \xrightarrow{d_{n-1}} & H_0(d_n, \Omega) & \longrightarrow & 0 \\
 & & \downarrow{\scriptstyle \pi_p^0} & & \downarrow{\scriptstyle \pi_p^1} & & & & \downarrow{\scriptstyle \pi_p^n} & & \\
\mathbb{R} & \longrightarrow & V_p^0 & \xrightarrow{d_0} & V_p^1 & \xrightarrow{d_1} & \cdots & \xrightarrow{d_{n-1}} & V_p^n & \longrightarrow & 0.
\end{array} \tag{1}$$

Remark 1. We use the index p for discrete spaces, so that it is explicit that we are interested in the p version of finite elements; nevertheless, the abstract theory we are going to present is valid for general Galerkin discretizations where V_p^k are finite dimensional subspaces of $H_0(d_k, \Omega)$.

Remark 2. In general, we are not going to assume that the full diagram (1) commutes. When we are interested in differential forms of degree k, it will be enough to consider a small portion of (1) in the vicinity of k-forms.

The coderivative operator $\delta_k = \star d_{n-k} \star$ maps $C^\infty(\Omega, \Lambda^k)$ to $C^\infty(\Omega, \Lambda^{k-1})$ and leads to the definition of the Hilbert space

$$H(\delta_k, \Omega) = \{\mathbf{v} \in L^2(\Omega, \Lambda^k) : \delta_k \mathbf{v} \in L^2(\Omega, \Lambda^{k-1})\}.$$

The spaces of differential forms when $n = 2, 3$ have been studied intensively. Table 1 recalls the representation of the involved quantities in terms of vector proxies.

Table 1 Identification between differential forms and vector proxies in $\mathbb{R}^2$ and $\mathbb{R}^3$

Differential form		Proxy representation	
		$d=2$	$d=3$
$k=0$	d_0	**grad**	**grad**
	$\mathbf{tr}_{\partial\Omega}\phi$	$\phi_{\vert\partial\Omega}$	$\phi_{\vert\partial\Omega}$
	$H_0(d_0,\Omega)$	$H_0^1(\Omega)$	$H_0^1(\Omega)$
	δ_1	$-\operatorname{div}$	$-\operatorname{div}$
$k=1$	d_1	curl	curl
	$\mathbf{tr}_{\partial\Omega}\mathbf{u}$	$(\mathbf{u}\cdot\mathbf{t})_{\vert\partial\Omega}$	$\mathbf{n}\times(\mathbf{u}\times\mathbf{n})_{\vert\partial\Omega}$
	$H_0(d_1,\Omega)$	$\mathbf{H}_0(\operatorname{curl})$	$\mathbf{H}_0(\operatorname{curl})$
	δ_2	$\overrightarrow{\operatorname{curl}}$	curl
$k=2$	d_2	0	div
	$\mathbf{tr}_{\partial\Omega}\mathbf{q}$	0	$(\mathbf{q}\cdot\mathbf{n})_{\vert\partial\Omega}$
	$H_0(d_2,\Omega)$	$L_0^2(\Omega)$	$\mathbf{H}_0(\operatorname{div})$
	δ_3		$-\mathbf{grad}$

3 The Hodge–Laplace Eigenvalue Problem

Given k with $0 \le k \le n$, we are interested in the following symmetric eigenvalue problem: find $\lambda \in \mathbb{R}$ and $\mathbf{u} \in H_0(d_k,\Omega)$ with $\mathbf{u} \neq 0$ such that

$$(d_k\mathbf{u}, d_k\mathbf{v}) = \lambda(\mathbf{u},\mathbf{v}) \quad \forall \mathbf{v} \in H_0(d_k,\Omega). \tag{2}$$

The interest for the eigenvalue problem (2) arose in [9]: the case $k=1$ (both for $n=2$ and $n=3$) corresponds to the Maxwell eigenvalue problem, since d_1 can be identified to the curl operator (see Table 1).

Problem (2) is strictly related to the so called Hodge–Laplace elliptic eigenvalue problem (see [4, 5]): find $\omega \in \mathbb{R}$ and $\mathbf{u} \in H_0(d_k,\Omega) \cap H(\delta_k,\Omega)$ with $\mathbf{u} \neq 0$ such that

$$(d_k\mathbf{u}, d_k\mathbf{v}) + (\delta_k\mathbf{u}, \delta_k\mathbf{v}) = \omega(\mathbf{u},\mathbf{v}) \quad \forall \mathbf{v} \in H_0(d_k,\Omega) \cap H(\delta_k,\Omega). \tag{3}$$

It is well-known that problem (3) is associated with a compact solution operator; this is consequence of the compact embedding of $H_0(d_k,\Omega) \cap H(\delta_k,\Omega)$ into $L^2(\Omega,\Lambda^k)$ (see [36]). Moreover, the eigensolutions of (3) split into two separate families: the first one consists of eigenvalues corresponding to eigenfunctions $\mathbf{u}$ with $d_k\mathbf{u}=0$ and the second one of eigenvalues corresponding to eigenfunctions $\mathbf{u}$ with $\delta_k\mathbf{u}=0$. The second family corresponds to all the solutions to our original eigenvalue problem (2) with positive frequencies. In addition, the zero frequency solves problem (2) with the infinite dimensional eigenspace $d_{k-1}(H_0(d_{k-1},\Omega))$.

Given a finite dimensional discretization V_p^k of $H_0(d_k,\Omega)$, the discrete version of (2) is: find $\lambda_p \in \mathbb{R}$ and $\mathbf{u}_p \in V_p^k$ with $\mathbf{u}_h \neq 0$ such that

$$(d_k\mathbf{u}_p, d_k\mathbf{v}) = \lambda_p(\mathbf{u}_h,\mathbf{v}) \quad \forall \mathbf{v} \in V_p^k. \tag{4}$$

One of the main issues for the convergence of the solutions of (4) towards those of (2) is the infinite dimensional kernel of (2). In general, we would like that the positive frequencies of (4) provide a good approximation of the positive frequencies of (2). From the above discussion, it follows that adding the condition $\lambda > 0$ to problem (2) is equivalent to adding the condition $\delta_k \mathbf{u} = 0$ to the solution of problem (2). This last property can also be proved by taking $\mathbf{v} = d_{k-1}\mathbf{t}$ in (2) with an arbitrary $\mathbf{t} \in H_0(d_{k-1}, \Omega)$. The variational equation with $\lambda \neq 0$ implies then $(\mathbf{u}, d_{k-1}\mathbf{t}) = 0$, that is $\delta_k \mathbf{u} = 0$.

In order to isolate the positive frequencies of problems (2) and (4) it is very convenient to consider equivalent mixed formulations.

3.1 First Mixed Formulation

A first mixed formulation of problem (2) can be obtained as a generalization of the so-called Kikuchi formulation for Maxwell's eigenvalue problem (see [29]). It uses $(k-1)$- and k-forms as follows: find $\lambda \in \mathbb{R}$ and $\mathbf{u} \in H_0(d_k, \Omega)$ with $\mathbf{u} \neq 0$ such that for $\mathbf{s} \in H_0(d_{k-1}, \Omega)$ it holds

$$\begin{cases} (d_k\mathbf{u}, d_k\mathbf{v}) + (d_{k-1}\mathbf{s}, \mathbf{v}) = \lambda(\mathbf{u}, \mathbf{v}) & \forall \mathbf{v} \in H_0(d_k, \Omega) \\ (d_{k-1}\mathbf{t}, \mathbf{u}) = 0 & \forall \mathbf{t} \in H_0(d_{k-1}, \Omega). \end{cases} \tag{5}$$

It can be easily shown that all eigensolutions of problem (5) have $\lambda > 0$ and solve the original problem (2), that $d_{k-1}\mathbf{s}$ is always equal to zero (take $\mathbf{v} = d_{k-1}\mathbf{s}$ in the first equation of (5)), and that all eigensolutions of (2) with positive eigenvalue solve (5) as well. When $k = 1$ (which is the case for Maxwell's eigenvalue problem), we additionally have that $\mathbf{s} = 0$ from $d_0\mathbf{s} = 0$ and the boundary conditions.

A discretization of (5) involves the discrete spaces $V_p^{k-1} \subset H_0(d_{k-1}, \Omega)$ and $V_p^k \subset H_0(d_k, \Omega)$ as follows: find $\lambda_p \in \mathbb{R}$ and $\mathbf{u}_p \in V_p^k$ with $\mathbf{u}_p \neq 0$ such that for $\mathbf{s}_p \in V_p^{k-1}$ it holds

$$\begin{cases} (d_k\mathbf{u}_p, d_k\mathbf{v}) + (d_{k-1}\mathbf{s}_p, \mathbf{v}) = \lambda_p(\mathbf{u}_p, \mathbf{v}) & \forall \mathbf{v} \in V_p^k \\ (d_{k-1}\mathbf{t}, \mathbf{u}_p) = 0 & \forall \mathbf{t} \in V_p^{k-1}. \end{cases} \tag{6}$$

We assume the fundamental inclusion

$$d_{k-1}(V_p^{k-1}) \subset V_p^k \tag{7}$$

which is a compatibility condition valid whenever diagram (1) is satisfied. Under hypothesis (7) the discrete problem (6) is equivalent to (4) in the sense that all solutions corresponding to positive frequencies are the same. Hence the convergence analysis of the solutions of (6) towards those of (5) can be used in order to pursue

our goal of studying the convergence of the positive solutions of (4) to the positive solutions of (2).

It follows from the theory of [13, Sect. 3] that the conditions we are going to present, ensure (and, in a sense, are necessary for) the convergence of the eigensolutions of (6) towards those of (5).

We need the discrete kernel of the δ_k operator (or, better, the kernel of the discrete δ_k operator), defined as follows:

$$\mathbb{K}_p^1 = \{\mathbf{v} \in V_p^k : (\mathbf{v}, d_k\mathbf{t}) = 0,\ \forall \mathbf{t} \in V_p^{k-1}\}.$$

Moreover, we introduce the solution spaces of the source problem associated with (5): V_0^k and V_0^{k-1} are the subspaces of $H_0(d_k, \Omega)$ and $H_0(d_{k-1}, \Omega)$, respectively, containing all the first and second components $\mathbf{u} \in H_0(d_k, \Omega)$ and $\mathbf{s} \in H_0(d_{k-1}, \Omega)$, respectively, of the solution of the source problem

$$\begin{cases} (d_k\mathbf{u}, d_k\mathbf{v}) + (d_{k-1}\mathbf{s}, \mathbf{v}) = (\mathbf{f}, \mathbf{v}) & \forall \mathbf{v} \in H_0(d_k, \Omega) \\ (d_{k-1}\mathbf{t}, \mathbf{u}) = 0 & \forall \mathbf{t} \in H_0(d_{k-1}, \Omega), \end{cases}$$

when $\mathbf{f}$ varies in $L^2(\Omega, \Lambda^k)$. Spaces V_0^k and V_0^{k-1} will be endowed with their natural norms.

Definition 1. The *ellipticity in the kernel* is satisfied if there exists a positive constant α, independent of p, such that

$$(d_k\mathbf{v}, d_k\mathbf{v}) \geq \alpha(\mathbf{v}, \mathbf{v}) \quad \forall \mathbf{v} \in \mathbb{K}_p^1.$$

Definition 2. The *weak approximability* of V_0^{k-1} is satisfied if there exists $\rho_1(p)$, tending to zero, such that for every $\mathbf{s} \in V_0^{k-1}$

$$\sup_{\mathbf{v} \in \mathbb{K}_p^1} \frac{(\mathbf{v}, d_{k-1}\mathbf{s})}{\|\mathbf{v}\|_{H(d_k,\Omega)}} \leq \rho_1(p)\|\mathbf{s}\|_{V_0^{k-1}}.$$

Definition 3. The *strong approximability* of V_0^k is satisfied if there exists $\rho_2(p)$, tending to zero, such that for every $\mathbf{u} \in V_0^k$ there exists $\mathbf{u}^I \in \mathbb{K}_p^1$ such that

$$\|\mathbf{u} - \mathbf{u}^I\|_{H(d_k,\Omega)} \leq \rho_2(p)\|u\|_{V_0^k}.$$

3.2 Second Mixed Formulation

We are now going to present an alternative mixed formulation of problem (2) which is a generalization of the one introduced in [10] and which makes use of the spaces $H_0(d_k, \Omega)$ and $W^{k+1} = d_k(H(d_k, \Omega)) \subset H_0(d_{k+1}, \Omega)$: find $\lambda \in \mathbb{R}$ and

$\mathbf{u} \in H_0(d_k, \Omega)$, with $\mathbf{u} \neq 0$ such that for $\boldsymbol{\psi} \in W^{k+1}$ it holds

$$\begin{cases} (\mathbf{u}, \mathbf{v}) + (d_k \mathbf{v}, \boldsymbol{\psi}) = 0 & \forall \mathbf{v} \in H_0(d_k, \Omega) \\ (d_k \mathbf{u}, \boldsymbol{\varphi}) = -\lambda(\boldsymbol{\psi}, \boldsymbol{\varphi}) & \forall \boldsymbol{\varphi} \in W^{k+1}. \end{cases} \tag{8}$$

It can be shown (see [10]) that all solutions of (8) have positive frequencies and correspond to the solutions of (2) with positive frequencies.

A discretization of (8) involves the space $V_p^k \subset H_0(d_k, \Omega)$ and a suitable discretization of $W^{k+1} \subset H_0(d_{k+1}, \Omega)$. Since we are not going to approximate (8) numerically, but we only use the mixed formulations for the numerical analysis of (2) and (4), the most natural choice for the approximation of W^{k+1} consists in taking

$$W_p^{k+1} = d_k(V_p^k). \tag{9}$$

In particular, equation (9) is the analogue of (7) for this mixed formulation. The discrete problem is: find $\lambda_p \in \mathbb{R}$ and $\mathbf{u}_p \in V_p^k$, with $\mathbf{u}_p \neq 0$ such that for $\boldsymbol{\psi}_p \in W_p^{k+1}$ it holds

$$\begin{cases} (\mathbf{u}_p, \mathbf{v}) + (d_k \mathbf{v}, \boldsymbol{\psi}_p) = 0 & \forall \mathbf{v} \in V_p^k \\ (d_k \mathbf{u}_p, \boldsymbol{\varphi}) = -\lambda_p(\boldsymbol{\psi}_p, \boldsymbol{\varphi}) & \forall \boldsymbol{\varphi} \in W_p^{k+1}. \end{cases} \tag{10}$$

Thanks to (9) is follows that all solutions of (10) correspond to the solutions of (4) with positive frequencies. As for the first mixed formulation, we can then analyze the convergence of problem (10) to (8) in order to study the convergence of the positive solutions of (4) towards the positive solutions of (2).

We now describe the conditions presented in [13, Sect. 4] which are sufficient (and in a sense necessary) for the convergence of (10) to (8). We consider the discrete kernel of the operator d_k, that is

$$\mathbb{K}_p^2 = \{\mathbf{v} \in V_p^k : (d_k \mathbf{v}, \boldsymbol{\varphi}) = 0 \ \forall \boldsymbol{\varphi} \in W_p^{k+1}\}.$$

Moreover, we need the solutions spaces W_0^k and W_0^{k+1} which contain all the first and second components $\mathbf{u} \in H(d_k, \Omega)$ and $\boldsymbol{\psi} \in W^{k+1}$, respectively, of the solution of the source problem

$$\begin{cases} (\mathbf{u}, \mathbf{v}) + (d_k \mathbf{v}, \boldsymbol{\psi}) = 0 & \forall \mathbf{v} \in H_0(d_k, \Omega) \\ (d_k \mathbf{u}, \boldsymbol{\varphi}) = -(\mathbf{g}, \boldsymbol{\varphi}) & \forall \boldsymbol{\varphi} \in W^{k+1}. \end{cases}$$

when $\mathbf{g}$ varies in $L^2(\Omega, \Lambda^{k+1})$. The spaces W_0^k and W_0^{k+1} are endowed with their natural norms.

Definition 4. The *weak approximability* of W_0^{k+1} is satisfied if there exists $\rho_3(p)$, tending to zero, such that

$$(d_k \mathbf{v}, \boldsymbol{\varphi}) \leq \rho_3(p) \|\mathbf{v}\|_{L^2(\Omega, \Lambda^k)} \|\boldsymbol{\varphi}\|_{W_0^{k+1}} \quad \forall \mathbf{v} \in \mathbb{K}_p^2 \ \forall \boldsymbol{\varphi} \in W_0^{k+1}.$$

Definition 5. The *strong approximability* of W_0^{k+1} is satisfied if there exists $\rho_4(p)$, tending to zero, such that for every $\boldsymbol{\psi} \in W_0^{k+1}$ there is $\boldsymbol{\psi}^I \in W_p^{k+1}$ with

$$\|\boldsymbol{\psi} - \boldsymbol{\psi}^I\|_{L^2(\Omega,\Lambda^{k+1})} \leq \rho_4(p)\|\boldsymbol{\psi}\|_{W_0^{k+1}}.$$

An operator $\Pi_p : W_0^k \to V_h^k$ is called *Fortin* operator if it satisfies

$$\begin{cases} (d_k(\mathbf{u} - \Pi_p\mathbf{u}), \boldsymbol{\varphi}) = 0 & \forall \mathbf{u} \in W_0^k \ \forall \boldsymbol{\varphi} \in W_p^{k+1} \\ \|\Pi_p\mathbf{u}\|_{H(d_k,\Omega)} \leq C\|\mathbf{u}\|_{W_0^k} & \forall \mathbf{u} \in W_0^k. \end{cases}$$

Definition 6. The *Fortid* property is satisfied if there exists a Fortin operator which converges to the identity in norm, that is, there exists $\rho_5(p)$, tending to zero, such that

$$\|\mathbf{u} - \Pi_p\mathbf{u}\|_{L^2(\Omega,\Lambda^k)} \leq \rho_5(p)\|\mathbf{u}\|_{W_0^k}.$$

3.3 Discrete Compactness Property

The eigenfunctions $\mathbf{u}$ of problem (2) corresponding to nonzero frequencies are characterized by the constraint $\delta_k\mathbf{u} = 0$. The discrete compactness property mimicks, at discrete level, the compactness of the subspace of $H_0(d_k, \Omega)$ consisting of functions with vanishing δ_k, into $L^2(\Omega, \Lambda^k)$. It makes use of discrete differential forms of order $k-1$ and k: V_p^{k-1} and V_p^k are finite dimensional internal approximations of $H_0(d_{k-1}, \Omega)$ and $H_0(d_k, \Omega)$, respectively.

Definition 7. The *discrete compactness property* is satisfied if every sequence $\{\mathbf{u}_p\}$ in V_p^k, bounded in $H_0(d_k, \Omega)$ and with

$$(\mathbf{u}_p, d_{k-1}\mathbf{t}) = 0 \quad \forall \mathbf{t} \in V_p^{k-1},$$

contains a subsequence which converges in $L^2(\Omega, \Lambda^k)$.

If the space V_p^{k-1} is a good approximation of $H_0(d_{k-1}, \Omega)$ then it is not difficult to see that the limit $\mathbf{u}$ in Definition 7 satisfies $(\mathbf{u}, d\mathbf{t}) = 0$ for all $\mathbf{t} \in H_0(d_{k-1}, \Omega)$, that is $\delta_k\mathbf{u} = 0$.

Definition 8. The *strong discrete compactness property* is satisfied if the limit $\mathbf{u}$ of the subsequence in Definition 7 satisfies $\delta_k\mathbf{u} = 0$.

The strong discrete compactness property is strictly related to the the standard discrete compactness property and the (CDK) property (completeness of discrete kernels) as it has been defined in [18] and used, for instance, in [9].

Before stating our main theorem, we need to make explicit a standard approximation property for the discrete space V_p^k:

$$\lim_p \inf_{\mathbf{u}_p \in V_p^k} \|\mathbf{u} - \mathbf{u}_p\|_{H(d_k,\Omega)} = 0 \quad \forall \mathbf{u} \in H_0(d_k,\Omega) \text{ with } \delta_k \mathbf{u} = 0. \tag{11}$$

Theorem 1. *Let us suppose that* (7) *and* (9) *are satisfied, so that the setting of Sects. 3.1 and 3.2 can be adopted. Then the following three sets of conditions are equivalent:*

1. *strong discrete compactness property (Definition 8) and approximation property* (11)*;*
2. *ellipticity in the kernel (Definition 1), weak approximability of V_0^{k-1} (Definition 2, and strong approximability of V_0^k (Definition 3);*
3. *weak approximability of W_0^{k+1} (Definition 4), strong approximability of W_0^{k+1} (Definition 5), and Fortid property (Definition 6).*

Proof. Due to the page restriction of the present paper, we cannot reproduce the full proof of this result. Nevertheless, the reader is referred to [11, Theorem 3] where the analogous result has been proved in a particular case ($k = 1$, $n = 3$ and the usual proxy representation where d_1 corresponds to the curl operator). The proofs of [11, Propositions 3–6] leading to [11, Theorem 3] can be repeated practically identical in our more general setting.

We take this opportunity to remark that [11, Proposition 3] which has been proved in [33, Corollary 4.2] should assume the strong discrete compactness property and not only the standard discrete compactness property. This change has no consequences for the final result.

The meaning of Theorem 1 is that each of the three equivalent conditions is a sufficient condition for the good approximation of the positive eigenvalues (and corresponding eigenspaces) of (4) to the positive eigenvalues (and corresponding eigenspaces) of (2) in the spirit of [6]. The task of evaluating the order of convergence is usually less difficult since it is possible to take advantage of the smoothness of the eigenfunctions. We refer the interested reader to [6] for the general case and to [32] for the case of mixed approximations. A review of the abstract theory will appear in [12].

4 The p Version of Edge Finite Elements

While the h version of edge finite elements for the approximation of Maxwell's eigenvalue problem has been the object of a rich literature, few results are available about the p and hp versions. Numerical results showed that the pure spectral method (one cubic element and p going to infinity) provides good results if suitable Nédélec finite elements are used (see [34]), on the other hand the first theoretical results about

the p version of edge finite elements are presented in [15], where the two dimensional triangular case is studied for the hp version. The analysis, however, relies on a conjectured estimate which has only been demonstrated numerically. In [14] the first proof of the discrete compactness in the case of the rectangular hp version of edge elements (allowing for 1-irregular hanging nodes) has been proposed.

A significant step forward for the analysis of the p version of edge finite element comes from the results of [20], where a regularized Poincaré lifting is introduced. This is one of the main ingredients for the analysis reported in [9], where the discrete compactness in p for a wide class of edge finite elements is proved.

We refer the interested reader to the details in [9] for the technical assumptions and the abstract setting. The main conclusion, related to edge finite elements for the approximation in p of Maxwell's eigenvalue problem, is that edge finite elements satisfy the discrete compactness in p in two dimensions (triangles and parallelograms) and in three dimensions (tetrahedra and parallelepipeds). The analysis relies on the already mentioned Poincaré lifting and on recently introduced projection-based interpolation operators (see, in particular, [17, 22–25]).

It is clear that after the results of [9] an important part of the analysis has been completed; nevertheless, there are still open problems that we hope can be solved in a near future. First of all, the technique of [9] does not apply to the general hp version in a straightforward way. There, we used a fixed mesh and the estimates depend on the mesh. Then, other finite element geometries might be used (prisms, pyramids, etc.) and the case of general quadrilaterals or hexehedra can be considered for which, so far, only relatively negative results concerning the approximation properties on distorted meshes are available (see [2, 16, 27]).

The approximation theory for eigenvalue problem in the framework of differential forms has been studied recently also in [5]. There, the discrete compactness property is studied by means of suitably constructed projection-based interpolation operators which satisfy the strong property of being bounded in L^2 and are constructed by a means of an extension-regularization procedure. It is not clear whether this assumption is met by the interpolation operators used for the present analysis; it would be interesting to further investigate this point and to see whether the assumptions in [5] are stronger than the discrete compactness property discussed in this paper.

References

1. P.M. Anselone. *Collectively compact operator approximation theory and applications to integral equations*. Prentice-Hall, NJ, 1971. With an appendix by Joel Davis, Prentice-Hall Series in Automatic Computation
2. D.N. Arnold, D. Boffi, and R.S. Falk. Quadrilateral $H(\mathrm{div})$ finite elements. *SIAM J. Numer. Anal.*, 42(6):2429–2451 (electronic), 2005
3. D.N. Arnold, R.S. Falk, and R. Winther. Differential complexes and stability of finite element methods I: The de Rham complex. In D. Arnold, P. Bochev, R. Lehoucq, R. Nicolaides, and M. Shaskov, editors, *Compatible Spatial Discretizations*, volume 142 of *The IMA Volumes in Mathematics and its Applications*, pages 23–46. Springer, Berlin, 2006

4. D.N. Arnold, R.S. Falk, and R. Winther. Finite element exterior calculus, homological techniques, and applications. *Acta Numer.*, 15:1–155, 2006
5. D.N. Arnold, R.S. Falk, and R. Winther. Finite element exterior calculus: from Hodge theory to numerical stability. To appear in *Bull. Amer. Math. Soc.*, pages 74, 2010
6. I. Babuška and J. Osborn. Eigenvalue problems. In *Handbook of numerical analysis, Vol. II*, Handb. Numer. Anal., II, pages 641–787. North-Holland, Amsterdam, 1991
7. D. Boffi. Fortin operator and discrete compactness for edge elements. *Numer. Math.*, 87(2): 229–246, 2000
8. D. Boffi. A note on the de Rham complex and a discrete compactness property. *Appl. Math. Lett.*, 14(1):33–38, 2001
9. D. Boffi, M. Costabel, M. Dauge, L. Demkowicz, and R. Hiptmair. Discrete compactness for the p-version of discrete differential forms. Submitted. hal-00420150, arXiv: 0909.5079v2
10. D. Boffi, P. Fernandes, L. Gastaldi, and I. Perugia. Computational models of electromagnetic resonators: analysis of edge element approximation. *SIAM J. Numer. Anal.*, 36(4):1264–1290, 1999
11. D. Boffi. Approximation of eigenvalues in mixed form, discrete compactness property, and application to hp mixed finite elements. *Comput. Methods Appl. Mech. Engrg.*, 196(37–40):3672–3681, 2007
12. D. Boffi. Finite element approximation of eigenvalue problems. *Acta Numerica*, 19:1–120, 2010
13. D. Boffi, F. Brezzi, and L. Gastaldi. On the convergence of eigenvalues for mixed formulations. *Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4)*, 25(1–2):131–154 (1998), 1997. Dedicated to Ennio De Giorgi
14. D. Boffi, M. Costabel, M. Dauge, and L. Demkowicz. Discrete compactness for the hp version of rectangular edge finite elements. *SIAM J. Numer. Anal.*, 44(3):979–1004 (electronic), 2006
15. D. Boffi, L. Demkowicz, and M. Costabel. Discrete compactness for p and hp 2D edge finite elements. *Math. Models Methods Appl. Sci.*, 13(11):1673–1687, 2003
16. D. Boffi, F. Kikuchi, and J. Schöberl. Edge element computation of Maxwell's eigenvalues on general quadrilateral meshes. *Math. Models Methods Appl. Sci.*, 16(2):265–273, 2006
17. W. Cao and L. Demkowicz. Optimal error estimate of a projection based interpolation for the p-version approximation in three dimensions. *Comput. Math. Appl.*, 50(3–4):359–366, 2005
18. S. Caorsi, P. Fernandes, and M. Raffetto. On the convergence of Galerkin finite element approximations of electromagnetic eigenproblems. *SIAM J. Numer. Anal.*, 38(2):580–607 (electronic), 2000
19. F. Chatelin. *Spectral approximation of linear operators*. Computer Science and Applied Mathematics. Academic Press Inc. [Harcourt Brace Jovanovich Publishers], New York, 1983. With a foreword by P. Henrici, With solutions to exercises by Mario Ahués
20. M. Costabel and A. McIntosh. On Bogovskii and regularized Poincaré integral operators for de Rham complexes on Lipschitz domains. *Math. Z.*, 2009. doi: 10.1007/s00209-009-0517-8, http://arxiv.org/abs/0808.2614v1
21. M. Costabel and M. Dauge. Computation of resonance frequencies for Maxwell equations in non-smooth domains. In M. Ainsworth, editor, *Computational Methods in Wave Propagation*, pages 127–164. Springer, New York, 2003
22. L. Demkowicz. Polynomial exact sequences and projection-based interpolation with applications to Maxwell equations. In D. Boffi and L. Gastaldi, editors, *Mixed Finite Elements, Compatibility Conditions, and Applications*, volume 1939 of *Lecture Notes in Mathematics*, pages 101–158. Springer, Berlin, 2008
23. L. Demkowicz and I. Babuška. p interpolation error estimates for edge finite elements of variable order in two dimensions. *SIAM J. Numer. Anal.*, 41(4):1195–1208, 2003
24. L. Demkowicz and A. Buffa. H^1, $\mathbf{H}(\mathbf{curl})$, and $\mathbf{H}(\mathrm{div})$-conforming projection-based interpolation in three dimensions. Quasi-optimal p-interpolation estimates. *Comput. Meth. Appl. Mech. Engr.*, 194:267–296, 2005
25. L. Demkowicz and J. Kurtz. Projection-based interpolation and automatic hp-adaptivity for finite element discretizations of elliptic and Maxwell problems. In *Proceedings of Journées d'Analyse Fonctionnelle et Numérique en l'honneur de Michel Crouzeix*, volume 21 of *ESAIM Proceedings*, pages 1–15, Les Ulis, 2007. EDP Science

26. J. Descloux, N. Nassaf, and J. Rappaz. On spectral approximation. Part I. The problem of convergence. *R.A.I.R.O. Numerical Analysis*, 12(2):97–112, 1978
27. R.S. Falk, P. Gatto, and P. Monk. Hexahedral H(div) and H(curl) finite elements. To appear in ESAIM: Mathematical Modelling and Numerical Analysis (M2AN)
28. R. Hiptmair. Finite elements in computational electromagnetism. *Acta Numer.*, 11:237–339, 2002
29. F. Kikuchi. Mixed and penalty formulations for finite element analysis of an eigenvalue problem in electromagnetism. In *Proceedings of the first world congress on computational mechanics (Austin, Tex., 1986)*, volume 64 (1–3), pages 509–521, 1987
30. F. Kikuchi. On a discrete compactness property for the Nédélec finite elements. *J. Fac. Sci. Univ. Tokyo Sect. IA Math.*, 36(3):479–490, 1989
31. F. Kikuchi. Theoretical analysis of Nédélec's edge elements. *Japan J. Ind. Appl. Math.*, 18(2):321–333, 2001
32. B. Mercier, J. Osborn, J. Rappaz, and P.-A. Raviart. Eigenvalue approximation by mixed and hybrid methods. *Math. Comp.*, 36(154):427–453, 1981
33. P. Monk and L. Demkowicz. Discrete compactness and the approximation of Maxwell's equations in $\mathbb{R}^3$. *Math. Comp.*, 70(234):507–523, 2001
34. P. Monk, Y. Wang, and B. Szabo. Computing cavity models using the p-version of the finite element method. *IEEE Trans. Magnetics*, 32:37–46, 1996
35. P. Monk. *Finite element methods for Maxwell's equations*. Numerical Mathematics and Scientific Computation. Oxford University Press, New York, 2003
36. R. Picard. An elementary proof for a compact imbedding result in generalized electromagnetic theory. *Math. Z.*, 187:151–161, 1984
37. F. Stummel. Diskrete Konvergenz linearer Operatoren. I. *Math. Ann.*, 190:45–92, 1970
38. G.M. Vaĭnikko. Discretely compact sequences. (russian). *Ž. Vyčisl. Mat. i Mat. Fiz.*, 14:575–583, 1974

Semi-Implicit DGFE Discretization of the Compressible Navier–Stokes Equations: Efficient Solution Strategy

Vít Dolejší and M. Holík

Abstract We deal with the numerical solution of the compressible Navier–Stokes equations with the aid of the semi-implicit discontinuous Galerkin method. We focus on the solution of the arising linear algebra systems and propose a new efficient strategy for the steady-state solutions. The efficiency is demonstrated by a set of numerical experiments.

1 Introduction

Our aim is to develop a sufficiently robust, efficient and accurate numerical scheme for the simulation of viscous compressible flows. The discontinuous Galerkin method (DGM) was employed in many papers for the discretization of compressible fluid flow problems, see, e.g., [2, 3, 5, 6, 9, 11, 13, 14, 16] and the references cited therein. DGM is based on a piecewise polynomial but discontinuous approximation which provides robust and high-order accurate approximations, particularly in transport dominated regimes. We employ the interior penalty Galerkin (IPG) methods.

In many physical applications, we are interested in the *steady-state* flow regimes when the solution is time independent. It is possible to solve the so-called *stationary Navier–Stokes equations* directly but a serious difficulty is a necessity to solve a system of strongly nonlinear algebraic equations, where the Newton-like method is employed usually, see, e.g., [14].

Another possibility is a solution of the *nonstationary Navier–Stokes equation* with the aid of the (pseudo-) time stabilization technique. Then it is possible to use an explicit time discretization where the main drawback is a high restriction of the size of the time step. On the other hand, a fully implicit time discretization

V. Dolejší (✉) and M. Holík
Charles University Prague, Faculty of Mathematics and Physics, Sokolovská 83, Prague, Czech Republic
e-mail: dolejsi@karlin.mff.cuni.cz

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_2,

leads to a necessity to solve a nonlinear system of algebraic equations at each time step. Therefore, in [6, 7, 11, 13], we developed the *semi-implicit time discretization method* which is based on a suitable linearization of the inviscid and viscous fluxes. The linear terms are treated implicitly whereas the nonlinear ones explicitly which leads to a linear algebraic problem at each time step. We call this approach the backward difference formula–discontinuous Galerkin finite element (BDF–DGFE) method. Fir a survey about semi-implicit approach see, e.g., in [15].

The BDF–DGFE method leads to a sequence of linear algebraic problems which should be solved by a suitable solver. Numerical experiments presented in [6] showed that the solution of linear algebra problem consume approximately 95–99% of the total computational time. Therefore, a significant reduce of computational time needed for the solution of these linear algebra problems is a necessary condition for a practical employment of the BDF–DGFE method. Moreover, the amount of the used computer memory has to be taken into account. Within this paper, we develop an efficient solution strategy for the mentioned algebraic problems, namely we deal with the choices stopping criterion and the size of the time step.

The content of the rest of the paper is the following. In Sect. 2, we introduce the system of the compressible Navier–Stokes equations. In Sect. 3, we recall the BDF–DGFE discretization of the Navier–Stokes equations from [6]. In Sect. 4, we discuss numerical solution of the arising linear algebra systems, propose an "optimal" strategy and demonstrate its efficiency in Sect. 5. Finally, we finish with some concluding remarks.

2 Compressible Flow Problem

Let $\Omega \subset I\!R^d$, $d = 2, 3$ be a bounded domain and $T > 0$. We set $Q_T = \Omega \times (0, T)$ and by $\partial\Omega$ denote the boundary of Ω. The system of Navier–Stokes equations describing a motion of viscous compressible fluids can be written in the dimensionless form

$$\frac{\partial \mathbf{w}}{\partial t} + \sum_{s=1}^{d} \frac{\partial \mathbf{f}_s(\mathbf{w})}{\partial x_s} = \sum_{s=1}^{d} \frac{\partial}{\partial x_s} \left(\sum_{k=1}^{d} \mathsf{K}_{sk}(\mathbf{w}) \frac{\partial \mathbf{w}}{\partial x_k} \right) \quad \text{in } Q_T, \tag{1}$$

where $\mathbf{w} = (\rho, \rho v_1, \ldots, \rho v_d, e)^{\mathrm{T}}$ is the *state vector*, $\mathbf{f}_s : I\!R^{d+2} \to I\!R^{d+2}$, $s = 1, \ldots, d$ are the inviscid (Euler) fluxes and $\mathsf{K}_{sk} : I\!R^{d+2} \to I\!R^{(d+2)\times(d+2)}$, $s, k = 1, \ldots, d$ represent the viscous terms. The forms of vectors $\mathbf{f}_s$, $s = 1, \ldots, d$ and matrices K_{sk} can be found, e.g., in [6] or [12, Sect. 4.3]. The system (1) is equipped with a suitable set of the initial and boundary conditions, see [5, 6]. The problem to solve the Navier–Stokes equations (1) equipped with the initial and boundary conditions will be denoted by (CFP) (compressible flow problem).

3 DGFE Discretization

Let $\mathcal{T}_h = \{K\}$ $(h > 0)$ be a partition of the domain Ω into a finite number of closed d-dimensional mutually disjoint (convex or non-convex) polyhedra K.

To each $K \in \mathcal{T}_h$, we assign a positive integer p_K (local polynomial degree). Then we define the vector $\mathsf{p} = \{p_K, K \in \mathcal{T}_h\}$. Over the triangulation $\mathcal{T}_h$ we define the space of discontinuous piecewise polynomial functions associated with the vector p by $S_{h\mathsf{p}} = \{v;\ v \in L^2(\Omega), v|_K \in P_{p_K}(K)\ \forall K \in \mathcal{T}_h\}$, where $P_{p_K}(K)$ denotes the space of all polynomials on K of degree $\leq p_K$, $K \in \mathcal{T}_h$. We seek the approximate solution in the space of vector-valued functions $\mathbf{S}_{h\mathsf{p}} = S_{h\mathsf{p}} \times \cdots \times S_{h\mathsf{p}}$ $(d + 2$ times).

The system of the Navier–Stokes equations (1) is discretized by the so-called backward difference formula–discontinuous Galerkin finite element (BDF–DGFE) method presented in [6], which leads to the following forms

$$\begin{aligned} c_h\,(\bar{\mathbf{w}}_h, \mathbf{w}_h, \boldsymbol{\varphi}_h) &: \mathbf{S}_{h\mathsf{p}} \times \mathbf{S}_{h\mathsf{p}} \times \mathbf{S}_{h\mathsf{p}} \to I\!R, \\ \tilde{c}_h\,(\bar{\mathbf{w}}_h, \boldsymbol{\varphi}_h) &: \quad \mathbf{S}_{h\mathsf{p}} \times \mathbf{S}_{h\mathsf{p}} \to I\!R, \end{aligned} \tag{2}$$

which are nonlinear with respect their first arguments but linear with respect to the other ones. It is possible to show (see, e.g., [5, 6]) that if $\mathbf{w} : \Omega \times (0, T) \to I\!R^{d+2}$ is a continuously differentiable function satisfying the Navier–Stokes equations (1) and the corresponding initial and boundary conditions then

$$\frac{\mathrm{d}}{\mathrm{d}t}\,(\mathbf{w}, \boldsymbol{\varphi}) + c_h\,(\mathbf{w}, \mathbf{w}, \boldsymbol{\varphi}) = \tilde{c}_h\,(\mathbf{w}, \boldsymbol{\varphi}) \quad \forall \boldsymbol{\varphi} \in \mathbf{S}_{h\mathsf{p}}, \tag{3}$$

where $(\cdot, \cdot)$ denotes the L^2-scalar product over Ω.

In [6], we introduced the following method. Let $0 = t_0 < t_1 < t_2 < \ldots t_r = T$ be a partition of the time interval $(0, T)$ and $\mathbf{w}_h^k \in \mathbf{S}_{h\mathsf{p}}$ denotes a piecewise polynomial approximation of $\mathbf{w}_h(t_k)$, $k = 0, 1, \ldots, r$.

Definition 1. We define the *approximate solution* of (CFP) by the 1-step BDF-DGFE scheme as functions $\mathbf{w}_{h,k}$, $k = 1, \ldots, r$, satisfying the conditions

a) $\mathbf{w}_{h,k} \in \mathbf{S}_{h\mathsf{p}}$, (4)

b) $\dfrac{1}{\tau_k}(\mathbf{w}_{h,k} - \mathbf{w}_{h,k-1}, \boldsymbol{\varphi}_h) + c_h(\mathbf{w}_{h,k-1}, \mathbf{w}_{h,k}, \boldsymbol{\varphi}_h) = \tilde{c}_h(\mathbf{w}_{h,k-1}, \boldsymbol{\varphi}_h)\ \ \forall\, \boldsymbol{\varphi}_h \in \mathbf{S}_{h\mathsf{p}}$

c) $\mathbf{w}_{h,0} \in S_{h\mathsf{p}}$ is an approximation of $\mathbf{w}^0$.

Remark 1. The 1-step BDF–DGFE scheme (4), (a–c) has only the first order of accuracy with respect to time which is sufficient for the seeking of the steady-state solutions. For n-step BDF–DGFE scheme $(n \geq 2)$ see [6, 8].

Remark 2. The resulting BDF–DGFE method is practically unconditionally stable, has a high order of accuracy with respect to the time and space coordinates and

at each time step we have to solve only one linear algebra problem, which will be discussed in the following section.

4 Solution of Linear Algebra Problems

4.1 Linear Algebra Representations

Since $\mathbf{S}_{hp}$ is a space of discontinuous piecewise polynomial functions, it is natural to construct its basis in such a way that the support of each basis function lies within one $K \in \mathscr{T}_h$. Then, let $B = \{\boldsymbol{\psi}_j,\ \boldsymbol{\psi}_j \in \mathbf{S}_{hp},\ j = 1, \ldots, \mathsf{dof}\}$ denote a basis of $\mathbf{S}_{hp}$ with dimension dof.

Therefore, a function $\mathbf{w}_{h,k} \in \mathbf{S}_{hp}$ can be written in the form

$$\mathbf{w}_{h,k}(x) = \sum_{j=1}^{\mathsf{dof}} \xi_{k,j} \boldsymbol{\psi}_j(x), \quad x \in \Omega,\ k = 0, 1, \ldots, r, \tag{5}$$

where $\xi_{k,j} \in I\!R,\ j = 1, \ldots, \mathsf{dof},\ k = 0, \ldots, r$. Moreover, for $\mathbf{w}_{h,k} \in \mathbf{S}_{hp}$, we define a vector of its basis coefficients by $\boldsymbol{\xi}_k = \{\xi_{k,j}\}_{j=1,\ldots,\mathsf{dof}} \in I\!R^{\mathsf{dof}},\ k = 0, 1, \ldots, r$. Using (5) we have an isomorphism

$$\mathbf{w}_{h,k} \in \mathbf{S}_{hp} \quad \longleftrightarrow \quad \boldsymbol{\xi}_k \in I\!R^{\mathsf{dof}}. \tag{6}$$

Finally, if B is an *orthonormal basis* (which can be simply constructed by an orthogonalization procedure element-wise) then we have

$$\|\mathbf{w}_{h,k}\|_{L^2(\Omega)} = \|\boldsymbol{\xi}_k\|_{\ell^2} \tag{7}$$

for any $\mathbf{w}_{h,k} \in \mathbf{S}_{hp}$ and the corresponding $\boldsymbol{\xi}_k \in I\!R^{\mathsf{dof}}$ via (6).

Then the problems (4) can be written in the matrix form:

$$\text{find } \boldsymbol{\xi}_k \in I\!R^{\mathsf{dof}} : \left(\frac{1}{\tau_k}\mathsf{M} + \mathscr{C}_h(\boldsymbol{\xi}_{k-1})\right)\boldsymbol{\xi}_k = \frac{1}{\tau_k}\mathbf{m}_k + \mathbf{q}(\boldsymbol{\xi}_{k-1}), \quad k = 1, \ldots, r, \tag{8}$$

where M is the block-diagonal *mass matrix* (if B is orthonormal basis with respect L^2 scalar product then M is the identity matrix) given by

$$\mathsf{M} = \{M^{i,j}\}_{i,j=1}^{\mathsf{dof}}, \quad M^{i,j} = (\boldsymbol{\psi}_i, \boldsymbol{\psi}_j), \tag{9}$$

the matrix $\mathscr{C}_h(\cdot)$ is the *flux matrix* corresponding to form $\boldsymbol{c}_h(\cdot,\cdot,\cdot)$ at t_k defined by

$$\mathscr{C}_h(\boldsymbol{\xi}_{k-1}) = \{C^{i,j}(\boldsymbol{\xi}_{k-1})\}_{i,j=1}^{\mathsf{dof}}, \quad C^{i,j}(\boldsymbol{\xi}_{k-1}) = \boldsymbol{c}_h\left(\mathbf{w}_{h,k-1},\ \boldsymbol{\psi}_j,\ \boldsymbol{\psi}_i\right), \tag{10}$$

$\mathbf{q} \in I\!R^{\mathsf{dof}}$ represents the right-hand-sides of (4), b) given by

$$\mathbf{q}(\boldsymbol{\xi}_{k-1}) = \{q^i(\boldsymbol{\xi}_{k-1})\}_{i=1}^{\mathsf{dof}}, \quad q^i(\boldsymbol{\xi}_{k-1}) = \tilde{c}_h\left(\mathbf{w}_{h,k-1},\ \boldsymbol{\psi}_i\right) \tag{11}$$

and

$$\mathbf{m}_k = \{m_k^i\}_{i=1}^{\mathsf{dof}}, \quad m_k^i = \left(\mathbf{w}_{h,k-1},\ \boldsymbol{\psi}_i\right). \tag{12}$$

In virtue of the local character of basis B it is easy to observe that the matrix $\mathscr{C}_h$ have a block structure.

Let us still mention that series of numerical experiments show that the Frobenius norm of the diagonal blocks of $\mathscr{C}_h$ is slightly higher than the norm of its off-diagonal blocks (in the same block-row). Moreover, the norm of the diagonal blocks of $\mathscr{C}_h$ is approximately 10^3 times higher than the Frobenius norm of the corresponding blocks of M.

4.2 General Solution Strategy

In case when we seek the steady state solution, problem (8) reduces to the problem:

$$\text{find } \boldsymbol{\xi} \in I\!R^{\mathsf{dof}}: \quad \mathscr{C}_h(\boldsymbol{\xi})\boldsymbol{\xi} = \mathbf{q}(\boldsymbol{\xi}). \tag{13}$$

However, problem (13) represents a system of strongly nonlinear algebraic equations whose direct solution is impossible. Then it is natural employ an iterative solver. The relation (13) offer to us to define a formal iterative process:

$$\begin{aligned} &\text{i) initiate } \boldsymbol{\xi}_0 \in I\!R^{\mathsf{dof}} \\ &\text{ii) find } \boldsymbol{\xi}_k \in I\!R^{\mathsf{dof}}: \quad \mathscr{C}_h(\boldsymbol{\xi}_{k-1})\boldsymbol{\xi}_k = \mathbf{q}(\boldsymbol{\xi}_{k-1}), \quad k = 1, \ldots, \\ &\text{iii) } \boldsymbol{\xi} = \lim_{k\to\infty} \boldsymbol{\xi}_k. \end{aligned} \tag{14}$$

However, numerical experiments show that this iterative process often fails which is caused by the fact that usually we start from an unphysical initial state (represented here by $\boldsymbol{\xi}_0$) and then negative density or pressure often appear.

A usual way how to avoid this obstacle is a use of the unsteady formulation (8) which can be also considered as a relaxation of method (14). It means that step (ii) in (14) is replaced by

$$\text{find } \boldsymbol{\xi}_k \in I\!R^{\mathsf{dof}}: \underbrace{\left(\frac{1}{\tau_k}\mathsf{M} + \mathscr{C}_h(\boldsymbol{\xi}_{k-1})\right)}_{=:\mathsf{A}_k(\boldsymbol{\xi}_{k-1})} \boldsymbol{\xi}_k = \underbrace{\frac{1}{\tau_k}\mathbf{m} + \mathbf{q}(\boldsymbol{\xi}_{k-1})}_{=:\mathbf{d}_k(\boldsymbol{\xi}_{k-1})}, \quad k = 1, \ldots, \tag{15}$$

where $\tau_k > 0,\ k = 1, \ldots,$ can be considered as the size of the time step or as the relaxation parameter. The relation (15) represents a sequence of systems of linear

algebraic equations which has to be solved by a suitable solver. There arise two fundamental questions:

1. How to choose $\tau_k,\ k = 1, \ldots, ?$
2. How to solve (15)?

It seems to be suitable to use an iterative solver for the solution of (15) since the solution from the old step $k-1$ can be used as the initial solution in the new step k. Moreover, it is sufficient to compute the solution of (15) only approximately since we are interested only in the limit vector $\boldsymbol{\xi} = \lim_{k\to\infty} \boldsymbol{\xi}_k$. Hence, the iterative solver for the solution of (15) can be stop after not too high number of iteration. In our case, we employ the restarted GMRES method with the block diagonal preconditioning (BDP). This approach is simple for an implementation, it is fast and requires a small amount of additional memory.

Based on the previous consideration, we propose the following general solution procedure:

Algorithm (A)

1. let $\boldsymbol{\xi}_0 \longleftrightarrow \mathbf{w}_h^0$ be given
2. for $k = 1$ to r
   a. set τ_k
   b. from $\boldsymbol{\xi}_{k-1}$ evaluate $\mathsf{A}_k(\boldsymbol{\xi}_{k-1})$, $\mathbf{d}_k(\boldsymbol{\xi}_{k-1})$
   c. solve $\mathsf{A}_k(\boldsymbol{\xi}_{k-1})\boldsymbol{\xi}_k = \mathbf{d}_k(\boldsymbol{\xi}_{k-1})$ by restarted GMRES with BDP by
      i. $\boldsymbol{\xi}_k^0 := \boldsymbol{\xi}_{k-1}$
      ii. $\boldsymbol{\xi}_k^{l+1} :=$ GMRES_iter$(\boldsymbol{\xi}_k^l),\ l = 1, \ldots, s_k$
      iii. $\boldsymbol{\xi}_k := \boldsymbol{\xi}_k^{s_k}$
3. $\boldsymbol{\xi} := \boldsymbol{\xi}_r$.

In the previous algorithm r denotes the total number of used time steps and $s_k,\ k = 1, \ldots, r$ the number of inner iterative loops of the GMRES solver for the time steps t_k. These values have to be chosen on the base of suitable stopping criteria which are discussed in the following sections.

4.2.1 Steady-State Criterion

Within this section we discuss the steady-state stopping criterion, i.e., when to stop the global loops in the algorithm **(A)** for $k = 1, \ldots, r$. The usual steady-state criterion, often used for explicit time discretization, is

$$\left\| \frac{\partial \mathbf{w}_h}{\partial t} \right\| \approx \eta_k := \frac{1}{\tau_k} \|\mathbf{w}_h^k - \mathbf{w}_h^{k-1}\|_{L^2(\Omega)} = \frac{1}{\tau_k} \|\boldsymbol{\xi}_k - \boldsymbol{\xi}_{k-1}\|_{\ell^2} \leq \mathrm{TOL}, \qquad (16)$$

where TOL is a given tolerance. However, this criterion makes not good sense for the semi-implicit time discretization when very large time steps can be employed. Then there exists a limit value of τ_k when $\mathsf{A}_k \approx \mathscr{C}_{hk}$ and $\mathbf{d}_k \approx \mathbf{q}_k$ are independent

of τ_k (in the finite precision arithmetic) whereas (16) depends on τ_k. Then by a very large choice of τ_k we can achieve very small value of the left-hand side of (16) although we are far from the steady-state solution.

Therefore, in virtue of (13) we employ the following *steady-state residual criterion*

$$\mathrm{SSres}(k) := \|\mathscr{C}_h(\boldsymbol{\xi}_k)\boldsymbol{\xi}_k - \mathbf{q}(\boldsymbol{\xi}_k)\|_{\ell^2} \leq \mathrm{TOL}, \tag{17}$$

which is independent of τ_k.

Another possibility when to stop the global loops in **(A)** follows from the physical background of the considered problem. Many often, we are interested in the so-called *aerodynamic coefficients* of the considered flow, namely coefficients of *drag* (c_D), *lift* (c_L) and *momentum* (c_M). Then the natural choice is stop global iterative loops when these coefficients achieve a given tolerance, e.g.,

$$\frac{\Delta c_x(k)}{|c_x(k)|} \leq \mathrm{tol}, \qquad \Delta c_x(k) := \max_{l=0.9k,\ldots,k} c_x(l) - \min_{l=0.9k,\ldots,k} c_x(l), \tag{18}$$

where tol is a given relative tolerance, subscript x takes the value D, L and M (drag, lift, momentum), $c_x(k)$ is the value of the corresponding aerodynamical coefficient at k^{th}-time step and the minimum and maximum in (18) are taken over last 10% of the number of time steps.

Whereas the tolerance TOL in the preconditioned residuum (17) has to be chosen empirically, the tolerance tol in (18) can be chosen only on the base of our accuracy requirements (without any previous numerical experiments), e.g., $\mathrm{tol} = 0.01$.

4.2.2 GMRES Stopping Criterion

Within this section we deal with the stopping criterion of the inner loop in **(A)**, i.e., when to step the GMRES iterative process at each time step $k = 1, \ldots, r$. It is clear that too weak criterion can decrease accuracy and on the other hand, too strong criterion decreases the efficiency. Usually, one uses residuum criterion

$$\|\mathsf{A}_k(\boldsymbol{\xi}_{k-1})\boldsymbol{\xi}_k - \mathbf{d}_k(\boldsymbol{\xi}_{k-1})\| \leq \mathrm{TOL} \tag{19}$$

or the preconditioned residuum criterion

$$\|\mathscr{Q}\mathsf{A}_k(\boldsymbol{\xi}_{k-1})\boldsymbol{\xi}_k - \mathscr{Q}\mathbf{d}_k(\boldsymbol{\xi}_{k-1})\| \leq \mathrm{TOL}, \tag{20}$$

where $\mathscr{Q}$ is the matrix of preconditioning and TOL is a given tolerance. However, there is a problem how to choose TOL since there is no indication from the theory.

Hence, inspired by the so-called *inexact Newton method* from [4] we propose the following stopping criterion for GMRES method:

$$\|\mathsf{A}_k(\boldsymbol{\xi}_{k-1})\boldsymbol{\xi}_k - \mathbf{d}_k(\boldsymbol{\xi}_{k-1})\| \leq \delta_k\|\mathsf{A}_k(\boldsymbol{\xi}_{k-1})\boldsymbol{\xi}_{k-1} - \mathbf{d}_k(\boldsymbol{\xi}_{k-1})\|, \tag{21}$$

where $\delta_k \in (0,1)$ is a given value, the left hand-side is the residuum and the term on the right hand side can be considered either as *consistency residuum* from the previous time step or *initial residuum* since the solution of the previous time step is taken as an initial solution on the next time step. Concerning δ_k, two choices were proposed and analyzed in [10]. However, numerical experiments presented in Sect. 5 show that for our purposes parameters δ_k can be chosen very simply.

4.2.3 Choice of the Time Step

The choice of the time step τ_k, $k = 1, \ldots, r$ exhibits another important issue in the efficient solution of the steady-state solution. At the beginning of computation, it is necessary to choose τ_k small in order to avoid fails of computations caused by the unphysical initial condition. On the other hand, when the solution is approaching to the steady-state, we are increasing the size of τ_k in order to accelerate the computational process. In other words we are decreasing the relaxation parameter (τ_k^{-1}). In [8] we proposed the *adaptive backward difference formulae* technique which adapts the size of the time step in order to keep the local discretization error under a given tolerance and to minimize a number of time step. However, numerical experiments show that the size of τ_k is very often underestimate when we seek steady state solutions, i.e., the time step can be chosen larger.

Therefore, we propose here a new rather *heuristic* adaptive choice of the time step according to the formula

$$\tau_1 := \frac{1}{2\Lambda_k}, \qquad \tau_{k+1} := \frac{1}{2\Lambda_k}\left(\frac{\eta_k}{\eta_1}\right)^{-\omega}, \quad k = 1, \ldots, r-1, \tag{22}$$

where η_k, $k = 1, \ldots, r$ is given by (16), $\omega > 0$ is a given constant usually chosen as $\omega = 3/2$ or $\omega = 2$ and

$$\Lambda_k = \max_{K \in \mathcal{T}_h} |K|^{-1} \max_{\Gamma \in \partial K} \max_{l=1,\ldots,d+2} \lambda_l(\mathbf{w}_h^k|_\Gamma)|\Gamma| \tag{23}$$

where $\lambda_l(\mathbf{w}_h^k|_\Gamma)$ is the spectral ration of the Jacobi matrix of inviscid fluxes evaluated on $\Gamma \in \partial K$, $K \in \mathcal{T}_h$. This means that at the first step, τ_1 is chosen in the same way as for an explicit time discretization with CFL $= 0.5$, see [12]. Moreover, τ_k is exponential increasing when η_k is decreasing.

5 Numerical Examples

In the previous section we presented the new solution strategy for the solving the steady-state solutions of the Navier–Stokes equations. However, there are still some undefined parameters whose choice will be discussed in the following. We show that these choices are very robust. Finally, we show a 3D illustrative example.

5.1 Numerical Study

Within this section we numerically study two still open questions:

- Choice of δ_k in (21),
- Choice of the number of restarts in GMRES solver.

Finally, we present a comparison the new strategy with the former BDF–DGFE method from [6].

We deal with a viscous compressible flow around NACA 0012 profile with inlet Mach number $M_{\text{inlet}} = 0.5$, angle of attack $\alpha = 2°$ and Reynolds number $Re = 5{,}000$. We employ a triangular grid having 2,394 elements (see Fig. 1) and a piecewise cubic polynomial approximation. The computational processes are stopped if condition (17) is valid with $\text{TOL} = 10^{-3}$ and condition (18) is valid with $\text{tol} = 10^{-2}$ for drag, lift and momentum coefficients.

Figure 2 shows the dependence of SSres defined by (17) on the number of time steps and the computational time in seconds for $\delta = \delta_k = 0.9, 0.5, 0.1, 0.02, 0.005$, $k = 1, 2 \ldots$. We see that small values of δ increase the computational time whereas there is almost negligible difference in computational time for $\delta \in [0.1; 0.9]$. Hence, we can simply put, e.g., $\delta = 0.5$ and this value will be (almost) optimal.

Moreover, Fig. 3 shows the dependence of SSres on the number of time steps and the computational time in seconds for different number of loops in GMRES method after which the GMRES is restarted (namely 20, 30, 40, 50, 60 loops). We observe that high number in inner loops within one restart is more efficient but the difference between 50 and 60 is again almost negligible. Hence, we use the restart after 50 loops in the following.

Furthermore, Table 1 shows a comparison of BDF–DGFE method presented in [6] with the new approach developed here, namely the number of time steps and computational time. The increase of efficiency (=decrease of computational time) is evident. This table also contains relative computational costs necessary for preparation and itself solution of linear algebra problems. For the new method, this ratio is equal almost to the optimal one 50%:50%.

Finally, Table 2 present the comparisons of relative computational costs necessary for preparation and itself solution of linear algebra problems carried out by P_1,

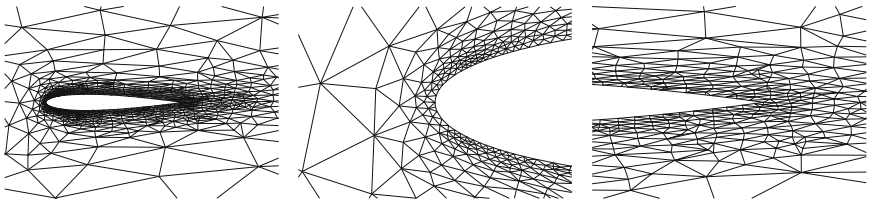


Fig. 1 The used triangular mesh around NACA0012 profile with details around leading and trailing edges

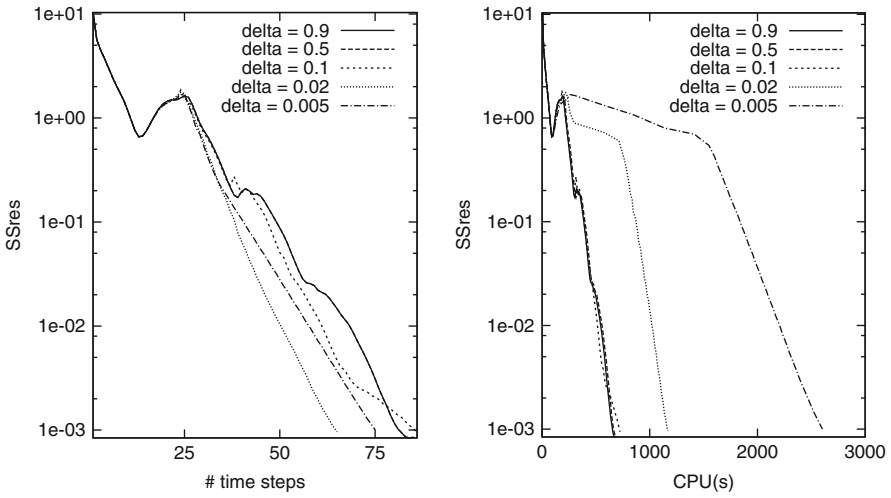


Fig. 2 Dependence of SSres on the number of time steps and the computational time for $\delta = 0.9$, 0.5, 0.1, 0.02, 0.005 in (17)

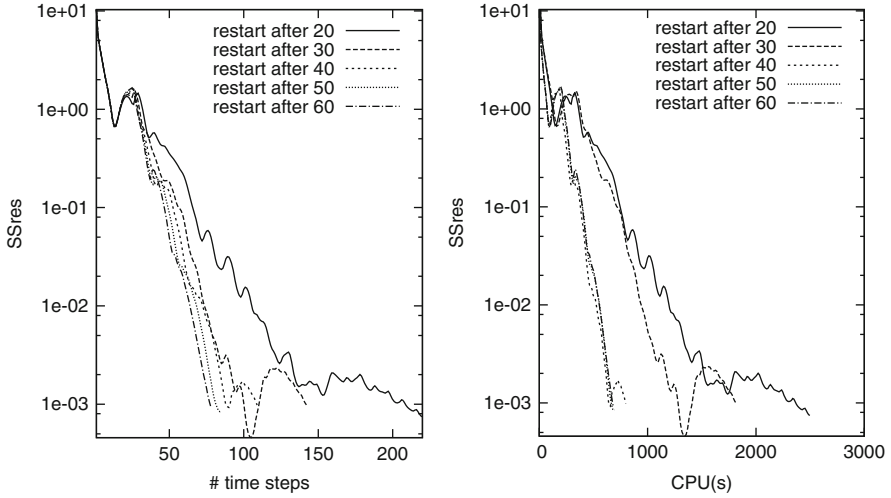


Fig. 3 Dependence of SSres on the number of time steps and the computational time for restart after 20, 30, 40, 50, 60 loops in GMRES method

P_2 and P_3 polynomial approximations for the mesh from Fig. 1 and a finner one. We simply observe that the ratios are close to the optimal one (50%:50%) and moreover, they are still better for finner grids (at least for P_1 and P_2) and higher degrees of polynomial approximations. Hence, our approach seems to be robust with respect to h and p.

Table 1 Comparison of BDF–DGFE method with the new approach, number of time steps, computational time and relative computational costs necessary for preparation and itself solution of linear algebra problems

Method	# time steps	CPU (s)	Preparing	Solving
BDF– DGFEM	273	10,774	12% CPU	88% CPU
New approach	85	695	42% CPU	58% CPU

Table 2 New approach, comparison of relative computational costs necessary for preparation and itself solution of linear algebra problems for two grids and different degree of polynomial approximations

P_k	$\#\mathcal{T}_h$	dof	Preparing $\mathsf{A}_k, \mathbf{d}_k$	Solving $\mathsf{A}_k \boldsymbol{\xi}_k = \mathbf{d}_k$
P_1	2,394	28,728	31%	69%
P_1	4,214	50,568	33%	67%
P_2	2,394	57,456	36%	64%
P_2	4,214	101,136	37%	63%
P_3	2,394	95,760	42%	58%
P_3	4,214	168,560	41%	59%

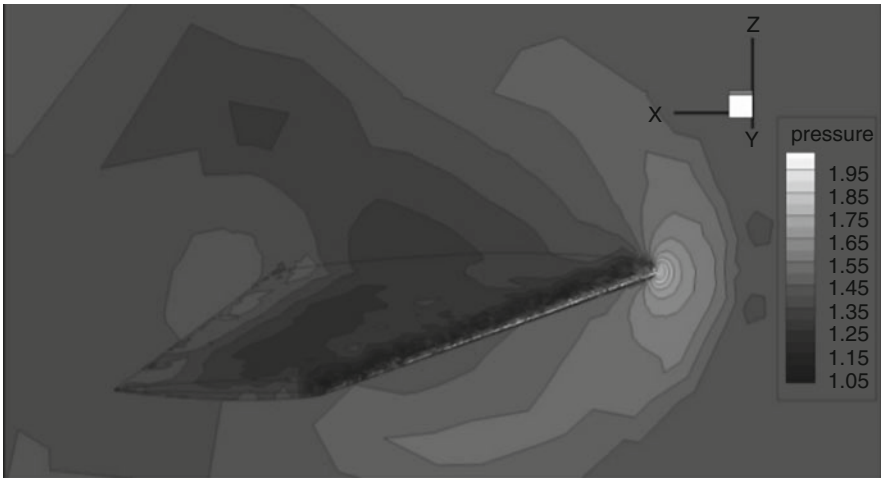


Fig. 4 ONERA M6 wing, distribution of the pressure

5.2 3D Test Case

Finally, we show an illustrative 3D laminar viscous flow around the ONERA M6 wing with the inlet Mach number $M_{\text{inlet}} = 0.71$, angle of attack $\alpha = 3.06°$ and the Reynolds number $Re = 5{,}000$ which was solved within the project ADIGMA [1]. Figure 4 shows a distribution of the pressure around the profiles. In order to obtain better resolution an adaptive mesh refinement has to be employed.

6 Conclusion

We developed an efficient technique for the solution of steady state viscous compressible flows. The key feature is a week stopping criterion of the linear algebra systems arising form the semi-implicit time discretization. Numerical experiments show that solution of these systems requires approximately the same computational time as the setting of these systems itself. Moreover, this approach is robust with respect to h and p.

Acknowledgements This work is a part of the research project MSM 0021620839 financed by the Ministry of Education of the Czech Republic and it was partly supported by the grant No. 201/08/0012 of the Grant Agency of the Czech Republic. The research of M. Holík was supported by the Grant No. 10209 of the Grant Agency of the Charles University Prague and the research project ADIGMA, No. 30719 financed within the 3rd Call of the 6th European Framework Programme.

References

1. ADIGMA: Adaptive higher-order variational methods for aerodynamic applications in industry, Specific Targeted Research Project no. 30719 supported by European Commision. URL: http://www.dlr.de/as/en/Desktopdefault.aspx/tabid-2035/2979_read-4582/
2. Bassi, F., Rebay, S.: A high order discontinuous Galerkin method for compressible turbulent flow. In: B. Cockburn, G.E. Karniadakis, C.W. Shu (eds.) Discontinuous Galerkin Method: Theory, Computations and Applications, Lecture Notes in Computational Science and Engineering 11, pp. 113–123. Springer, Berlin (2000)
3. Baumann, C.E., Oden, J.T.: A discontinuous hp finite element method for the Euler and Navier–Stokes equations. Int. J. Numer. Methods Fluids **31**(1), 79–95 (1999)
4. Dembo, R.S., Eisenstat, S.C., Steihaug, T.: Inexact newton methods. SIAM J. Numer. Anal. **19**, 400–408 (1982)
5. Dolejší, V.: On the discontinuous Galerkin method for the numerical solution of the Navier–Stokes equations. Int. J. Numer. Methods Fluids **45**, 1083–1106 (2004)
6. Dolejší, V.: Semi-implicit interior penalty discontinuous Galerkin methods for viscous compressible flows. Commun. Comput. Phys. **4**(2), 231–274 (2008)
7. Dolejší, V., Feistauer, M.: Semi-implicit discontinuous Galerkin finite element method for the numerical solution of inviscid compressible flow. J. Comput. Phys. **198**(2), 727–746 (2004)
8. Dolejší, V., Kůs, P.: Adaptive backward difference formula – discontinuous Galerkin finite element method for the solution of conservation laws. Int. J. Numer. Methods Eng. **73**(12), 1739–1766 (2008)
9. Dumbser, M., Munz, C.D.: Building blocks for arbitrary high-order discontinuous Galerkin methods. J. Sci. Comput. **27**, 215–230 (2006)
10. Eisenstat, S.C., Walker, H.: Choosing the forcing terms in inexact newton method. SIAM J. Sci. Comput. **17**(1), 16–32 (1996)
11. Feistauer, M., Kučera, V.: On a robust discontinuous galerkin technique for the solution of compressible flow. J. Comput. Phys. **224**(1), 208–221 (2007)
12. Feistauer, M., Felcman, J., Straškraba, I.: Mathematical and Computational Methods for Compressible Flow. Oxford University Press, Oxford (2003)
13. Feistauer, M., Kučera, V., Prokopová, J.: Discontinuous Galerkin solution of compressible flow in time dependent domains. Mathematics and Computers in Simulations (2009). doi: 10.1016/j.matcom.2009.01.020

14. Hartmann, R., Houston, P.: Symmetric interior penalty DG methods for the compressible Navier–Stokes equations I: Method formulation. Int. J. Numer. Anal. Model. **1**, 1–20 (2006)
15. Keppens, R., Tóth, G., Botchev, M.A., van der Ploeg, A.: Implicit and semi-immplicit schemes: Algorithms. Int. J. Numer. Meth. Fluids **30**, 335–352 (1999)
16. Klaij, C.M., van der Vegt, J., der Ven, H.V.: Pseudo-time stepping for space-time discontinuous Galerkin discretizations of the compressible Navier–Stokes equations. J. Comput. Phys. **219**(2), 622–643 (2006)

Some Numerical Approaches for Weakly Random Homogenization

Claude Le Bris

Abstract We overview a series of recent works addressing homogenization problems for some materials seen as small random perturbations of periodic materials (in a sense made precise in the body of the text). These recent works are joint works with several collaborators: Blanc (Paris 6), Lions (Collège de France), Legoll, Anantharaman, Costaouec (Ecole Nationale des Ponts et Chaussées and INRIA). The theory, developed in [C. R. Acad. Sci. Série I, 343, 717–724 (2006), Journal de Mathématiques Pures et Appliquées, 88, 34–63 (2007)], is only outlined. Next a collection of numerical appropriate approaches introduced in [Note aux Comptes Rendus de l'Académie des Sciences (2009), Thèse de l' Université Paris Est, C. R. Acad. Sci. Série I, 348, 99–103 (2010)] is presented. The theoretical considerations and the numerical tests provided here show that for the materials with only a small amount of randomness that are considered, a dedicated approach is far more efficient than a direct, stochastic approach.

1 Introduction

Multiscale approaches are increasingly popular in computational materials science. Although much effort has been devoted lately to the development of appropriate, computationally efficient approaches, there is still room for improvement, given the enormous variety of the field.

The motivation for the works summarized in the present review is contained in the following four-fold observation:

1. *A new feature that becomes ubiquitous in computational materials science is randomness*. Most of the simulations performed in the past decades, including

C. Le Bris
CERMICS, École Nationale des Ponts et Chaussées, 6 & 8, avenue Blaise Pascal, 77455 Marne-La-Vallée Cedex 2 and INRIA Rocquencourt, MICMAC project, Domaine de Voluceau, B.P. 105, 78153 Le Chesnay Cedex, France
e-mail: lebris@cermics.enpc.fr

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_3,

the most recent development along the multiscale paradigm, consider idealized materials. Such materials are flawless, and most of the time perfectly periodic. In sharp contrast, real materials have defects, and have several characteristic length-scales that differ from one another by orders of magnitude. Their qualitative and quantitative response to environment might therefore differ a lot from the idealized scenario. Think for instance of solid materials consisting of grains, each grain being a particular assembly of monocrystals, each of them in turn possibly separated by interfaces and possibly embedding dislocations.

2. *Multiscale simulations, already computationally expensive* per se *may admittedly become prohibitively expensive in the presence of randomness.* A good example (the topic of the present review article) is random homogenization, which is *infinitely* more expensive than periodic homogenization (basically because it requires solving corrector problems posed on the entire space, see (2) and (14) below). Alternative approaches are thus interesting.
3. *The very definition of a random material is still mostly vague.* Given a microscopic picture of a material, it is indeed unclear to decide whether the microstructures are periodically repeated, whether some type of stationary ergodic character is encoded in the microstructures, or whether a much more general type of modeling should be adopted. Defining the geometric assumption that will allow to efficiently simulate the material computationally is a challenge in its own rights.
4. *In many practical situations, the random material under consideration is not far from being a periodic material.* At zero order of approximation, the material can be considered periodic, and it is only at a higher order that randomness plays a role. A good example is provided by materials that are industrially produced, where the defect of periodicity typically owes to failures in the synthesis process. See Fig. 1. Despite its smallness, the microscopic amount of randomness might affect the macroscale at order one, and it is indeed the interesting issue to quantitatively model this effect.

Considering the above, our purpose here is to outline a modeling strategy that accounts for the presence of randomness in a multiscale computation, but specifically addresses the case when the amount of randomness present in the system is small, in a sense to be made precise below. The weakly random material is thus considered as a small perturbation of a periodic material. Based on this interpretation, an efficient numerical strategy is then devised. It only aims at computing an *approximation* of the response of the material, given that the randomness is weak. But, as shown in the sequel, the strategy is computationally much less expensive than a direct stochastic approach.

The context in which we develop our approach is homogenization theory, and more precisely homogenization of simple, second order elliptic equations in divergence form with highly oscillatory coefficients. This particular case is to be thought of as a prototypical case. Although we have not developed our theory and computations for other, more general equations and settings, we are convinced that the same line of approach (namely small amount of randomness as compared to a reference periodic setting, plus expansion in the randomness amplitude, and simplified computations) can be useful in many contexts.

Fig. 1 Composite material (extracted from [18], reproduced with permission of the author): it is evident from the picture (a two-dimensional cut of a three-dimensional material) that the cross section of the fibers of the materials are not arranged periodically. On the other hand, it would not be fair to say that the material is entirely disordered. Some types of ordering, at different lengthscales, can be identified on the picture

The article is articulated as follows. Section 2 recalls some basics of the theory of periodic and stochastic homogenization, and introduces some elements on a variant recently studied by the author and collaborators. Section 3 first presents the bottom line of the approach: Taylor expanding the corrector and the homogenized matrix with respect to the small parameter measuring the amount of randomness in the system. The approach is then applied, under two different variants, to some academic cases which we hope to be representative of generic practical situations. The article concludes with Sect. 4 briefly discussing related problems and techniques.

2 Some Elements of Homogenization Theory

2.1 Periodic Homogenization

To begin with, we recall some basic ingredients of elliptic homogenization theory in the periodic setting. We refer e.g., to the monographs [4, 8, 12] for more details on homogenization theory.

We consider, in a regular domain $\mathscr{D}$ in $\mathbb{R}^d$, the problem

$$\begin{cases} -\text{div}\left[A_{per}\left(\frac{x}{\varepsilon}\right)\nabla u^\varepsilon\right] = f \quad \text{in} \quad \mathscr{D}, \\ u^\varepsilon = 0 \quad \text{on} \quad \partial\mathscr{D}, \end{cases} \tag{1}$$

where the matrix A_{per} is symmetric and $\mathbb{Z}^d$-periodic. We manipulate for simplicity *symmetric* matrices, but the discussion carries over to non symmetric matrices up to slight modifications.

The corrector problem associated to (1) reads, for p fixed in $\mathbb{R}^d$,

$$\begin{cases} -\mathrm{div}(A_{per}(y)(p+\nabla w_p)) = 0, \\ w_p \text{ is } \mathbb{Z}^d\text{-periodic.} \end{cases} \tag{2}$$

It has a unique solution up to the addition of a constant. Then, the homogenized coefficients read

$$\begin{aligned} [A_*]_{ij} &= \int_Q (e_i + \nabla w_{e_i}(y))^T A_{per}(y)(e_j + \nabla w_{e_j}(y))dy \\ &= \int_Q (e_i + \nabla w_{e_i}(y))^T A_{per}(y) e_j dy, \end{aligned} \tag{3}$$

where Q is the unit cube. The main result of periodic homogenization theory is that, as ε goes to zero, the solution u^ε to (1) converges to u^* solution to

$$\begin{cases} -\mathrm{div}\,[A_*\nabla u^*] = f \quad \text{in} \quad \mathscr{D}, \\ u^* = 0 \quad \text{on} \quad \partial\mathscr{D}. \end{cases} \tag{4}$$

The convergence holds in $L^2(\mathscr{D})$, and weakly in $H_0^1(\mathscr{D})$. The correctors w_{e_i} (for e_i the canonical vectors of $\mathbb{R}^d$) may then also be used to "correct" u^* in order to identify the behavior of u^ε in the strong topology $H_0^1(\mathscr{D})$. Several other convergences on various products involving $A_{per}\left(\frac{x}{\varepsilon}\right)$ and u^ε also hold. All this is well documented.

The practical interest of the approach is evident. No small scale ε is present in the homogenized problem (4). At the price of only computing d periodic problems (2) (as many problems as dimensions in the ambient space, take indeed p the vectors of the canonical basis of $\mathbb{R}^d$), the solution to problem (1) can be efficiently approached for ε small. A direct attack of problem (1) would require taking a meshsize smaller than ε. The difficulty has been circumvented. Of course, many improvements and alternatives exist in the literature.

The proof of the above result can be performed in several ways. One approach is the *energy method* by Murat and Tartar (see [14, 17]). Another possible approach is to use the notion of *two-scale convergence* introduced by G. Nguetseng and developed by G. Allaire (see [1, 15]).

2.2 Classical Random Homogenization

The present section introduces the classical *stationary ergodic setting*. We choose to present the theory in a *discrete* stationary setting, which is more appropriate for our specific purpose in the next sections. Random homogenization is more often presented in the *continuous* stationary setting. Although the two settings are different

(neither of them being an extension of the other), the modifications needed to pass from one setting to the other are tiny, and summarized in Remark 1 below.

Throughout this article, $(\Omega, \mathscr{F}, \mathbb{P})$ denotes a probability space. For any random variable $X \in L^1(\Omega, d\mathbb{P})$, we denote by $\mathbb{E}(X) = \int_\Omega X(\omega) d\mathbb{P}(\omega)$ its expectation value. We fix $d \in \mathbb{N}^*$, and assume that the group $(\mathbb{Z}^d, +)$ acts on Ω. We denote by $(\tau_k)_{k\in\mathbb{Z}^d}$ this action, and assume that it preserves the measure $\mathbb{P}$, i.e.,

$$\forall k \in \mathbb{Z}^d, \quad \forall A \in \mathscr{F}, \quad \mathbb{P}(\tau_k A) = \mathbb{P}(A). \tag{5}$$

We assume that τ is *ergodic*, that is,

$$\forall A \in \mathscr{F}, \quad \Big(\forall k \in \mathbb{Z}^d, \quad \tau_k A = A\Big) \Rightarrow (\mathbb{P}(A) = 0 \quad \text{or} \quad 1). \tag{6}$$

In addition, we define the following notion of stationarity: any $F \in L^1_{\text{loc}}(\mathbb{R}^d, L^1(\Omega))$ is said to be *stationary* if

$$\forall k \in \mathbb{Z}^d, \quad F(x + k, \omega) = F(x, \tau_k \omega) \text{ almost everywhere in } x, \text{ almost surely.} \tag{7}$$

In this setting, the ergodic theorem [13, 16] can be stated as follows:

Theorem 1 (Ergodic theorem, [13,16]). *Let $F \in L^\infty(\mathbb{R}^d, L^1(\Omega))$ be a stationary random variable in the sense of (7). For $k = (k_1, k_2, \dots k_d) \in \mathbb{R}^d$, we set $|k|_\infty = \sup_{1\le i\le d} |k_i|$. Then*

$$\frac{1}{(2N+1)^d} \sum_{|k|_\infty \le N} F(x, \tau_k \omega) \underset{N\to\infty}{\longrightarrow} \mathbb{E}(F(x, \cdot)) \quad \textit{in } L^\infty(\mathbb{R}^d), \textit{ almost surely.} \tag{8}$$

This implies that (denoting by Q the unit cube in $\mathbb{R}^d$)

$$F\left(\frac{x}{\varepsilon}, \omega\right) \overset{*}{\underset{\varepsilon\to 0}{\longrightarrow}} \mathbb{E}\Big(\int_Q F(x, \cdot) dx\Big) \quad \textit{in } L^\infty(\mathbb{R}^d), \textit{ almost surely.} \tag{9}$$

It is useful to intuitively define stationarity and ergodicity in terms of material modeling. Pick two points x and $y \neq x$ at the microscale in the material. The particular local environment seen from x (that is, the microstructure present at x) is generically different from what is seen from y (that is, the microstructure present at y). However, the *average* local environment in x is identical to that in y (considering the various realizations of the random material). In mathematical terms, the *law* of microstructures is the same at all points. This is *stationarity*. On the other hand, *ergodicity* means that considering all the points in the material amounts to fixing a point x in this material and considering all the possible microstructures present there.

Remark 1. Alternatively to the above discrete setting, it is possible to define a continuous ergodic setting, the reader might be more familiar with. We fix $d \in \mathbb{N}^*$,

and assume that the group $(\mathbb{R}^d, +)$ acts on Ω. We denote by $(\tau_x)_{x\in\mathbb{R}^d}$ this action. We assume that it preserves the measure $\mathbb{P}$, that it is ergodic, both properties being expressed using a straightforward adaptation of (5) and (6) respectively. The notion of stationarity is defined by $F(x + y, \omega) = F(x, \tau_y\omega)$, for all $y \in \mathbb{R}$, almost everywhere in $x \in \mathbb{R}$ and almost surely. To understand the difference between the discrete and the continuous settings, note for instance that a $\mathbb{Z}^d$-periodic function F is a particular case of (7), when F is assumed to be deterministic. In contrast, it is an example of the continuous setting for a *genuinely* random function F, Ω being the d dimensional torus and $\tau_x y \equiv x + y$.

In the continuous setting, the ergodic theorem [13,16] holds. The conclusions (8) and (9) are respectively replaced by:

$$\frac{1}{|B_R|}\int_{B_R} F(x,\tau_y\omega)dy \underset{R\to\infty}{\longrightarrow} \mathbb{E}\left(F(x,\cdot)\right) = \mathbb{E}(F) \quad \text{in } L^\infty(\mathbb{R}^d), \text{ almost surely,} \tag{10}$$

and

$$F\left(\frac{x}{\varepsilon},\omega\right) \overset{*}{\underset{\varepsilon\to 0}{\longrightarrow}} \mathbb{E}\left(F\right) \quad \text{in } L^\infty(\mathbb{R}^d), \text{ almost surely.} \tag{11}$$

□

We now fix $\mathscr{D}$ an open, smooth and bounded subset of $\mathbb{R}^d$, and A a square matrix of size d, which is assumed stationary in the sense defined above, and which is assumed to enjoy the classical assumptions of uniform ellipticity and boundedness. Then we consider the boundary value problem

$$\begin{cases} -\mathrm{div}\left(A\left(\frac{x}{\varepsilon},\omega\right)\nabla u^\varepsilon\right) = f \quad \text{in} \quad \mathscr{D}, \\ u^\varepsilon = 0 \quad \text{on} \quad \partial\mathscr{D}. \end{cases} \tag{12}$$

Standard results of stochastic homogenization [4, 12] apply and allow to find the homogenized problem for problem (12). These results generalize the periodic results recalled in Sect. 2.1. The solution u^ε to (12) converges to the solution to (4) where the homogenized matrix is now defined as:

$$[A_*]_{ij} = \mathbb{E}\left(\int_Q (e_i + \nabla w_{e_i}(y,\cdot))^T A(y,\cdot)e_j\, dy\right), \tag{13}$$

where for any $p \in \mathbb{R}^d$, w_p is the solution (unique up to the addition of a (random) constant) in $\{w \in L^2_{\mathrm{loc}}(\mathbb{R}^d, L^2(\Omega)), \nabla w \in L^2_{\mathrm{unif}}(\mathbb{R}^d, L^2(\Omega))\}$ to

$$\begin{cases} -\mathrm{div}[A(y,\omega)(p + \nabla w_p(y,\omega))] = 0, \quad \text{a.s. on} \quad \mathbb{R}^d \\ \nabla w_p \quad \text{is stationary in the sense of (7),} \\ \mathbb{E}\left(\int_Q \nabla w_p(y,\cdot)\, dy\right) = 0. \end{cases} \tag{14}$$

We have used above the notation L^2_{unif} for the *uniform* L^2 space, that is the space of functions for which, say, the L^2 norm on a ball of unit size is bounded above independently from the center of the ball.

A striking difference between the stochastic setting and the periodic setting can be observed comparing (2) and (14). In the periodic case, the corrector problem is posed on a bounded domain (namely, the periodic cell Q), since the corrector w_p is periodic. In sharp contrast, the corrector problem (14) of the random case is posed on the whole space $\mathbb{R}^d$, and cannot be reduced to a problem posed on a bounded domain. The reason is, condition $\mathbb{E}\left(\int_Q \nabla w_p(y,\cdot)\,dy\right) = 0$ in (14) is a *global* condition. It indeed equivalently reads, because of the ergodic Theorem, a.s. $-\lim_{R\longrightarrow+\infty} \frac{1}{|B_R|}\int_{B_R} \nabla w_p(y,\cdot)\,dy = 0$ for any sequence of balls B_R of radii R. The fact that the random corrector problem is posed on the entire space has far reaching consequences for numerical practice. Truncations of problem (14) have to be considered, and the actual homogenized coefficients are only correct in the asymptotic regime. The present series of works is somehow motivated by the above observation, as already pointed out in the introduction.

Remark 2. In fact, the situation considered here is simple: it is the linear elliptic case. It is well known that, even in the periodic setting, the difficulties we mention for the random setting already arise in the periodic setting when the operator is, for instance, nonlinear. Then determining the periodic homogenized problem cannot always be reduced to a simple computation on one single periodic cell of the problem.

2.3 A Variant

A specific stochastic setting has been introduced and studied in [5, 7]. It is *not* a particular case of the classical stationary settings defined above. As briefly mentioned in the introduction, it is motivated by the consideration of random geometries (we mean, materials) that have some relation to the periodic setting. Here, the periodic setting is taken as a *reference* configuration, somewhat similarly to the classical mathematical formalization of continuum mechanics where a reference configuration is used to define the state of the material under study. Another related idea, in a completely different context, is the consideration of a reference element for finite element computations. In all cases, the real situation is seen *via* a *mapping* from the reference configuration to the actual configuration.

We fix some $\mathbb{Z}^d$-periodic, square matrix A_{per} of size d, assumed to satisfy

$$\exists \gamma > 0 / \forall \xi \in \mathbb{R}^d,\ \ \xi^T A_{per}(y)\xi \geq \gamma|\xi|^2,\ \ \text{almost everywhere in } y \in \mathbb{R}^d, \quad (15)$$

$$\forall i, j \in \{1, 2, \ldots, d\}, \quad [A_{per}]_{ij} \in L^\infty(\mathbb{R}^d). \quad (16)$$

We consider the following problem:

$$\begin{cases} -\mathrm{div}\left(A_{per}\left(\Phi^{-1}\left(\frac{x}{\varepsilon},\omega\right)\right)\nabla u^\varepsilon\right) = f \quad \text{in} \quad \mathscr{D}, \\ u^\varepsilon = 0 \quad \text{on} \quad \partial\mathscr{D}, \end{cases} \tag{17}$$

where the function $\Phi(\cdot,\omega)$ is assumed to be a diffeomorphism from $\mathbb{R}^d$ to $\mathbb{R}^d$ for $\mathbb{P}$-almost every ω. The diffeomorphism is assumed to additionally satisfy

$$\operatorname*{EssInf}_{\omega\in\Omega,\, x\in\mathbb{R}^d} [\det(\nabla\Phi(x,\omega))] = \nu > 0, \tag{18}$$

$$\operatorname*{EssSup}_{\omega\in\Omega,\, x\in\mathbb{R}^d} (|\nabla\Phi(x,\omega)|) = M < \infty, \tag{19}$$

$$\nabla\Phi(x,\omega) \quad \text{is stationary in the sense of (7).} \tag{20}$$

Such a Φ is called a *random stationary diffeomorphism*.

The following result is proved in [5,7]:

Theorem 2. *Let $\mathscr{D}$ be a bounded smooth open subset of $\mathbb{R}^d$, and let $f \in H^{-1}(\mathscr{D})$. Let A_{per} be a square matrix which is $\mathbb{Z}^d$-periodic and satisfies (15) and (16). Let Φ be a random stationary diffeomorphism satisfying hypotheses (18–20). Then the solution $u^\varepsilon(x,\omega)$ to (17) satisfies the following properties:*

1. *$u^\varepsilon(x,\omega)$ converges to some $u_0(x)$ strongly in $L^2(\mathscr{D})$ and weakly in $H^1(\mathscr{D})$, almost surely;*
2. *the function u_0 is the solution to the homogenized problem:*

$$\begin{cases} -\mathrm{div}(A_*\nabla u_0) = f \quad \textit{in} \quad \mathscr{D}, \\ u_0 = 0 \quad \textit{on} \quad \partial\mathscr{D}. \end{cases} \tag{21}$$

In (21), the homogenized matrix A_* is defined by:

$$[A_*]_{ij} = \det\left(\mathbb{E}\left(\int_Q \nabla\Phi(z,\cdot)dz\right)\right)^{-1} \times \mathbb{E}\left(\int_{\Phi(Q,\cdot)} (e_i + \nabla w_{e_i}(y,\cdot))^T A_{per}(\Phi^{-1}(y,\cdot))e_j\, dy\right), \tag{22}$$

where for any $p \in \mathbb{R}^d$, w_p is the solution (unique up to the addition of a (random) constant) in $\left\{w \in L^2_{\mathrm{loc}}(\mathbb{R}^d, L^2(\Omega)), \nabla w \in L^2_{\mathrm{unif}}(\mathbb{R}^d, L^2(\Omega))\right\}$ to

$$\begin{cases} -\mathrm{div}\left[A_{per}(\Phi^{-1}(y,\omega))(p + \nabla w_p)\right] = 0, \\ w_p(y,\omega) = \tilde{w}_p(\Phi^{-1}(y,\omega),\omega), \quad \nabla\tilde{w}_p \quad \textit{is stationary in the sense of (7)}, \\ \mathbb{E}\left(\int_{\Phi(Q,\cdot)} \nabla w_p(y,\cdot)dy\right) = 0. \end{cases} \tag{23}$$

3 Numerical Approaches for an Approximation at First Order

3.1 Small Perturbations of the Periodic Setting

It has been shown in [7] that, when Φ in (17) is a perturbation of the Identity map

$$\Phi(x,\omega) = x + \eta\,\Psi(x,\omega) + O(\eta^2), \tag{24}$$

the solution to the corrector problem (23) may be developed in powers of the small parameter η. It reads $\widetilde{w}_p(x,\omega) = w_p^0(x) + \eta w_p^1(x,\omega) + O(\eta^2)$, where w_p^0 solves

$$-\operatorname{div}\left[A_{per}\,(p+\nabla w_p^0)\right] = 0, \quad w_p^0 \text{ is } Q\text{-periodic}, \tag{25}$$

and where w_p^1 solves

$$\begin{cases} -\operatorname{div}\left[A_{per}\nabla w_p^1\right] = \operatorname{div}\left[-A_{per}\nabla\Psi\nabla w_p^0 - (\nabla\Psi^T - (\operatorname{div}\Psi)\mathrm{Id})\,A_{per}\,(p+\nabla w_p^0)\right], \\ \nabla w_p^1 \text{ is stationary and } \mathbb{E}\left(\displaystyle\int_Q \nabla w_p^1\right) = 0. \end{cases} \tag{26}$$

The problem (26) in w_p^1 is random in nature, but it is in fact easy to see, taking the expectation, that $\overline{w}_p^1 = \mathbb{E}(w_p^1)$ is Q-periodic and solves the *deterministic* problem

$$\begin{aligned} &-\operatorname{div}\left[A_{per}\,\nabla\overline{w}_p^1\right] \\ &\quad = \operatorname{div}\left[-A_{per}\,\mathbb{E}(\nabla\Psi)\,\nabla w_p^0 - (\mathbb{E}(\nabla\Psi^T) - \mathbb{E}(\operatorname{div}\Psi)\mathrm{Id})\,A_{per}\,(p+\nabla w_p^0)\right]. \end{aligned} \tag{27}$$

This is useful because, on the other hand, the knowledge of w_p^0 and $\overline{w}_p^1$ suffices to obtain a first order expansion (in η) of the homogenized matrix. Define $A_{ij}^0 = \int_Q (e_i + \nabla w_{e_i}^0)^T A_{per}\,e_j$ and

$$\begin{aligned} A_{ij}^1 = &-\int_Q \mathbb{E}(\operatorname{div}\Psi)\,A_{ij}^0 + \int_Q (e_i+\nabla w_{e_i}^0)^T A_{per}\,e_j\,\mathbb{E}(\operatorname{div}\Psi) \\ &+ \int_Q \left(\nabla\overline{w}_{e_i}^1 - \mathbb{E}(\nabla\Psi)\nabla w_{e_i}^0\right)^T A_{per}\,e_j, \end{aligned}$$

we then have

$$A_* = A^0 + \eta A^1 + O(\eta^2). \tag{28}$$

As subsequently shown in [10], a similar approach can be applied to the corrector problems once *discretized* by a finite element approach. Given a mesh $\mathscr{T}_h^{(Q)}$ of Q of size h, reproduced by periodicity on $Q_N = [0,N]^d$, we define the discrete variational formulation

$$\left\{\begin{array}{l}\text{Find } \widetilde{w}_p^{h,N}(\cdot,\omega) \in V_h^{\rm per}(Q_N) \quad \text{such that, for all } \widetilde{v}_h \in V_h^{\rm per}(Q_N),\\ \displaystyle\int_{Q_N} \det(\nabla\Phi)(\nabla\widetilde{v}_h)^T(\nabla\Phi)^{-T} A_{per}\big(p + (\nabla\Phi)^{-1}\nabla\widetilde{w}_p^{h,N}(\cdot,\omega)\big) = 0 \\ \qquad\qquad \text{almost surely,}\end{array}\right. \tag{29}$$

where $V_h^{\rm per}(Q_N)$ is the set of Q_N-periodic functions that have their restriction to Q_N in a typical finite element space built from the mesh $\mathscr{T}_h^N$ (obtained by periodization). Note that the problem is formulated in terms of $\widetilde{w}_p$ (rather than w_p) because the gradient of $\widetilde{w}_p$ is stationary. The matrix

$$\begin{aligned}&\big[A_*{}^{h,N}\big]_{ij}(\omega)\\ &= \det\left(\frac{1}{|Q_N|}\int_{Q_N}\nabla\Phi\right)^{-1}\frac{1}{|Q_N|}\int_{Q_N}\det(\nabla\Phi)\big(e_i + (\nabla\Phi)^{-1}\nabla\widetilde{w}_{e_i}^{h,N}\big)^T A_{per}e_j\end{aligned} \tag{30}$$

is then considered. Using the same expansion (24) as in the above "continuous" case, a formal expansion $\widetilde{w}_p^{h,N} = w_p^{0,h,N} + \eta w_p^{1,h,N} + O(\eta^2)$ of the discrete corrector is performed and inserted in (29). The function $w_p^{0,h,N}$ is then shown to be independent of N (it is henceforth denoted $w_p^{0,h}$), while $w_p^{0,h}$ and $w_p^{1,h,N}$ are respectively solutions to

$$\begin{array}{l}\text{Find } w_p^{0,h} \in V_h^{\rm per}(Q) \text{ such that, for all } v_h \in V_h^{\rm per}(Q),\\ \displaystyle\int_Q (\nabla v_h)^T A_{per}\,(p + \nabla w_p^{0,h}) = 0,\end{array} \tag{31}$$

and

$$\left\{\begin{array}{l}\text{Find } w_p^{1,h,N}(\cdot,\omega) \in V_h^{\rm per}(Q_N) \text{ such that, for all } v_h \in V_h^{\rm per}(Q_N), \quad \text{and almost surely,}\\ \displaystyle\int_{Q_N} (\nabla v_h)^T A_{per}\,\nabla w_p^{1,h,N}\\ \qquad = \int_{Q_N}(\nabla v_h)^T\Big[A_{per}\nabla\Psi\nabla w_p^{0,h} + (\nabla\Psi^T - (\mathrm{div}\Psi)\mathrm{Id})A_{per}(p + \nabla w_p^{0,h})\Big].\end{array}\right. \tag{32}$$

Equations (31) and (32) are of course discretized formulations of (25) and (26), respectively. Similarly to what has been proven in the continuous setting in [7] (and briefly recalled above), it is possible to show that there exists a constant $C(h,N,\omega)$ such that, for η sufficiently small,

$$\eta^{-2}\left\|\nabla\widetilde{w}_p^{h,N}(\cdot,\omega) - \nabla w_p^{0,h} - \eta\nabla w_p^{1,h,N}(\cdot,\omega)\right\|_{L^2(Q_N)} \le |Q_N|^{1/2}\,C(h,N,\omega), \tag{33}$$

and

$$\eta^{-2}|A_*{}^{h,N}(\omega) - A^{0,h} - \eta A^{1,h,N}(\omega)| \le C(h,N,\omega), \tag{34}$$

where ${A_*}^{h,N}$ is defined by (30), $(A^{0,h})_{ij} = \int_Q \left(e_i + \nabla w_{e_i}^{0,h}\right)^T A_{per}\, e_j$ and

$$(A^{1,h,N})_{ij} = -(A^{0,h})_{ij} \frac{1}{|Q_N|} \int_{Q_N} \operatorname{div} \Psi + \frac{1}{|Q_N|} \int_{Q_N} \left(e_i + \nabla w_{e_i}^{0,h}\right)^T A_{per} e_j \operatorname{div} \Psi + \frac{1}{|Q_N|} \int_{Q_N} \left(\nabla w_{e_i}^{1,h,N} - \nabla \Psi\, \nabla w_{e_i}^{0,h}\right)^T A_{per}\, e_j.$$

Again as in the continuous setting, knowing only the expectation $\overline{w}_p^{1,h,N} = \mathbb{E}(w_p^{1,h,N})$ which solves, for all $v_h \in V_h^{\mathrm{per}}(Q_N)$,

$$\int_{Q_N} (\nabla v_h)^T A_{per}\, \nabla \overline{w}_p^{1,h,N} = \int_{Q_N} (\nabla v_h)^T \big[A_{per}\, \mathbb{E}(\nabla \Psi)\, \nabla w_p^{0,h} + \big(\mathbb{E}(\nabla \Psi)^T - \mathbb{E}(\operatorname{div} \Psi)\mathrm{Id}\big) A_{per} \left(p + \nabla w_p^{0,h}\right)\big] \tag{35}$$

is sufficient to determine the first order correction to the homogenized matrix. A simple argument shows that $\overline{w}_p^{1,h,N}$ is independent from N (it is henceforth denoted by $\overline{w}_p^{1,h}$), Q-periodic, and solution to (35) with $N = 1$, which is a converging discretization of (27) when h vanishes. The matrix $A^{1,h} = \mathbb{E}(A^{1,h,N})$ is similarly independent of N, and can be computed only using $\nabla \overline{w}_p^{1,h}$.

The question arises to know how large the (random) constant $C(h, N, \omega)$ in (34) is. Too large a constant would indeed mean that the first order expansion in η, although appealing theoretically, is useless practically to get an accurate approximation of the homogenized matrix. This is the purpose of [10] to examine this issue in a simple testcase, representative of some generality.

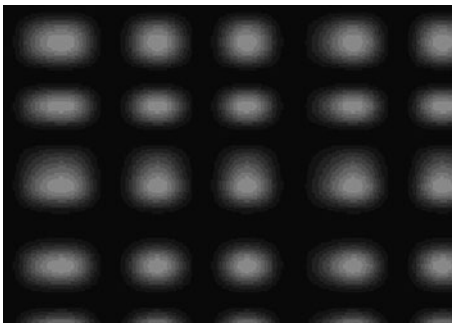
We work in dimension 2, with coordinates $x = (x_1, x_2)$, and consider two families $(X_k)_{k\in\mathbb{Z}}$ and $(Y_k)_{k\in\mathbb{Z}}$ of scalar, identically distributed, independent random variables. Their common law is the uniform law $\mathscr{U}([a, b])$ on the range $[a, b]$. We choose the diffeomorphism $\Phi(x) = x + \eta\Psi(x, \omega)$, with $\Psi(x, \omega) = (\psi_X(x_1, \omega), \psi_Y(x_2, \omega))$, where ψ_X is defined by

$$\psi_X(x_1, \omega) = \sum_{k\in\mathbb{Z}} 1_{[k,k+1[}(x_1) \left(\sum_{q=0}^{k-1} X_q(\omega) + 2X_k(\omega) \int_k^{x_1} \sin^2(2\pi t)\, dt \right),$$

and ψ_Y is defined similarly. The periodic matrix A_{per} is defined by

$$\forall x \in Q, A_{per}(x) = a_{\mathrm{per}}(x)\mathrm{Id}_2, \quad a_{\mathrm{per}}(x_1, x_2) = \beta + (\alpha - \beta) \sin^2(\pi x_1) \sin^2(\pi x_2).$$

The idea is to consider a $\mathbb{Z}^2$-periodic material, where thermal conductivity (modeled by the matrix $A_{per} \circ \Phi^{-1}$) smoothly varies from α to $\beta \le \alpha$. Conductivity is maximum at the center of the cell Q, and minimum on its boundary. Note that the map ψ_X is not stationary, but its gradient is. This is a prototypical example of



η	$(A_*{}^{h,N})_{11}$	$(e^{h,N})_{11}$
0.1	3.073 ± 0.00928	-4.233 ± 0.216
0.01	2.839 ± 0.00111	-5.009 ± 0.254
0.001	2.812 ± 0.000113	-5.104 ± 0.259
0.0001	2.809 ± 0.0000113	-5.114 ± 0.259

Fig. 2 *Left*: value of $A_{per} \circ \Phi^{-1}(x, \omega)$ for a particular random realization on the domain $Q_{N=5}$ ($\eta = 0.05$). This intuitively models a periodic structure (disks centered on a periodic lattice) slightly perturbed by a random diffeomorphism close to Identity. *Right*: values of $(A_*^{h,N})_{11}$ and $(e^{h,N})_{11}$ in function of η. All data are extracted from [10]

the setting developed in [7], which is not covered by *classical* stochastic homogenization theory since $A_{per} \circ \Phi^{-1}$ is not stationary. As shown by Fig. 2 (left) where $A_{per} \circ \Phi^{-1}(x, \omega)$ is displayed for a particular realization of the randomness, this is however a quite intuitive setting which deserves specific attention. The specific values chosen for the parameters are: $a = -2.25$, $b = 5.75$, $\alpha = 10$, and $\beta = 1$, $h = 1/3$, $N = 20$. The number of realizations is 10. The numerical results are obtained using the finite element software FreeFem++. They are displayed on the table of Fig. 2 (right). The left column shows the result obtained for the $(1, 1)$ entry of the homogenized matrix, with the interval of confidency. The right column gives the value of the error estimator

$$e^{h,N}(\omega) := \eta^{-2}\big(A_*{}^{h,N}(\omega) - A^{0,h} - \eta A^{1,h,N}(\omega)\big),$$

again for the $(1, 1)$ entry. The values found for other entries of the homogenized matrix lead to similar conclusions. Note that, for the purpose of analysis and with a view to reducing variance (see the details in [10]), we have used the random value $A^{0,h} + \eta A^{1,h,N}(\omega)$ in the right hand side of the estimator. In practice, $A^{0,h} + \eta A^{1,h}$ would be used, instead of $A^{0,h} + \eta A^{1,h,N}(\omega)$, as an approximation for $A_*{}^{h,N}$.

The conclusion is that the constant $C(h, N, \omega)$ is small (say of the order of 5 in this particular case) and that the first order approximation $A^{0,h} + \eta A^{1,h}$ of the homogenized matrix $A_*{}^{h,N}$ is thus a practically accurate numerical approach (provided the first order precision is judged satisfactory for the application considered). In terms of computational efficiency, the gain is enormous. Solving the couple of *periodic* problems (31) and (35) to respectively get $w_p^{0,h}$ and $\overline{w}_p^{1,h}$ is much less expensive than solving the original *stochastic* corrector problem (29).

3.2 Rare but Possibly Large Perturbations

We now consider a slightly different perturbative approach. It could be presented in the setting of random diffeormophisms introduced in Sect. 2.3 above, but for clarity we present it in the more classical setting of Sect. 2.2.

As above, we consider our random material as a small perturbation of a periodic material. The matrix that models its response is thus expanded as

$$A_\eta(x,\omega) = A_{per}(x) + b_\eta(x,\omega)C_{per}(x), \tag{36}$$

where, with evident notation, A_{per} is a periodic matrix modeling the unperturbed material, and where C_{per} is a periodic matrix modeling the perturbation. The amplitude of the perturbation, which used to be modeled by a *deterministic* coefficient η in the previous section, is now a scalar *random* field $b_\eta(x,\omega)$. We assume that this field satisfies

$$\|b_\eta\|_{L^\infty(Q;L^p(\Omega))} \underset{\eta\to 0}{\to} 0, \tag{37}$$

for some $1 \leq p < \infty$. For well-posedness of the problem, we also assume there exists $0 < \alpha \leq \beta$ such that for almost all $x \in \mathbb{R}^d$ and for almost all $\omega \in \Omega$,

$$\forall \xi \in \mathbb{R}^d,\ \forall \eta > 0, \quad \alpha|\xi|^2 \leq A_\eta(x,\omega)\xi\cdot\xi \quad and \quad |A_\eta(x,\omega)\xi| \leq \beta|\xi|.$$

Condition (37) states that the perturbation in (36) is small *on average*. However, it does not prevent the perturbation to be large, once in a while, because we only have $p < \infty$ (Note that the setting of the previous section corresponds to a situation where $p = \infty$). Whereas the idea underlying the setting of the previous section was *perturb the periodic material possibly often but only slightly*, the intuitive image behind the present setting is *perturb the periodic material only rarely, but then possibly largely*. The comparison of Fig. 2 (left) and Fig. 3 is self explanatory.

When the exponent p in (37) is strictly larger than one, a theory similar to that of the previous section can be developed. Assuming that $m_\eta := \|b_\eta\|_{L^\infty(Q;L^p(\Omega))} \longrightarrow 0$ as η vanishes, it may be proved, up to the extraction of a subsequence, that the homogenized tensor $A_{\eta,*}$ admits a first order expansion in terms of the small "coefficient" m_η. The coefficients are easily expressed using periodic corrector problems built from the matrices A_{per} and C_{per}. The remainder in the expansion can indeed be shown to be $o(m_\eta)$ in a certain sense and under appropriate assumptions. We refer to [2, 3] for the details. There are some cases when the expansion in fact does not converge. We now address such a case, very different in nature.

Consider the prototypical case where b_η is uniform in each cell of $\mathbb{Z}^d$ and writes

$$b_\eta(x,\omega) = \sum_{k\in\mathbb{Z}^d} \mathbf{1}_{\{Q+k\}}(x) B^k_\eta(\omega), \tag{38}$$

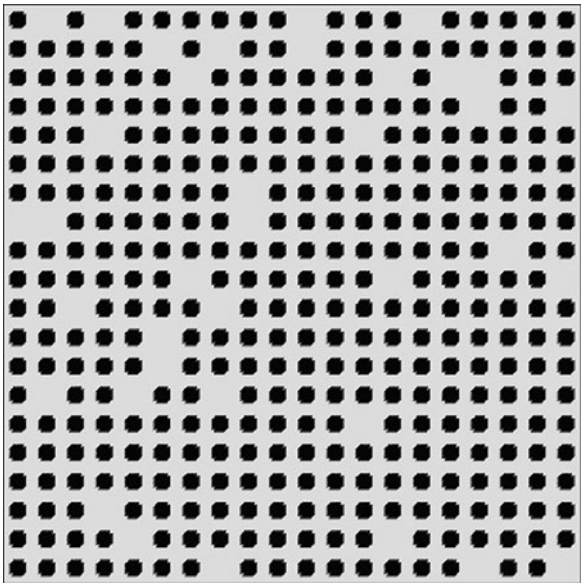


Fig. 3 A typical random realization of the Bernoulli law for the perturbed periodic material

where the B_η^k are independent identically distributed random variables. Their common law is assumed to be a Bernoulli law of parameter η. This setting satisfies condition (37) for all $p \geq 1$. The difficulty with a possible expansion in "powers" of b_η is intuitively that, a Bernoulli variable B, being valued in $\{0, 1\}$, is such that $B^p = B$ for all p. So all terms in the expansion are potentially of the same order. A different strategy is needed. We now explain an alternative, *formal* approach, for which we do not know any rigorous foundation to date. Although definite conclusions on the validity of the approach have yet to be obtained, the numerical tests we performed show its practical correctness and efficiency.

Heuristically, on the cube $Q_N = [0, N]^d$ and at order 1 in η, the probability to get the perfect periodic material (entirely modeled by the matrix A_{per}) is $(1-\eta)^{N^d} \approx 1 - N^d\eta + O(\eta^2)$, while the probability to obtain the unperturbed material on all cells except one (where the material has matrix $A_{per} + C_{per}$) is $N^d\,(1-\eta)^{N^d-1}\eta \approx N^d\eta + O(\eta^2)$. All other configurations, with more than two cells perturbed, yield contributions of orders higher than or equal to η^2. This gives the intuition that the first order correction indeed comes from the difference between the material perfectly periodic except on one cell and the perfect material itself. It is therefore claimed in [2, 3] that $A_{\eta,*} = A_{per,*} + \eta A_{1,*} + o(\eta)$ where $A_{per,*}$ is the homogenized matrix for the unperturbed periodic material and

$$A_{1,*}\, e_i = \lim_{N\to+\infty} \int_{Q_N} \Big[(A_{per} + \mathbf{1}_Q C_{per})\big(\nabla w_i^N + e_i\big) - A_{per}\,\big(\nabla w_i^0 + e_i\big)\Big], \tag{39}$$

where w_i^0 is the corrector for A_{per}, and w_i^N solves

$$\begin{aligned} &-\operatorname{div}\left(\left(A_{per}(x)+\mathbf{1}_Q C_{per}(x)\right)\left(\nabla w_i^N(x)+e_i\right)\right)=0 \\ &\quad \text{in} \quad Q_N, \quad w_i^N \; Q_N-\text{periodic}. \end{aligned} \tag{40}$$

Note that the integral appearing in the right-hand side of (39) is *not* normalized: it a priori scales as the volume N^d of Q_N and has finite limit only because of cancellation effects between the two terms in the integrand. This is very similar in nature to the modeling of *defects* in Statistical Physics: a flawless (periodic) environment is substracted to the actual environment and acts as a normalization.

There actually exists a formal generalization of (39) that allows for recovering the setting of the previous cases. The approach of the present section therefore appears to be the most general approach to the modeling of "small" random perturbations. We again refer to [2, 3] for more details.

The approach has been tested in [3]. The matrix A_{per} is taken scalar. In each periodic cell, it has constant value 1,020 in the central circular inclusion and constant value 20 in the surrounding region. The matrix C_{per} has value $-1{,}000$ in the inclusions and 0 outside. The coefficient b_η is of the form (38), with B_η a Bernoulli variable with parameter $\eta = 0.1$. The results are shown on Fig. 4 below. On the

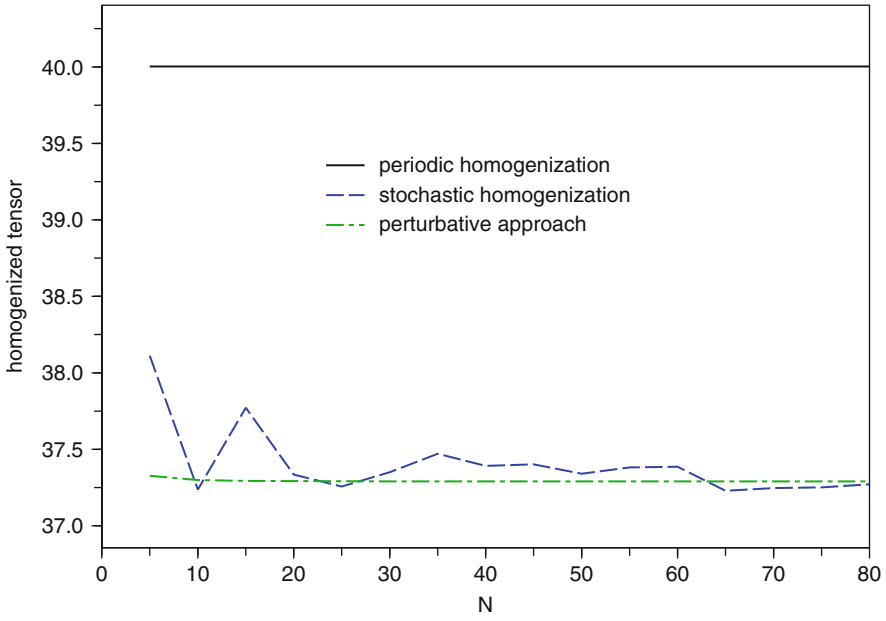


Fig. 4 Comparison of the actual random coefficient, which converges to the homogenized coefficient in the limit of large cube sizes N, (curve labelled "stochastic homogenization") with the unperturbed periodic homogenized coefficient (curve labelled "periodic homogenization") and the first order expansion (curve labelled "perturbative approach"). The asymptotic limit is almost instantaneously found by the perturbative approach

cube $Q_N = [0, N]^2$ with N increasingly large, an approximation of $A_{\eta,*}$ is directly computed. Alternatively, expression (39) is employed to calculate the first order term $A_{1,*}$ of the expansion. The values $A_{\eta,*}$ and $A_{per,*} + \eta A_{1,*}$ are then compared to one another. The process is completed for several realizations of the random material. Only a particular realization is shown on Fig. 4 but all realizations yield qualitatively similar behaviours. It is observed that, using the perturbative approach, the large N limit for cubes of size N is already very well approached for small values of N. As in the previous section, the computational efficiency of the approach is clear: solving the two periodic problems with coefficients A_{per} and $A_{per} + \mathbf{1}_Q C_{per}$ for a limited size N is much less expensive than solving the original, random corrector problem for a much larger size N.

4 Related Problems and Techniques

We conclude this article with some comments.

First, it is useful to mention that the variant of stochastic homogenization described in Sect. 2.3 has originally been introduced in [6, 7] for an apparently different context, related to atomistic modeling of materials and the limit of atomistic models to derive models for continuum mechanics. Although the two topics of Atomistic to Continuum limits and homogenization of partial differential equations look different at first sight, they actually share similarities, as two sides of the general paradigm of change of scales.

Second, the set of techniques presented above is specific to the case of periodic settings slightly perturbed by random perturbations. Although we believe this allows to treat many situations, the situation where randomness is intense is still, of course, of major interest. In that case, there seems to be no hope of simplifying the problem. A corrector problem of the type (14) (or of the type (23) when random diffeomorphisms are employed), posed on the entire ambient space, needs to be solved, for each vector p of the canonical basis. And, the average giving the homogenized matrix needs then to be computed. As in any situation where randomness is present, numerical practice shows that variance issues come into the picture and complicate the already huge computational task. A companion article [11] presents some techniques recently introduced to improve the efficiency of computations of homogenization problems that require the solution of corrector problems posed on the entire space.

Acknowledgements The work of the author is partially supported by ONR under contract Grant 00014-09-1-0470.

References

1. G. Allaire, *Homogenization and two-scale convergence*, SIAM J. Math. Anal., 23 (6), pp 1482–1518, 1992
2. A. Anantharaman, Thèse de l' Université Paris Est, Ecole des Ponts
3. A. Anantharaman, C. Le Bris, *Homogénéisation d'un matériau périodique faiblement perturbé aléatoirement*, [Homogenization of a weakly randomly perturbed periodic material], C. R., Math., Acad. Sci. Paris 348, No. 9–10, 529–534, 2010
4. A. Bensoussan, J. L. Lions, G. Papanicolaou, **Asymptotic analysis for periodic structures**, Studies in Mathematics and its Applications, 5. North-Holland, Amsterdam-New York, 1978
5. X. Blanc, C. Le Bris, P.-L. Lions, *Une variante de la théorie de l'homogénéisation stochastique des opérateurs elliptiques [A variant of stochastic homogenization theory for elliptic operators]*, C. R. Acad. Sci. Série I, 343, pp 717–724, 2006
6. X. Blanc, C. Le Bris, P.-L. Lions, *The energy of some microscopic stochastic lattices*, Arch. Rat. Mech. Anal., 184, pp 303–339, 2007
7. X. Blanc, C. Le Bris, P.-L. Lions, *Stochastic homogenization and random lattices*, Journal de Mathématiques Pures et Appliquées, 88, pp 34–63, 2007
8. D. Cioranescu, P. Donato, **An introduction to homogenization.** Oxford Lecture Series in Mathematics and its Applications, 17. The Clarendon Press, Oxford University Press, New York, 1999
9. R. Costaouec, Thèse de l' Université Paris Est, Ecole des Ponts
10. R. Costaouec, C. Le Bris, F. Legoll, *Approximation numérique d'une classe de problèmes en homogénéisation stochastique*, [Numerical approximation of a class of problems in stochastic homogenization], C. R. Acad. Sci. Série I, 348, pp 99–103, 2010
11. R. Costaouec, C. Le Bris, F. Legoll, *Variance reduction in stochastic homogenization: proof of concept, using antithetic variables*, Bol. Soc. Esp. Mat. Apl., 50, pp 9–27, 2010
12. V. V. Jikov, S. M. Kozlov, O. A. Oleinik, **Homogenization of differential operators and integral functionals**. Springer, Berlin, 1994
13. U. Krengel, **Ergodic theorems**, de Gruyter Studies in Mathematics, vol. 6, de Gruyter, 1985
14. F. Murat, *Compacité par compensation*, Ann. Scuola Norm. Sup. Pisa. Cl. Sci. 5 (4) pp 485–507, 1978
15. G. Nguetseng, *A general convergence result for a functional related to the theory of homogenization*, SIAM J. Math. Anal. 20 (3), pp 608–623, 1989
16. A. N. Shiryaev, **Probability**, Graduate Texts in Mathematics, vol. 95, Springer, Berlin, 1984
17. L. Tartar, *Compensated compactness and applications to partial differential equations,* Nonlinear analysis and mechanics: Heriot–Watt Symposium, Vol. IV, pp. 136–212, Res. Notes in Math., 39, Pitman, MA, 1979
18. M. Thomas, *Propriétés thermiques de matériaux composites : caractérisation expérimentale et approche microstructurale*, [Thermal properties of composite materials: experimental characterization and microstructural approach], Thèse de l' Université de Nantes, Laboratoire de Thermocinétique, CNRS - UMR 6607, 2008

Goal Oriented, Anisotropic, A Posteriori Error Estimates for the Laplace Equation

Frederic Alauzet, Wissam Hassan, and Marco Picasso

Abstract A posteriori error estimates are presented for the Laplace equation and meshes with large aspect ratio. Error estimates are presented in the natural H^1 seminorm or in the framework of goal oriented error control. The proposed estimator relies on anisotropic interpolation estimates derived by Formaggia and Perotto [Numer. Math. 89(4), 641–667 (2001), Numer. Math. 94(1), 67–92 (2003)] and on Zienckiewicz–Zhu [Int. J. Numer. Meth. Eng. 33(7), 1331–1364 (1992), Int. J. Numer. Meth. Eng. 24(2), 337–357 (1987)] post-processing techniques, thus avoids approximations of the Hessian of the solution. All the constant involved in the error estimates are independent of the mesh size and aspect ratio, which should enable the use of anisotropic, adaptive finite elements.

1 Introduction

A posteriori error estimates aim to link the error between the true solution u and the finite element approximation u_h with a computable quantity – the so-called error estimator η. Then, the error estimator can be used as a refining – or coarsening – criteria in adaptive finite element algorithms. The subject was initiated by Babuska and Rheinboldt [3] and mesh adaptation is nowadays a classical feature in finite element software, see for instance [5, 16, 31].

F. Alauzet
INRIA Rocquencourt, Projet Gamma, B.P. 105, 78153 Le Chesnay Cedex, France
e-mail: frederic.alauzet@inria.fr

W. Hassan
IACS, Station 8, Ecole Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland
e-mail: wissam.hassan@epfl.ch supported by Dassault Aviation

M. Picasso (✉)
IACS, Station 8, Ecole Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland
and INRIA Rocquencourt, Projet Gamma, B.P. 105, 78153 Le Chesnay Cedex, France
e-mail: marco.picasso@epfl.ch

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_4,

Interpolation estimates are usually needed in order to derive a posteriori error estimates. In the simplest case, namely continuous, piecewise linear finite elements on triangles, Lagrange interpolation [13] is needed to prove a priori error estimates whereas Clément interpolation [14] is needed to derive a posteriori error estimates. In the classical setting of isotropic meshes, both interpolation estimates hold under the so-called regularity assumption which requires that

$$\exists C > 0 \qquad \forall h > 0 \qquad \forall K \in \mathscr{T}_h \qquad \frac{h_K}{\rho_K} \leq C.$$

Hereabove, $\mathscr{T}_h$ denotes a mesh of the calculation domain Ω into triangles K with diameter h_K less than h and ρ_K is the largest circle contained in K. However, in practice, anisotropic finite elements are used with success in order to solve complex problems such as fluid flow around bodies, see for instance [2, 9, 21, 22]. Recently, the theory of finite elements was updated in order to comply with the use of anisotropic finite elements. Hereafter, we will consider the contributions of Formaggia and Perotto [19, 20], however similar results have been obtained for Lagrange interpolation in [8, 12, 22] and for Clément interpolation [23].

In the classical setting of isotropic meshes satisfying the regularity assumption, the interpolation estimate for the Lagrange interpolation operator r_h with polynomial degree one write [13]:

$$\exists C > 0 \qquad \forall h > 0 \qquad \forall K \in \mathscr{T}_h \qquad \forall v \in H^2(K)$$
$$\int_K |\nabla(v - r_h v)|^2 \leq C \frac{h_K^4}{\rho_K^2} \int_K \left(\left(\frac{\partial^2 v}{\partial x_1^2} \right)^2 + \left(\frac{\partial^2 v}{\partial x_1 \partial x_2} \right)^2 + \left(\frac{\partial^2 v}{\partial x_2^2} \right)^2 \right),$$

where C does not depend on the mesh aspect ratio. If, for instance, v depends only on x_2, then the right hand side of the above estimate blows up when the mesh is refined only in the x_2 direction, since the mesh aspect ratio h_K/ρ_K increases. On the other side, anisotropic interpolation estimates [19, 20] are as follows:

$$\int_K |\nabla(v - r_h v)|^2 \leq C \left(\frac{\lambda_{1,K}^4}{\lambda_{2,K}^2} \int_K (\mathbf{r}_{1,K}^T H(v) \mathbf{r}_{1,K})^2 \right.$$
$$\left. + 2\lambda_{1,K}^2 \int_K (\mathbf{r}_{1,K}^T H(v) \mathbf{r}_{2,K})^2 + \lambda_{2,K}^2 \int_K (\mathbf{r}_{2,K}^T H(v) \mathbf{r}_{2,K})^2 \right).$$

Hereabove $H(v)$ is the Hessian matrix

$$H(v) = \begin{pmatrix} \dfrac{\partial^2 v}{\partial x_1^2} & \dfrac{\partial^2 v}{\partial x_1 \partial x_2} \\ \dfrac{\partial^2 v}{\partial x_1 \partial x_2} & \dfrac{\partial^2 v}{\partial x_2^2} \end{pmatrix},$$

$\mathbf{r}_{1,K}$ and $\mathbf{r}_{2,K}$ denote orthogonal unit vectors in the direction of maximum and minimum streching, respectively, while $\lambda_{1,K}$ and $\lambda_{2,K}$ denote the stretching amplitudes in the direction of maximum and minimum streching, respectively. The precise definition of these quantities is proposed in the next section.

If, for instance, v depends only on x_2 and if the mesh is refined only in the x_2 direction, then $\mathbf{r}_{1,K} = (1\ 0)^T$ and $\mathbf{r}_{2,K} = (0\ 1)^T$ and the above anisotropic interpolation estimate reduces to

$$\int_K |\nabla(v - r_h v)|^2 \leq C \lambda_{2,K}^2 \int_K \left(\frac{\partial^2 v}{\partial x_2^2} \right)^2$$

so that convergence is achieved as soon as $\lambda_{2,K}$ goes to zero, no matter what the aspect ratio $\lambda_{1,K}/\lambda_{2,K}$ is. The same argument holds when the isovalues of v and the mesh are rotated with any angle.

Based on anisotropic interpolation estimates, a priori and a posteriori error estimates have been revisited and adaptive algorithms having meshes with large aspect ratio have been used with success in CFD, see for instance [2, 9, 15, 21, 22, 25]. Most of the paper involving anisotropic adaptive meshes deal with an estimate of the Hessian matrix, thus approaching the second derivatives of the exact (unknown) solution using the computed solution. In [27] an anisotropic error estimator involving the gradient matrix rather than the Hessian matrix was proposed for elliptic and parabolic problems in the energy norm. A lower bound was proved in [26, 28] for the Laplace problem. Goal oriented, anisotropic a posteriori error estimates involving gradients were proposed in [15] for advection-diffusion-reaction, but only an upper bound was proved. In this paper we prove a lower bound for goal oriented, anisotropic a posteriori error estimates in the frame of the Laplace equation.

2 The Laplace Equation with Anisotropic Finite Elements

Given a polygonal domain $\Omega \subset \mathbb{R}^2$, given $f \in L^2(\Omega)$, we are searching for $u : \Omega \to \mathbb{R}$ such that

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

For any $0 < h < 1$, let $\mathcal{T}_h$ be a conforming triangulation of $\overline{\Omega}$ into triangles K with diameter h_K less than h. Let V_h be the usual finite element space of continuous, piecewise linear functions on the triangles of $\mathcal{T}_h$, zero valued on $\partial\Omega$. The simplest finite element approximation of (1) therefore consists in seeking $u_h \in V_h$ such that

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

We now describe the mesh anisotropy using the framework of Formaggia and Perotto [19,20]. Again, alternative descriptions are available [8,12,22,23]. For any triangle K of the mesh, let $T_K : \hat{K} \to K$ be the affine transformation which maps the reference triangle $\hat{K}$ into K. Let M_K be the Jacobian of T_K that is

$$\mathbf{x} = T_K(\hat{\mathbf{x}}) = M_K \hat{\mathbf{x}} + \mathbf{t}_K.$$

Since M_K is invertible, it admits a singular value decomposition $M_K = R_K^T \Lambda_K P_K$, where R_K and P_K are orthogonal and where Λ_K is diagonal with positive entries. In the following we set

$$\Lambda_K = \begin{pmatrix} \lambda_{1,K} & 0 \\ 0 & \lambda_{2,K} \end{pmatrix} \qquad \text{and} \qquad R_K = \begin{pmatrix} \mathbf{r}_{1,K}^T \\ \mathbf{r}_{2,K}^T \end{pmatrix}, \tag{3}$$

with the choice $\lambda_{1,K} \geq \lambda_{2,K}$. A simple example of such a transformation is when stretching occurs only in the x_1 direction. Let the vertices of the reference triangle $\hat{K}$ be $(0,0)$, $(1,0)$, $(0,1)$ and let the mapping T_K be defined by $x_1 = H\hat{x}_1$, $x_2 = h\hat{x}_2$, with $H \geq h$, thus

$$M_K = \begin{pmatrix} H & 0 \\ 0 & h \end{pmatrix}, \quad \lambda_{1,K} = H, \quad \lambda_{2,K} = h, \quad \mathbf{r}_{1,K} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \mathbf{r}_{2,K} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

A geometrical interpretation of the decomposition $M_K = R_K^T \Lambda_K P_K$ is the following. Consider the case when $\hat{K}$ is the unit equilateral reference triangle and consider the set of points lying on the unit circle, that is the points $\hat{\mathbf{x}}$ satisfying $\hat{\mathbf{x}}^T \hat{\mathbf{x}} = 1$. Since $\hat{\mathbf{x}} = M_K^{-1}(\mathbf{x} - \mathbf{t}_K)$, we have

$$1 = (\mathbf{x} - \mathbf{t}_K)^T M_K^{-T} M_K^{-1} (\mathbf{x} - \mathbf{t}_K) = (\mathbf{x} - \mathbf{t}_K)^T R_K^T \Lambda_K^{-2} R_K (\mathbf{x} - \mathbf{t}_K),$$

thus the unit circle is mapped into an ellipse with directions $\mathbf{r}_{1,K}$ and $\mathbf{r}_{2,K}$, the amplitude of stretching being $\lambda_{1,K}$ and $\lambda_{2,K}$.

In the frame of anisotropic meshes, the classical minimum angle condition is not required. However, for each vertex, the number of neighbouring vertices should be bounded from above, uniformely with respect to the mesh size h. Also, for each triangle K of the mesh, there is a restriction related to the patch Δ_K, the set of triangles having a vertex common with K. More precisely, the diameter of the reference patch $T_K^{-1}(\Delta_K)$ must be uniformly bounded independently of the mesh geometry. This assumption excludes some too distorted reference patches and implies that the local geometric quantities $\lambda_{i,K}$, $\mathbf{r}_{i,K}$, $i = 1,2$, vary smoothly on neighbouring triangles. In practice, no restrictions have been added in order to satisfy this condition and the anisotropic mesh generators that have been used [7,17,18] seem to satisfy the uniform boundedness of the reference patch.

3 Anisotropic, A Posteriori Error Estimates in the Energy Norm

Let us introduce the anisotropic error estimator proposed in [27] and corresponding to the error in the energy norm $\|\nabla(u-u_h)\|_{L^2(\Omega)}$. For all $K \in \mathscr{T}_h$, let

$$\Pi_K f = \frac{1}{|K|}\int_K f$$

be the $L^2(K)$ projection of f onto the constants. Let ℓ_i, $i = 1,2,3$ be the triangle three edges, let $[\cdot]$ denote the jump of the bracketed quantity across ℓ_i, with the convention $[\cdot] = 0$ for an edge ℓ_i on the boundary $\partial\Omega$. Then, the anisotropic error estimator corresponding to the energy norm on triangle K is defined by

$$\left(\eta_K^{EN}\right)^2 = \left(\|\Pi_K f\|_{L^2(K)} + \frac{1}{2}\sum_{i=1}^{3}\left(\frac{|\ell_i|}{\lambda_{1,K}\lambda_{2,K}}\right)^{1/2}\|[\nabla u_h \cdot \mathbf{n}]\|_{L^2(\ell_i)}\right)\omega_K(u-u_h). \tag{4}$$

Here $\mathbf{n}$ is the edge unit normal (in arbitrary direction), and $\omega_K(v)$ is defined for all $v \in H^1(\Omega)$ by

$$\omega_K^2(v) = \lambda_{1,K}^2\left(\mathbf{r}_{1,K}^T G_K(v)\mathbf{r}_{1,K}\right) + \lambda_{2,K}^2\left(\mathbf{r}_{2,K}^T G_K(v)\mathbf{r}_{2,K}\right), \tag{5}$$

where $G_K(v)$ denotes the 2×2 matrix defined by

$$G_K(v) = \begin{pmatrix} \int_{\Delta_K}\left(\frac{\partial v}{\partial x_1}\right)^2 dx & \int_{\Delta_K}\frac{\partial v}{\partial x_1}\frac{\partial v}{\partial x_2}dx \\ \int_{\Delta_K}\frac{\partial v}{\partial x_1}\frac{\partial v}{\partial x_2}dx & \int_{\Delta_K}\left(\frac{\partial v}{\partial x_2}\right)^2 dx \end{pmatrix}. \tag{6}$$

The following upper and lower bounds in the energy norm have been proved in [28]. Similar results can be found in [26].

Theorem 1. *There exists a constant C_1 independent of the mesh size and aspect ratio such that*

$$\int_\Omega |\nabla(u-u_h)|^2 \le C_1\Big(\sum_{K\in\mathscr{T}_h}(\eta_K^{EN})^2 + \sum_{K\in\mathscr{T}_h}\lambda_{1,K}^2\|f-\Pi_K f\|_{L^2(K)}^2\Big). \tag{7}$$

Moreover, if the mesh is such that there exists a constant C_2 independent of the mesh size and aspect ratio such that

$$\sum_{K\in\mathscr{T}_h}\frac{\omega_K^2(u-u_h)}{\lambda_{2,K}^2} \le C_2\int_\Omega |\nabla(u-u_h)|^2. \tag{8}$$

then, there exists a constant C_3 independent of the mesh size and aspect ratio such that

$$\sum_{K\in\mathcal{T}_h}(\eta_K^{EN})^2 \le C_3\Big(\int_\Omega |\nabla(u-u_h)|^2 + \sum_{K\in\mathcal{T}_h}\lambda_{1,K}^2\|f-\Pi_K f\|^2_{L^2(K)}\Big). \quad (9)$$

Remark 1. In the isotropic setting, $\lambda_{1,K}\simeq\lambda_{2,K}\simeq h_K$, assuming that f is smooth enough, estimates (7) and (9) yield

$$\int_\Omega |\nabla(u-u_h)|^2 \le C_1\sum_{K\in\mathcal{T}_h}\eta_K^2 + h.o.t. \quad \text{and} \quad \sum_{K\in\mathcal{T}_h}\eta_K^2 \le C_3\int_\Omega |\nabla(u-u_h)|^2 + h.o.t.,$$

where $h.o.t.$ denotes a high order term that behaves as $O(h^4)$ and where

$$\eta_K^2 = h_K^2\|\Pi_K f\|^2_{L^2(K)} + \frac{1}{2}\sum_{i=1}^{3}|\ell_i|\,\|\,[\nabla u_h\cdot\mathbf{n}]\,\|^2_{L^2(\ell_i)}$$

is the classical explicit, residual based error estimator studied for instance in [4,33]. In the isotropic setting, assumption (8) is not necessary but, in turn, the constants C_1 and C_3 hereabove depend on the mesh aspect ratio.

Remark 2. The estimator (4) is not a usual error estimator since u is still involved. However, if we can guess $u-u_h$, then (4) can be used to derive a computable quantity. This idea has been used in [27,28] and an efficient anisotropic error indicator has also been obtained replacing the derivatives

$$\frac{\partial(u-u_h)}{\partial x_i} \text{ in (6) by } \frac{\partial u_h}{\partial x_i} - \Pi_h\frac{\partial u_h}{\partial x_i}, \quad i=1,2, \quad (10)$$

where Π_h is an approximate $L^2(\Omega)$ projection onto V_h. More precisely, from constant values of $\partial u_h/\partial x_i$ on triangles, we build values at vertices P using the formula

$$\Pi_h\left(\frac{\partial u_h}{\partial x_i}\right)(P) = \frac{1}{\sum\limits_{\substack{K\in\mathcal{T}_h\\ P\in K}}|K|}\sum_{\substack{K\in\mathcal{T}_h\\ P\in K}}|K|\left(\frac{\partial u_h}{\partial x_i}\right)_{|K} \qquad i=1,2.$$

Approximating $\partial(u-u_h)/\partial x_i$ by $(I-\Pi_h)\partial u_h/\partial x_i$ is at the base of the celebrated Zienkiewicz–Zhu error estimator [35,36] and can be justified theoretically whenever superconvergence occurs, that is when $\nabla u - \Pi_h\nabla u_h$ is better than $O(h)$. For instance, it is proved in [1,30] that the Zienkiewicz–Zhu error estimator is asymptotically exact on parallel meshes, see also [10] for 3D results. Superconvergence has also be obtained for 2D mildly structured meshes in [34] but excludes for instance the chevron pattern, for which the Zienkiewicz–Zhu is not asymptotically exact, see

[30]. On general unstructured meshes, the Zienkiewicz–Zhu error estimator is only proved to be equivalent to the true error, see for instance [11,30] for isotropic meshes and [24] for anisotropic meshes. Numerical results show that the good properties of the Zienkiewicz–Zhu error estimator are underestimated by theoretical results.

Remark 3. Assumption (8) is true provided there exists a constant C independent of the mesh aspect ratio such that, for all $K \in \mathscr{T}_h$,

$$\lambda_{1,K}^2 \left(\mathbf{r}_{1,K}^T G_K (u - u_h) \mathbf{r}_{1,K} \right) \leq C \lambda_{2,K}^2 \left(\mathbf{r}_{2,K}^T G_K (u - u_h) \mathbf{r}_{2,K} \right), \tag{11}$$

in other words, when the error gradient in the direction of maximum stretching is less than the error gradient in the direction of minimum stretching. This is for instance the case when the error is equidistributed in both the directions of minimum and maximum stretching. This is precisely the goal of the adaptive algorithm described in [27, 28]. Numerical results reported in [27, 28] have shown that the effectivity index is aspect ratio independent for adapted meshes.

4 Goal Oriented, Anisotropic, A Posteriori Error Estimates

We now present an anisotropic error estimator for goal oriented a posteriori error estimates based on first order derivatives rather than second order derivatives. This error estimator has already been introduced in [15] for advection-diffusion-reaction problems but only an upper bound was proved. Hereafter, we propose a lower bound proceeding as in [28]. We refer [6, 29] for goal oriented, isotropic a posteriori error estimates.

In order to simplify the presentation we consider the linear functional J defined for all $v \in L^1(\Omega)$ by

$$J(v) = \int_{\Omega_0} v,$$

where $\Omega_0 \subset \Omega$. Our goal is now to control $J(u - u_h)$ and we introduce the dual problem: find $z \in H_0^1(\Omega)$ such that

$$\int_\Omega \nabla z \cdot \nabla v = J(v) \qquad \forall v \in H_0^1(\Omega). \tag{12}$$

We also need the corresponding finite element approximation of z namely $z_h \in V_h$ such that

$$\int_\Omega \nabla z_h \cdot \nabla v_h = J(v_h) \qquad \forall v \in V_h. \tag{13}$$

The error estimator corresponding to the goal oriented error $J(u - u_h)$ on triangle K is now defined by

$$(\eta_K^{GO})^2 = \left(\|\Pi_K f\|_{L^2(K)} + \frac{1}{2} \sum_{i=1}^{3} \left(\frac{|\ell_i|}{\lambda_{1,K}\lambda_{2,K}} \right)^{1/2} \|[\nabla u_h \cdot \mathbf{n}]\|_{L^2(\ell_i)} \right) \omega_K(z - z_h), \tag{14}$$

and proceeding as in [28], we can prove the following.

Proposition 1. *There is a constant C independent of the mesh size and aspect ratio such that*

$$J(u - u_h) \leq C \left(\sum_{K \in \mathcal{T}_h} (\eta_K^{GO})^2 + \sum_{K \in \mathcal{T}_h} \|f - \Pi_K f\|_{L^2(K)} \omega_K(z - z_h) \right). \tag{15}$$

Proof. Let I_h be Clément's interpolant [14]. From Proposition 3.2 in [19], there exists a constant C depending only on the reference element $\hat{K}$ such that, for all $v \in H^1(\Omega)$, for all $K \in \mathcal{T}_h$

$$\|v - I_h v\|_{L^2(K)} \leq C \omega_K(v). \tag{16}$$

Moreover, proceeding as in the proof of Proposition 2 in [20], there exists a constant C depending only on the reference element $\hat{K}$ such that, for all $v \in H^1(\Omega)$, for all $K \in \mathcal{T}_h$, for $i = 1, 2, 3$

$$\|v - I_h v\|_{L^2(\ell_i)} \leq C \left(\frac{|\ell_i|}{\lambda_{1,K}\lambda_{2,K}} \right)^{1/2} \omega_K(v).$$

Using (1) (2) (12) (13) we have

$$\begin{aligned} J(u - u_h) &= \int_\Omega f(z - z_h - v_h) - \int_\Omega \nabla u_h \cdot \nabla(z - z_h - v_h) \\ &= \sum_{K \in \mathcal{T}_h} \left(\int_K (\Pi_K f + \Delta u_h)(z - z_h - v_h) + \frac{1}{2} \int_{\partial K} [\nabla u_h \cdot \mathbf{n}](z - z_h - v_h) \right) \\ &\quad + \sum_{K \in \mathcal{T}_h} \int_K (f - \Pi_K f)(z - z_h - v_h), \end{aligned}$$

for all $v_h \in V_h$. We then choose $v_h = I_h(z - z_h)$, use Cauchy–Schwarz inequality and the above anisotropic interpolation estimates to obtain the result.

The proof of the following result is as in [28].

Proposition 2. *There exists a function $\varphi \in H_0^1(\Omega)$ and a constant C independent of the mesh size and aspect ratio such that, for all $K \in \mathcal{T}_h$ we have*

$$\int_{\ell_i} [\nabla u_h \cdot \mathbf{n}] \varphi = \left(\frac{|\ell_i|}{\lambda_{1,K}\lambda_{2,K}} \right)^{1/2} \left(\int_{\ell_i} [\nabla u_h \cdot \mathbf{n}]^2 \right)^{1/2} \omega_K(z - z_h), \quad i = 1, 2, 3, \tag{17}$$

$$\int_K (\Pi_K f)\varphi = \left(\int_K (\Pi_K f)^2\right)^{1/2} \omega_K(z - z_h), \tag{18}$$

$$\int_K |\nabla\varphi|^2 \leq C\,\frac{\omega_K^2(z - z_h)}{\lambda_{2,K}^2}. \tag{19}$$

Proceeding as in [28], we then prove what follows.

Proposition 3. *There exists a constant C independent of the mesh size and aspect ratio such that*

$$\sum_{K\in\mathcal{T}_h} (\eta_K^{GO})^2 \leq C \sum_{K\in\mathcal{T}_h} \Big(\|\nabla(u-u_h)\|_{L^2(K)} + \lambda_{1,K}\|f - \Pi_K f\|_{L^2(K)}\Big)\frac{\omega_K(z - z_h)}{\lambda_{2,K}}. \tag{20}$$

From the three above Propositions, we can now state the main result of the paper, which can be compared to Theorem 1.

Theorem 2. *There exists a constant C_1 independent of the mesh size and aspect ratio such that*

$$J(u - u_h) \leq C_1\Big(\sum_{K\in\mathcal{T}_h} (\eta_K^{GO})^2 + \sum_{K\in\mathcal{T}_h} \lambda_{1,K}\|f - \Pi_K f\|_{L^2(K)}^2 + \sum_{K\in\mathcal{T}_h} \lambda_{1,K}\|\nabla(z - z_h)\|_{L^2(\Delta_K)}^2\Big). \tag{21}$$

Moreover, if the mesh is such that there exists a constant C_2 independent of the mesh size and aspect ratio such that

$$\frac{\omega_K^2(z - z_h)}{\lambda_{2,K}^2} \leq C_2\|\nabla(z - z_h)\|_{L^2(\Delta_K)}^2 \qquad \forall K \in \mathcal{T}_h, \tag{22}$$

then, there exists a constant C_3 independent of the mesh size and aspect ratio such that

$$\sum_{K\in\mathcal{T}_h} (\eta_K^{GO})^2 \leq C_3\Big(\|\nabla(u - u_h)\|_{L^2(\Omega)}\|\nabla(z - z_h)\|_{L^2(\Omega)} + \sum_{K\in\mathcal{T}_h} \lambda_{1,K}\|f - \Pi_K f\|_{L^2(K)}^2 + \sum_{K\in\mathcal{T}_h} \lambda_{1,K}\|\nabla(z - z_h)\|_{L^2(\Delta_K)}^2\Big). \tag{23}$$

Proof. To obtain (21), it suffices to notice that

$$\omega_K(z - z_h) \leq \lambda_{1,K}\|\nabla(z - z_h)\|_{L^2(\Delta_K)}$$

in (15) and to use Young's inequality. On the other side, if we use assumption (22) in (20) and apply Young's inequality, we obtain (23).

The three following remarks are similar to Remarks 1–3.

Remark 4. In the isotropic setting, $\lambda_{1,K} \simeq \lambda_{2,K} \simeq h_K$, assuming that $f \in H^1(\Omega)$ and $z \in H^2(\Omega)$, estimates (21) and (23) write

$$J(u-u_h) \leq C_1 \sum_{K\in\mathscr{T}_h} \eta_K^2 + h.o.t.$$

and

$$\sum_{K\in\mathscr{T}_h} \eta_K^2 \leq C_3 \|\nabla(u-u_h)\|_{L^2(\Omega)} \|\nabla(z-z_h)\|_{L^2(\Omega)} + h.o.t., \tag{24}$$

where $h.o.t.$ denotes a high order term that behaves as $O(h^3)$ and where

$$\eta_K^2 = \left(h_K^2 \|\Pi_K f\|_{L^2(K)}^2 + \frac{1}{2} \sum_{i=1}^{3} |\ell_i| \| [\nabla u_h \cdot \mathbf{n}] \|_{L^2(\ell_i)}^2 \right)^{1/2} \|\nabla(z-z_h)\|_{L^2(\Delta_K)}.$$

In general, the last term in the above definition is estimated using interpolation results, hierarchical techniques, or post-processing [6, 32]. In the isotropic setting, assumption (22) is not necessary but, in turn, the constants C_1 and C_3 hereabove depend on the mesh aspect ratio. Moreover, whenever u and z are in $H^2(\Omega)$, then (24) writes

$$\sum_{K\in\mathscr{T}_h} \eta_K^2 \leq C_4 h^2 + h.o.t.,$$

where C_4 is independent of h, thus the error estimator is of optimal $O(h^2)$ order.

Remark 5. The estimator (4) is not a usual error estimator since z is still involved. However, if we can guess $z-z_h$, then (4) can be used to derive a computable quantity. Following [27, 28], we propose to replace the derivatives

$$\frac{\partial(z-z_h)}{\partial x_i} \text{ in (6) by } \frac{\partial z_h}{\partial x_i} - \Pi_h \frac{\partial z_h}{\partial x_i}, \quad i = 1, 2, \tag{25}$$

where Π_h is an approximate $L^2(\Omega)$ projection onto V_h.

Remark 6. Assumption (22) is true provided there exists a constant C independent of the mesh aspect ratio such that, for all $K \in \mathscr{T}_h$,

$$\lambda_{1,K}^2 \left(\mathbf{r}_{1,K}^T G_K (z-z_h) \mathbf{r}_{1,K} \right) \leq C \lambda_{2,K}^2 \left(\mathbf{r}_{2,K}^T G_K (z-z_h) \mathbf{r}_{2,K} \right),$$

in other words, when the dual error gradient in the direction of maximum stretching is less than the dual error gradient in the direction of minimum stretching. This is for instance the case when the dual error is equidistributed in both the directions of minimum and maximum stretching. This condition should be enforced in the adaptive algorithm for goal oriented anisotropic meshes.

Acknowledgements Wissam Hassan is supported by Dassault Aviation. The authors thank Didier Alleau, Jean-Pierre Figeac, Nicolas Flandrin, Alain Naim and Gilbert Rogé from Dassault Aviation for fruitful discussions.

References

1. Ainsworth, M., Oden, J.T.: A posteriori error estimation in finite element analysis. Comput. Meth. Appl. Mech. Engrg. **142**(1–2), 1–88 (1997)
2. Alauzet, F.: High-order methods and mesh adaptation for Euler equations. Internat. J. Numer. Meth. Fluids **56**(8), 1069–1076 (2008)
3. Babuška, I., Rheinboldt, W.C.: Error estimates for adaptive finite element computations. SIAM J. Numer. Anal. **15**(4), 736–754 (1978)
4. Babuška, I., Durán, R., Rodríguez, R.: Analysis of the efficiency of an a posteriori error estimator for linear triangular finite elements. SIAM J. Numer. Anal. **29**(4), 947–964 (1992)
5. Bangerth, W., Hartmann, R., Kanschat, G.: deal.II—a general-purpose object-oriented finite element library. ACM Trans. Math. Software **33**(4), Art. 24, 27 (2007)
6. Becker, R., Rannacher, R.: An optimal control approach to a posteriori error estimation in finite element methods. Acta Numer. **10**, 1–102 (2001)
7. Borouchaki, H., Laug, G.: The BL2D Mesh Generator : Beginner's Guide, User's and Programmer's Manual. Technical report RT-0194, Institut National de Recherche en Informatique et Automatique (INRIA), Rocquencourt, 78153 Le Chesnay, France (1996)
8. Borouchaki, H., George, P.L., Hecht, F., Laug, P., Saltel, E.: Delaunay mesh generation governed by metric specifications. I. Algorithms. Finite Elem. Anal. Des. **25**(1–2), 61–83 (1997). Adaptive meshing, Part 1
9. Bourgault, Y., Picasso, M., Alauzet, F., Loseille, A.: On the use of anisotropic a posteriori error estimators for the adaptative solution of 3D inviscid compressible flows. Internat. J. Numer. Meth. Fluids **59**(1), 47–74 (2009)
10. Brandts, J., Křížek, M.: Gradient superconvergence on uniform simplicial partitions of polytopes. IMA J. Numer. Anal. **23**(3), 489–505 (2003)
11. Carstensen, C.: All first-order averaging techniques for a posteriori finite element error control on unstructured grids are efficient and reliable. Math. Comp. **73**(247), 1153–1165 (electronic) (2004)
12. Chen, L., Sun, P., Xu, J.: Optimal anisotropic meshes for minimizing interpolation errors in L^p-norm. Math. Comp. **76**(257), 179–204 (electronic) (2007)
13. Ciarlet, P.G.: Basic error estimates for elliptic problems. In: Handbook of numerical analysis, Vol. II, Handb. Numer. Anal., II, pp. 17–351. North-Holland, Amsterdam (1991)
14. Clément, P.: Approximation by finite element functions using local regularization. RAIRO Analyse Numérique **9**(R-2), 77–84 (1975)
15. Dedè, L., Micheletti, S., Perotto, S.: Anisotropic error control for environmental applications. Appl. Numer. Math. **58**(9), 1320–1339 (2008)
16. Demkowicz, L.: Computing with hp-adaptive finite elements. Vol. 1. Chapman & Hall/CRC Applied Mathematics and Nonlinear Science Series. Chapman & Hall/CRC, Boca Raton, FL (2007). One and two dimensional elliptic and Maxwell problems, With 1 CD-ROM (UNIX)
17. Distene S.A.S., Pôle Teratec - BARD-1, Domaine du Grand Rué, 91680 Bruyères-le-Chatel, France: MeshAdapt : A mesh adaptation tool, User's manual Version 3.0 (2003)
18. Dobrzynski, C., Frey, P.J., Mohammadi, B., Pironneau, O.: Fast and accurate simulations of air-cooled structures. Comput. Meth. Appl. Mech. Engrg. **195**(23–24), 3168–3180 (2006)
19. Formaggia, L., Perotto, S.: New anisotropic a priori error estimates. Numer. Math. **89**(4), 641–667 (2001)
20. Formaggia, L., Perotto, S.: Anisotropic error estimates for elliptic problems. Numer. Math. **94**(1), 67–92 (2003)

21. Frey, P.J., Alauzet, F.: Anisotropic mesh adaptation for CFD computations. Comput. Meth. Appl. Mech. Engrg. **194**(48–49), 5068–5082 (2005)
22. Habashi, W.G., Dompierre, J., Bourgault, Y., Ait-Ali-Yahia, D., Fortin, M., Vallet, M.G.: Anisotropic mesh adaptation: towards user-independent, mesh-independent and solver-independent CFD. I. General principles. Internat. J. Numer. Meth. Fluids **32**(6), 725–744 (2000)
23. Kunert, G.: An a posteriori residual error estimator for the finite element method on anisotropic tetrahedral meshes. Numer. Math. **86**(3), 471–490 (2000)
24. Kunert, G., Nicaise, S.: Zienkiewicz-Zhu error estimators on anisotropic tetrahedral and triangular finite element meshes. M2AN Math. Model. Numer. Anal. **37**(6), 1013–1043 (2003)
25. Loseille, A.: Adaptation de maillage anisotrope 3d multi-échelles et ciblée à une fonctionnelle pour la mécanique des fluides. application la prédiction haute-fidélité du bang sonique. PhD Thesis, Paris VI (2008)
26. Micheletti, S., Perotto, S.: Reliability and efficiency of an anisotropic Zienkiewicz–Zhu error estimator. Comput. Meth. Appl. Mech. Engrg. **195**(9–12), 799–835 (2006)
27. Picasso, M.: An anisotropic error indicator based on Zienkiewicz–Zhu error estimator: application to elliptic and parabolic problems. SIAM J. Sci. Comput. **24**(4), 1328–1355 (electronic) (2003)
28. Picasso, M.: Adaptive finite elements with large aspect ratio based on an anisotropic error estimator involving first order derivatives. Comput. Meth. Appl. Mech. Engrg. **196**(1–3), 14–23 (2006)
29. Rannacher, R.: Adaptive Galerkin finite element methods for partial differential equations. J. Comput. Appl. Math. **128**(1–2), 205–233 (2001). Numerical analysis 2000, Vol. VII, Partial differential equations
30. Rodríguez, R.: Some remarks on Zienkiewicz–Zhu estimator. Numer. Meth. Partial Differential Equations **10**(5), 625–635 (1994)
31. Schmidt, A., Siebert, K.G.: Design of adaptive finite element software, *Lecture Notes in Computational Science and Engineering*, vol. 42. Springer, Berlin (2005). The finite element toolbox ALBERTA, With 1 CD-ROM (Unix/Linux)
32. Suttmeier, F.T.: Reliable, goal-oriented postprocessing for FE-discretizations. Numer. Meth. Partial Differential Equations **21**(2), 387–396 (2005)
33. Verfürth, R.: A review of a posteriori error estimation and adaptive mesh-refinement techniques. Wiley-Teubner (1996)
34. Xu, J., Zhang, Z.: Analysis of recovery type a posteriori error estimators for mildly structured grids. Math. Comp. **73**(247), 1139–1152 (electronic) (2004)
35. Zienkiewicz, O.C., Zhu, J.Z.: A simple error estimator and adaptive procedure for practical engineering analysis. Int. J. Numer. Meth. Engrg. **24**(2), 337–357 (1987)
36. Zienkiewicz, O.C., Zhu, J.Z.: The superconvergent patch recovery and a posteriori error estimates. I. The recovery technique. Int. J. Numer. Meth. Engrg. **33**(7), 1331–1364 (1992)

Part II
Contributed Papers

Energy Stability of the MUSCL Scheme

Qaisar Abbas, Edwin van der Weide, and Jan Nordström

Abstract We analyze the energy stability of the standard MUSCL scheme. The analysis is possible by reformulating the MUSCL scheme in the framework of summation-by-parts (SBP) operators including an artificial dissipation. The effect of different slope limiters is studied. It is found that all the slope limiters do not lead to the correct sign of the entries in the dissipation matrix. The implication of that is discussed. The analysis is done for both linear and nonlinear scalar problems.

1 Introduction

For problems involving shocks which arise in computational fluid mechanics and related areas, the MUSCL scheme [10] is a very effective approach to resolve discontinuities. This scheme ensures the monotonicity of the solution for the whole computational time and it is arguably computationally less expensive compared to relevant counterparts like the WENO schemes [5] for approximately the same accuracy.

Q. Abbas (✉)
Department of Information Technology, Scientific Computing, Uppsala University, SE-751 05 Uppsala, Sweden
e-mail: qaisar.abbas@it.uu.se

E. van der Weide
Faculty of Engineering Technology, University of Twente, 7500 AE Enschede, The Netherlands
e-mail: vdweide@gmail.com

J. Nordström
School of Mechanical, Industrial and Aeronautical Engineering, University of the Witvatersrand, PO WITS 2050, Johannesburg, South Africa
and
Department of Aeronautics and Systems Integration, FOI, The Swedish Defence Research Agency, SE-164 90 Stockholm, Sweden
and
Department of Information Technology, Scientific Computing, Uppsala University, SE-751 05 Uppsala, Sweden
e-mail: jan.nordstrom@it.uu.se

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_5,

In this paper, we have reformulated the MUSCL scheme in summation-by-parts (SBP) form including an artificial dissipation operator. Related work can be found in [11, 12], where the WENO scheme has been formulated in a similar way. The SBP operators are well-established theoretically [4] and their usefulness is proven for practical applications, see [7–9].

In this work we will investigate the MUSCL scheme to see if the scheme is energy stable, i.e., stable in the L_2-norm, see [11, 12]. We consider both scalar linear and nonlinear hyperbolic problems in one dimension. Our analysis is based on theoretical as well as numerical observations.

2 The MUSCL Scheme in SBP Form

Consider the unsteady one-dimensional conservation law

$$u_t + f(u)_x = 0, \qquad 0 \le x \le 1, \quad t \ge 0. \tag{1}$$

Define a uniform grid $x_j = j\Delta x$, $j = 0, \ldots, N$, with $\Delta x = 1/N$. On the grid, define a flux $F(U)$, where $U = [U_0(t), U_1(t) \ldots, U_N(t)]^T$ is the discrete approximation of the solution u in (1). The second order upwind discretization of (1) using the MUSCL approach [10] results in

$$U_t + RES_i = 0, \quad RES_i = \frac{1}{\Delta x}\left(F_{i+\frac{1}{2}} - F_{i-\frac{1}{2}}\right). \tag{2}$$

In (2), $F_{i+\frac{1}{2}}$ is the flux function at the interface $i + \frac{1}{2}$. More details on the computation of numerical flux function can be found in [1].

Similarly a second order discretization of the flux function in (1), obeying the SBP property [4] and with the introduction of artificial dissipation on SBP form [6] leads to

$$U_t + D_2 F = -P^{-1}\widetilde{D}_1^T B_M \widetilde{D}_1 U, \tag{3}$$

where D_2 is the central finite difference operator on SBP form given by

$$D_2 = P^{-1}Q, \qquad Q + Q^T = B, \quad P = P^T, \tag{4}$$

$$D_2 = \frac{1}{2\Delta x}\begin{bmatrix} -2 & 2 & & & \\ -1 & 0 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 0 & 1 \\ & & & -2 & 2 \end{bmatrix}, \quad P^{-1} = \frac{1}{\Delta x}\begin{bmatrix} 2 & & & & \\ & 1 & & & \\ & & \ddots & & \\ & & & 1 & \\ & & & & 2 \end{bmatrix},$$

and $B = diag(-1, 0, \ldots, 0, 1)$. The term $-P^{-1}\widetilde{D}_1^T B_M \widetilde{D}_1 U$ in (3) is an artificial dissipation operator. It will be shown below that the matrix B_M can be constructed

such that (3) corresponds to the standard second order MUSCL formulation [10] which means that the formulations given by (2) and (3) are equivalent. $\widetilde{D}_1$ is a two point difference operator and the matrix B_M is a diagonal matrix, see (5).

$$\widetilde{D}_1 = \begin{bmatrix} -1 & 1 & & & \\ & -1 & 1 & & \\ & & \ddots & \ddots & \\ & & & -1 & 1 \\ & & & -1 & 1 \end{bmatrix}, \quad B_M = \begin{bmatrix} b_0 & & & & 0 \\ & b_1 & & & \\ & & \ddots & & \\ & & & b_{N-1} & \\ 0 & & & & 0 \end{bmatrix}. \tag{5}$$

2.1 Explicit Form of B_M

At an interior point i, we have

$$\left\{-P^{-1}\widetilde{D}_1^T B_M \widetilde{D}_1 U\right\}_i = -\frac{1}{\Delta x}(b_{i-1}\Delta U_{i-1} - b_i \Delta U_i), \tag{6}$$

where $\Delta U_i = U_{i+1} - U_i$. In combination with the central discretization of the convective term, this leads to the following formulation of the residual for an internal node

$$\Delta x RES_i = \frac{1}{2}(F_{i+1} - F_{i-1}) + b_{i-1}\Delta U_{i-1} - b_i \Delta U_i. \tag{7}$$

For the boundary nodes x_0 and x_N, the residuals are

$$\Delta x RES_0 = \Delta F_0 - \widetilde{P}_0^{-1} b_0 \Delta U_0, \quad \Delta x RES_N = \Delta F_{N-1} + \widetilde{P}_N^{-1} b_{N-1}\Delta U_{N-1}, \tag{8}$$

where $\widetilde{P}_0^{-1} = \widetilde{P}_N^{-1} = 2$. Comparing (2) and (7), it is clear that both schemes are identical if

$$b_i \Delta U_i = \frac{1}{2}(F_{i+1} + F_i) - F_{i+\frac{1}{2}}. \tag{9}$$

It can be shown that the b_i in (9) becomes

$$b_i = \frac{1}{2}\left\{ \left|A_{i+\frac{1}{2}}\right| \left(1 - \frac{\phi_i}{2} - \frac{\psi_{i+1}}{2}\right) + A_{R_{i+\frac{1}{2},i+1}} \frac{\psi_{i+1}}{2} - A_{L_{i+\frac{1}{2},i}} \frac{\phi_i}{2} \right\}, \tag{10}$$

where ϕ_i and ψ_{i+1} are the slope limiters involved in the fluxes. They are related in the following way.

$$\phi_i = \phi(r_i) = \phi\left(\frac{\Delta U_{i-1}}{\Delta U_i}\right) = \frac{\Delta U_{i-1}}{\Delta U_i}\phi\left(\frac{1}{r_i}\right) = \frac{\Delta U_{i-1}}{\Delta U_i}\psi_i. \tag{11}$$

Also $A = \frac{\partial F}{\partial U}$ is a Jacobian matrix evaluated at the Roe average states. The property of a Roe average state is that $f_2 - f_1 = A_{Roe}(u_2 - u_1)$.

2.2 Energy Stability

In this section we define the two versions of the energy stability, that we will work with in the analysis below.

Definition 1. Consider (3) and (12). The scheme defined by (3) is *pointwise energy stable* if $b_i \geq 0$ for all $i = 0, 1 \ldots, N$. The scheme defined by (3) is *energy stable in the mean* if $(DU)^T B_M (DU) \geq 0$, where $DU = [(DU)_0, (DU)_1, \ldots, (DU)_N]^T$.

Remark 1. Pointwise energy stable schemes lead to energy stable schemes in the mean. The reverse is not true.

2.3 Energy Estimates

To investigate whether the scheme defined in (3) is energy stable or not, we start by considering the linear constant coefficient case with $F = aU$ and use the energy method. Multiplying (3) with $U^T P$, adding its transpose and using (4) leads to

$$\frac{d}{dt}||U||_P^2 + aU^T BU = -2(\widetilde{D}_1 U)^T B_M (\widetilde{D}_1 U). \tag{12}$$

where $||U||_P^2 = U^T PU$. For a bounded solution and energy stability we must have $\frac{d}{dt}||U||_P^2 \leq 0$. The boundary terms $U^T BU = U_0^2 - U_N^2$ can be bounded using the SAT boundary treatment [3] and are ignored from now on. The right-hand side of (12) is negative if the matrix B_M is positive semi-definite. The matrix B_M for a linear problem becomes

$$b_i = \frac{1}{2}\{|A| - \phi_i A^+ + \psi_{i+1} A^-\}, \tag{13}$$

where A^+ contains the positive eigenvalues of A and A^- the negative ones,

$$A^+ = \frac{1}{2}(A + |A|), \quad A^- = \frac{1}{2}(A - |A|). \tag{14}$$

For a scalar problem with $F = aU$, (13) reduces to

$$b_i = \frac{1}{2}a\{1 - \phi_i\}, \quad a > 0, \quad \text{and} \quad b_i = \frac{1}{2}|a|\{1 - \psi_{i+1}\}, \quad a < 0. \tag{15}$$

From the theory of the slope limiters [2] we have that $0 \leq \phi_i, \psi_{i+1} \leq 2$. It is obvious that any limiter which takes values greater than 1, will lead to $b_i \leq 0$ in

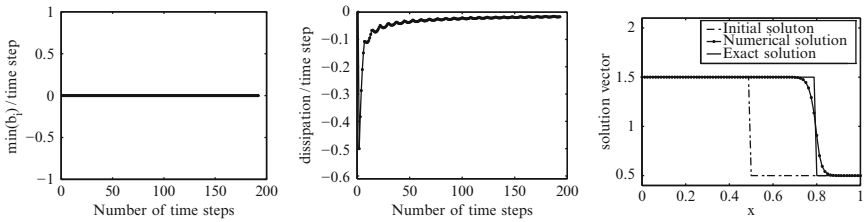
the computational domain and hence no pointwise energy stability. In [11, 12], the authors modified the WENO scheme by correcting this anomaly of the scheme. We will discuss below whether that is necessary and meaningful.

3 Numerical Results

Consider (10)–(12). It is obvious that the sign of b_i depends on the slope limiters involved in the MUSCL scheme. If the solution is smooth, we have $\phi_i = \psi_{i+1} = 1$, and for all A, b_i will be zero. For problems with discontinuities , we could have $0 \leq \phi_i,\ \psi_{i+1} \leq 2$, which decides the sign of b_i in non-smooth regions.

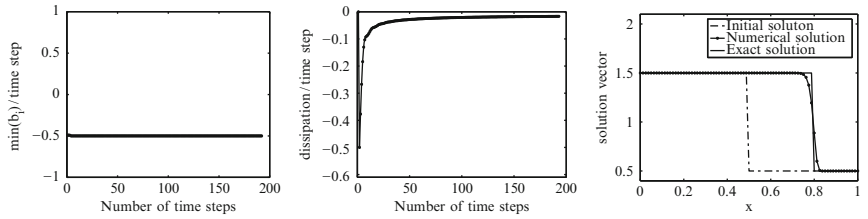
We consider a linear problem ($f = u$) first with a step discontinuity as initial data and analyze four different limiters. All the results are shown for $N = 80$ and $t = 0.3$. In Figs. 1–4 we have shown the minimum of b_i and $-(D_1U)^T B_M (D_1U)$ at each time step for minmod, VanLeer, superbee and MC limiters. The minmod limiter has $b_i \geq 0$ for all time and hence is pointwise stable. All other limiters lead to $b_i < 0$ at few points near the discontinuity. It means that these limiters do not lead to pointwise stability. It is also found that $-(D_1U)^T B_M (D_1U) \leq 0$ for all limiters for the whole computational time which gives energy stability in the mean.

Next we consider the Burger equation with $f = \frac{u^2}{2}$ in (1) and repeat the same analysis with minmod, VanLeer, superbee and MC limiters. It is found that all the tested limiters have some $b_i < 0$ but the minmod limiter is almost zero for all time leading to pointwise stability, see Figs. 5–8. It can also be seen that all schemes are stable in the mean.



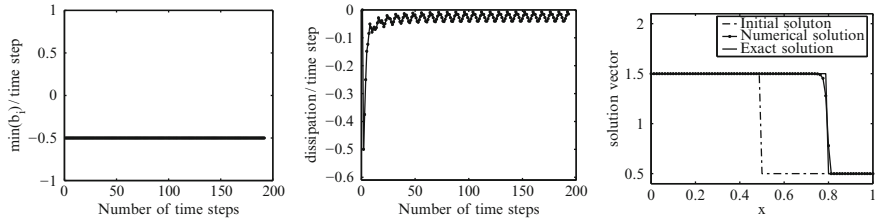
(a) $min(b_i)$ per time step (b) $-(D_1U)^T B_M (D_1U)$ (c) solution, $N = 80, t = 0.3$

Fig. 1 Results from the MUSCL in SBP form using the *minmod limiter*, $f = u$



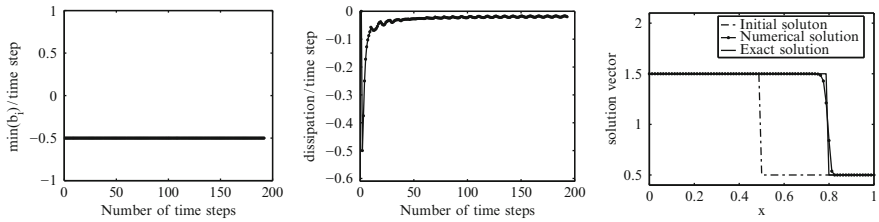
(a) $min(b_i)$ per time step (b) $-(D_1U)^T B_M (D_1U)$ (c) solution, $N = 80, t = 0.3$

Fig. 2 Results from the MUSCL in SBP form using the *Van Leer limiter*, $f = u$



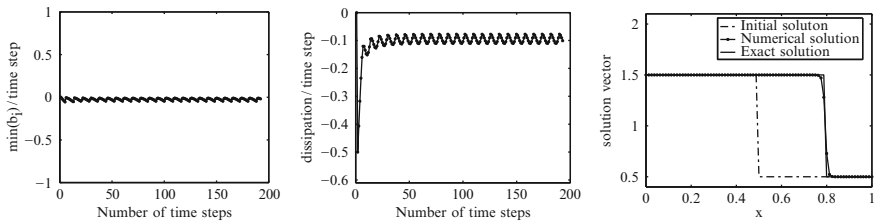
(a) $min(b_i)$ per time step (b) $-(D_1U)^T B_M (D_1U)$ (c) solution, $N = 80, t = 0.3$

Fig. 3 Results from the MUSCL in SBP form using the *Superbee limiter*, $f = u$



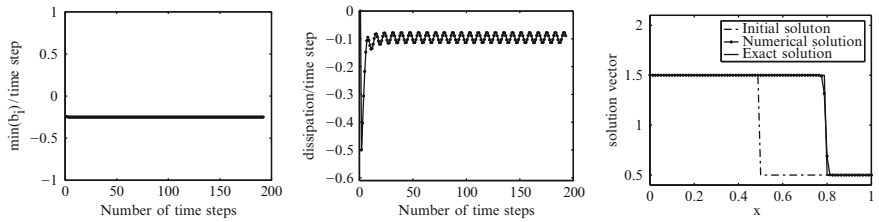
(a) $min(b_i)$ per time step (b) $-(D_1U)^T B_M (D_1U)$ (c) solution, $N = 80, t = 0.3$

Fig. 4 Results from the MUSCL in SBP form using the *MC limiter*, $f = u$



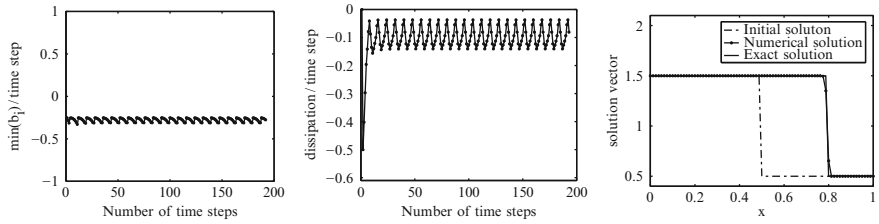
(a) $min(b_i)$ per time step (b) $-(D_1U)^T B_M (D_1U)$ (c) solution, $N = 80, t = 0.3$

Fig. 5 Results from the MUSCL in SBP form using the *minmod limiter*, $f = \frac{u^2}{2}$



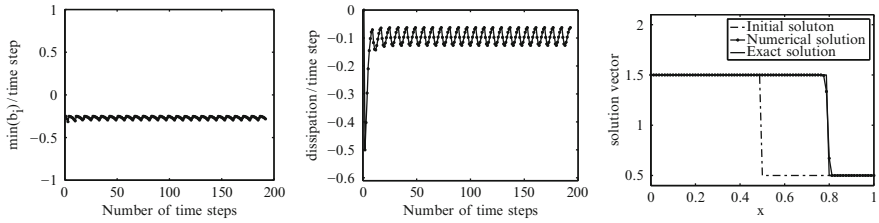
(a) $min(b_i)$ per time step (b) $-(D_1U)^T B_M (D_1U)$ (c) solution, $N = 80, t = 0.3$

Fig. 6 Results from the MUSCL in SBP form using the *Van Leer limiter*, $f = \frac{u^2}{2}$



(a) $min(b_i)$ per time step (b) $-(D_1U)^T B_M (D_1U)$ (c) solution, $N = 80, t = 0.3$

Fig. 7 Results from the MUSCL in SBP form using the *Superbee limiter*, $f = \frac{u^2}{2}$



(a) $min(b_i)$ per time step (b) $-(D_1U)^T B_M (D_1U)$ (c) solution, $N = 80, t = 0.3$

Fig. 8 Results from the MUSCL in SBP form using the *MC limiter*, $f = \frac{u^2}{2}$

Table 1 Analysis of different limiters for the linear problem ($a = 1, t = 0.3$)

Limiters	l_2-error, min($b_i \leq 0$)	l_2-error, min($b_i = 0$)
minmod	0.0711	0.0711
Van Leer	0.0578	0.0642
Superbee	0.0400	0.0634
MC	0.0504	0.0639

It is not clear whether pointwise stability is necessary or if stability in the mean is enough. If we replace $b_i < 0$ in the matrix B_M with $b_i = 0$ at each time step, we find that it leads to an additional and excessive amount of dissipation in the discontinuity/shock region, see Table 1 for l_2-error of solutions. By demanding the pointwise stability, clearly the sharpness of the shock decreases.

4 Conclusion

We have expressed the MUSCL scheme as a combination of an SBP operator and an artificial dissipation operator. This form allows us to use the energy method for analyzing stability. Our main interest was to look at the behavior of dissipation matrix B_M in (5), which is crucial for the stability of the scheme and also influence the sharpness of the shock.

As the matrix depends on the slope limiters of the MUSCL scheme, it was found most of the tested limiters except minmod limiter do not lead to pointwise stability while all limiters are stable in the mean.

By making the schemes pointwise stable by replacing $b_i < 0$ in the matrix B_M with $b_i = 0$ resulted in an additional and excessive dissipation for all the limiters. It was shown that the error in the calculations increased and the sharpness of the shock decreased. This procedure was used in [11, 12] but seems questionable.

References

1. Abbas Q., van der Weide E., Nordström J.: Accurate and stable calculations involving shocks using a new hybrid scheme. In Proc. 19th AIAA CFD Conference, volume 2009–3985 of Conference Proceeding Series, AIAA (2009)
2. Berger M.H., Aftosmis M.J., Murman S.M.: Analysis of slope limiters on irregular grids. Technical Report NAS-05-007, NAS Technical Report (2005)
3. Carpenter M.H., Gottlieb D., Abarbanel S.: The stability of numerical boundary treatments for compact high-order finite difference schemes. J. Comput. Phys. **108(2)**, 272–295 (1994)
4. Carpenter M.H., Nordström J., Gottlieb D.: A stable and conservative interface treatment of arbitrary spatial accuracy. J. Comput. Phys. **148(2)**, 341–365 (1999)
5. Jiang G., Shu C.W.: Efficient implementation of weighted ENO schemes. J. Comput. Phys. **126**, 202–228 (1996)
6. Mattsson K., Svärd M., Nordström J.: Stable and accurate artificial dissipation. J. Sci. Comput. **21(1)**, 57–79 (2004)
7. Nordström J., Gong J., van der Weide E., Svärd M.: A stable and conservative high order multi-block method for the compressible Navier-Stokes equations. J. Comput. Phys. **228**, 9020–9035 (2009)
8. Svärd M., Carpenter M.H., Nordström J.: A stable high-order finite difference scheme for the compressible Navier-Stokes equations, far-field boundary conditions. J. Comput. Phys. **225(1)**, 1020–1038 (2007)
9. Svärd M., Nordström J., A stable high-order finite difference scheme for the compressible Navier-Stokes equations: wall boundary conditions. J. Comput. Phys. **227**, 4805–4824 (2008)
10. van Leer B.: Towards the ultimate conservative difference scheme, V. A second order sequel to Godunov's method. J. Comput. Phys. **32**, 101–136 (1979)
11. Yamaleev N.K., Carpenter M.H.: Third-order energy stable WENO scheme. J. Comput. Phys. **228**, 3025–3047 (2009)
12. Yamaleev N.K., Carpenter M.H.: A systematic methodology for constructing high-order energy stable WENO schemes. J. Comput. Phys. **228**, 4248–4272 (2009)

Numerical Stabilization of the Melt Front for Laser Beam Cutting

Torsten Adolph, Willi Schönauer, Markus Niessen, and Wolfgang Schulz

Abstract The Finite Difference Element Method (FDEM) is a black-box solver that solves by a finite difference method arbitrary nonlinear systems of elliptic and parabolic partial differential equations (PDEs) on an unstructured FEM grid in 2D or 3D. For each node we generate difference formulas of consistency order q with a sophisticated algorithm. An unprecedented feature for such a general black-box is the error estimate that is computed together with the solution. In this paper we present the numerical simulation of the laser beam cutting of a metal sheet. This is a free boundary problem where we compute the temperature and the form of the melt front in the metal sheet. During the cutting process, the numerical stabilization of the melt front is a great challenge.

1 Finite Difference Element Method FDEM

FDEM is an unprecedented generalization of the FDM on an unstructured FEM mesh. It is a black-box solver for arbitrary nonlinear systems of 2D and 3D elliptic or parabolic PDEs. With certain restrictions it can be used also for hyperbolic PDEs. If the unknown solution is $u(t, x, y, z)$, the operator for PDEs and BCs (boundary conditions) is (2.4.1) and (2.4.2) in [7]:

$$Pu \equiv P(t, x, y, z, u, u_t, u_x, u_y, u_z, u_{xx}, u_{yy}, u_{zz}, u_{xy}, u_{xz}, u_{yz}) = 0. \qquad (1)$$

T. Adolph (✉) and W. Schönauer
Karlsruhe Institute of Technology, Steinbuch Centre for Computing, Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen, Germany
e-mail: torsten.adolph@kit.edu, willi.schoenauer@kit.edu

M. Niessen and W. Schulz
Fraunhofer Institute for Laser Technology, ILT Aachen, Steinbachstraße 15, 52074 Aachen, Germany
e-mail: markus.niessen@ilt.fraunhofer.de, wolfgang.schulz@ilt.fraunhofer.de

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_6,

For a system of m PDEs, u and Pu have m components:

$$u = (u_1, \ldots, u_m)^T, \quad Pu = (P_1 u, \ldots, P_m u)^T. \tag{2}$$

As we have a black-box solver, the PDEs and BCs and their Jacobian matrices of type (2.4.6) in [7] must be entered as Fortran code in prescribed frames.

The geometry of the domain of solution is entered as a FEM mesh with triangles in 2D and tetrahedra in 3D. The domain may be composed of subdomains with different PDEs and non-matching grid. From the element list and its inverted list, we select for each node more than the necessary number of nodes for difference formulas of a given consistency order q. By a sophisticated algorithm, the necessary optimal nodes are determined from this set, see Sect. 2.2 in [7]. From the difference of formulas of different consistency order, we get an estimate of the discretization error. If we want e.g., the discretization error for u_x, and $u_{x,d,q}$ denotes the difference formula of consistency order q, the error estimate d_x is defined by

$$d_x := u_{x,d,q+2} - u_{x,d,q}, \tag{3}$$

i.e., by the difference to the order $q+2$. This has a built-in self-control: If this is not a "better" formula, the error estimate shows large error.

With such error estimates, we can explicitly compute the error of the solution by the error equation (2.4.8) in [7]. The knowledge of the error allows local mesh refinement and control of the space consistency order, see Sect. 2.5 in [7]. There we also explain the computation of the time step size Δt and the selection of the consistency order p in time direction for parabolic problems. There is also computed a global error estimate in time direction that gives the history of the discretization and linearization errors in time.

A special problem for a black-box solver is the efficient parallelization with MPI. The user enters his domain by the FEM mesh, and we use a 1D domain decomposition with overlap to distribute the data to the processors, see Sect. 2.8 in [7]. A detailed report on the parallelization is [1]. The resulting large and sparse linear system is solved by the LINSOL program package [6] that is also efficiently parallelized for iterative methods of CG type and (I)LU preconditioning.

2 The One-Phase Problem

Laser cutting is a thermal separation process widely used in shaping and contour cutting applications. The cutting process is described with a spatial 3D Free Boundary Problem for the motion of one-phase boundary. In cutting, the interaction of the gaseous, liquid and solid domains takes place only across spatially 2D surfaces. The solid volume $\Omega(\tau)$ (Fig. 1) is bounded by the absorption front $\Gamma_+(\tau)$ and the bottom surface $\Gamma_-(\tau)$. A part of the absorption front is a free boundary called the melt front $\Gamma_m(\tau)$. Hence, adding the various effects of the gas and melt flow only changes the boundary values, while the structure of the model for the solid, called the one-phase

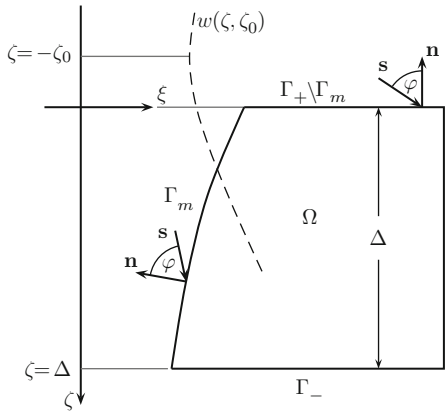


Fig. 1 The absorption front $\Gamma_+(\tau)$ consists of two parts: The melt front $\Gamma_m(\tau)$ and a part of the surface $\Gamma_+(\tau)\backslash\Gamma_m(\tau)$ of the solid metal $\Omega(\tau)$ where $\theta < 1$ holds. The laser beam propagates in ζ-direction, the laser beam axis has the position $\xi = 0$ and the focal position of the laser beam is at $\zeta = -\zeta_0$. The metal has thickness Δ and its upper surface is at $\zeta = 0$. $w(\zeta, \zeta_0)$ is the Gaussian beam width (5)

model, remains unchanged. To find the properties of a comprehensive cutting model relies on the detailed analysis of the one-phase model, which was first formulated in 1997 by Enß et al. [4].

The phase boundary of the one-phase Free Boundary Problem can move into the material and erosion takes place or remains unchanged. Resolidification is not allowed. The intensity $I = I_0(t) f((\mathbf{x} - \mathbf{v_0}t)/w_0)$, $\mathbf{x} \in \mathbb{R}^2$, of the laser beam is characterized by its maximum value $I_0(t)$, the spatial distribution f $(0 \le f \le 1)$ and the laser beam radius w_0. The laser beam is directed top to bottom with local direction $\mathbf{s} = \mathbf{s}(\mathbf{x}, t)$ $(|\mathbf{s}| = 1)$ of the Poynting vector $\mathbf{S} = I_0 f \mathbf{s}$. The metal sheet moves with the velocity v_0 in negative x-direction, $\mathbf{v_0} = v_0 \mathbf{e}_x$. The absorbed heat flux $q_a = -A_p\, \mathbf{n} \cdot \mathbf{S}$ depends on the degree of absorption A_p, where $\mathbf{n} = \mathbf{n}(\mathbf{x}, t)$ is normal with respect to the absorption front $\Gamma_+(t)$. The absorption front $\Gamma_+(t)$ can be subdivided into two regions: the melt front $\Gamma_m(t)$, where the temperature equals the melting point $T = T_m$ and erosion takes place, and the rest $\Gamma_+(t)\backslash\Gamma_m(t)$ $(T < T_m)$. The melt front $\Gamma_m(t)$ is the *free boundary* of the one-phase problem.

For the numerical solution of the 2D one-phase problem in the x, z-plane, we introduce the scalings

$$t = \frac{w_0}{v_0}\tau, \quad x = w_0\xi, \quad z = w_0\zeta, \quad \theta = \frac{T - T_a}{T_m - T_a} \tag{4}$$

where T_a and T_m are the ambient- and the melting temperatures. If ζ_R denotes the Rayleigh range, the Gaussian beam width is given by

$$w(\zeta, \zeta_0) = w_0 \sqrt{1 + \left(\frac{\zeta + \zeta_0}{\zeta_R}\right)^2} \tag{5}$$

With these definitions the one-phase Free Boundary Problem has the following form: Find the solution of the heat conduction equation that reads in non-dimensional form

$$\frac{\partial\theta}{\partial\tau} - \frac{\partial\theta}{\partial\xi} - \frac{1}{Pe}\left(\frac{\partial^2\theta}{\partial\xi^2} + \frac{\partial^2\theta}{\partial\zeta^2}\right) = 0, \quad \mathbf{x} = (\xi,\zeta) \in \Omega(\tau) \tag{6}$$

with the boundary conditions

$$Q_a - \left(\frac{\partial\theta}{\partial\xi}n_\xi + \frac{\partial\theta}{\partial\zeta}n_\zeta\right) = 0, \quad \mathbf{x} \in \Gamma_+(\tau) \setminus \Gamma_m(\tau), \tag{7}$$

$$\frac{\partial\theta}{\partial\xi}n_\xi + \frac{\partial\theta}{\partial\zeta}n_\zeta = 0, \quad \mathbf{x} \in \Gamma_-(\tau), \tag{8}$$

$$\theta|_{\xi\to\infty} = 0, \quad \mathbf{x} \in \Omega(\tau). \tag{9}$$

In (6), Pe is the dimensionless Péclet number relating the rate of advection of a flow to its rate of diffusion. In (7), Q_a is the dimensionless heat flux. The position of the melt front $\mathbf{x}(\tau) = (\xi(\tau),\zeta(\tau)) \in \Gamma_m(\tau)$ is determined by the velocity $\mathrm{u}_p^{(m)}$ which we get from the Stefan condition

$$Q_a - \left(\frac{\partial\theta}{\partial\xi}n_\xi + \frac{\partial\theta}{\partial\zeta}n_\zeta\right) = Pe\, h_m \mathrm{u}_p^{(m)}, \quad \mathbf{x} \in \Gamma_m(\tau), \tag{10}$$

$$\theta = 1, \quad \mathbf{x} \in \Gamma_m(\tau). \tag{11}$$

The velocity $\mathrm{u}_p^{(m)}$ is normal with respect to the melt front $\Gamma_m(\tau) \subset \partial\Omega$. In (10), h_m is the inverse Stefan-number.

The dimensionless heat flux Q_a is the product of the cosine μ of the angle of incidence φ, the absorption coefficient $A(\mu)$, the maximum intensity $\gamma(\tau)$ and the distribution $f(\mathbf{x})$ of the laser radation:

$$Q_a = \mu\, A(\mu)\, \gamma(\tau)\, f(\mathbf{x}), \quad \mathbf{x} \in \Gamma_+(\tau) \tag{12}$$

In comparison with the well known Stefan problem, as represented by Elliot et al. [3] or Fasano et al. [5], there is a different and more complicated situation here, as the energy transfer takes place directly at the free boundary and as the heat flux absorbed at the free boundary depends on the angle of incidence and via the intensity distribution $f(\mathbf{x})$ on the position.

3 Numerical Solution of the One-Phase Problem

To fulfil the Stefan condition (10), we compute the velocity $\mathrm{u}_p^{(m)}$ from (10) for each node on the melt front. Multiplied by the time step width $\Delta\tau$ we get the new position of the nodes. But after some time steps the melt front becomes non-smooth and the

errors become large. So we have to stabilize the melt front numerically during the computation. This stabilization is done by the approximation of the melt front by a polynomial of order q_m by the method of least squares. We choose $q_m = 5$, as it gives the best results.

In each time step, we have to carry out the so-called grid iteration: In each iteration step, we first compute the new temperature distribution on the given grid. With the new temperature values we compute the values $u_p^{(m)}$ in each node of the melt front and by $u_p^{(m)}$ we get the new position of the melt front. Then the nodes on each grid line $\zeta = const$ are distributed equidistantly between the melt front and the right end of the metal sheet. On this new grid, we again compute the temperature and so on. The grid iteration is stopped if the grid does not move anymore, and we continue with the next time step. If the maximum displacement of the melt front falls below a bound $\Delta s = 10^{-6}$, the melt front is stable in time, and the whole computation is stopped.

We carry out the computations on two distributed memory parallel computers: The first one is the HP XC4000 with 2.6 GHz AMD Opteron processors and InfiniBand 4X interconnect installed at the Steinbuch Centre for Computing of the Karlsruhe Institute of Technology, Germany, the second one is the SGI Altix 4700 with 1.6 GHz Intel Itanium2 Montecito Dual Core processors and NUMAlink 4 interconnect at the Leibniz Supercomputing Centre in Garching, Germany.

The solution domain Ω is characterized by $\Delta = 4\,\text{mm}/w_0$ and $\xi_r = 16\,\text{mm}/w_0$. The starting form of the melt front is a parabola through the nodes $(0.6\,\text{mm}/w_0, 0)$ and $(0, \Delta)$ with a slope of zero at $\zeta = \Delta$.

The grid we use is of the following type: We have horizontal grid lines with $\zeta = const$ where the distance between the lines Δh_ζ becomes the smaller the more we approach $\zeta = 0$. The nodes are distributed equidistantly in ξ-direction between the melt front and the right boundary on each line $\zeta = const$.

We use a grid with $481 \times 161 = 77{,}441$ nodes and 153,600 elements for the structuring of the space for each computation, we compute on 8 processors and we compute with consistency order $q = 2$.

We solve the laser cutting problem with three different parameter sets where we vary the focal radius w_0, the sheet velocity v_0 and the power of the laser beam P_L, see Table 1.

In Table 2 we present the results of the three computations. In the first column we give the maximum of the temperature θ_{max}, in the second and third column you see the maximum and the mean relative estimated error, respectively. In the fourth column we give the number of iteration steps of all time steps, and in the last two

Table 1 Parameter sets for the stationary one-phase problem

Symbol	Value for parameter set			Unit	Description
	1	2	3		
w_0	600	600	100	μm	Focal radius
v_0	0.025	0.005	1.0	m s^{-1}	Sheet velocity
P_L	2	1	20	kW	Power of laser beam

Table 2 Results for the stationary one-phase problem for the three parameter sets of Table 1

Set	θ_{max}	Relative est. error		No. of it. steps	CPU time [s]	
		Max.	Mean		HP XC4000	SGI Altix 4700
1	1.001	$0.218 \cdot 10^{-1}$	$0.420 \cdot 10^{-4}$	81	300.35	285.93
2	1.000	$0.460 \cdot 10^{-2}$	$0.318 \cdot 10^{-4}$	44	164.09	164.67
3	1.889	$0.374 \cdot 10^{0}$	$0.794 \cdot 10^{-3}$	268	978.58	968.00

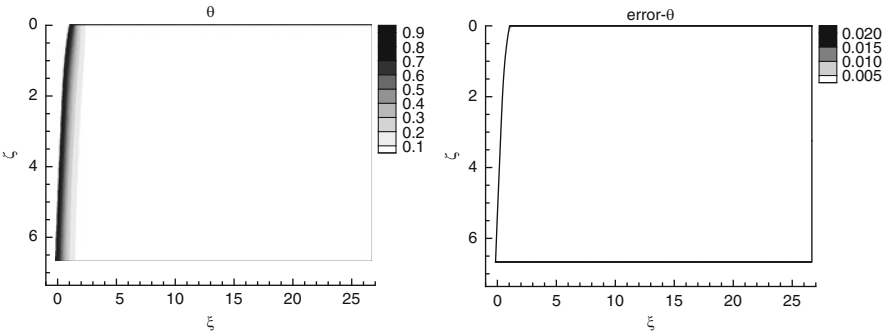


Fig. 2 Contour plot of the temperature and its error for parameter set 1

columns we give the CPU time of master processor 1 for the computations on the HP XC4000 and the SGI Altix 4700, respectively.

For the first two parameter sets, the maximum temperature is quite close to the expected maximum of $\theta_{max} = 1$ which is the boundary condition on the melt front. The maximum errors are 2.2% and 0.5%, respectively, the mean errors are very small. The third parameter set is extremely critical: We choose a small focal radius, a high velocity of the metal sheet and a high laser power. The maximum temperature is 1.889 which is totally wrong. The error estimate clearly shows this, but the errors are only local ones as the mean error is very small for this set, too.

For this parameter set we carried out three more computations on finer grids with $961 \times 321 = 308,481$ nodes, $1921 \times 641 = 1,231,361$ nodes and $3841 \times 1281 = 4,920,321$ nodes, i.e., we halved the mesh size in ξ- and ζ-direction from one grid to the other. Then it holds $\theta_{max} = 1.303$, $\theta_{max} = 1.044$ and $\theta_{max} = 1.012$, respectively. The maximum relative estimated errors are 20.7%, 6.9% and 1.9% for the three computations, and the mean errors are 0.0093%, 0.0011% and 0.0003%.

Figure 2 shows the temperature plot for parameter set 1. The relative estimated error is shown in the right picture of Fig. 2 from which you can see that the maximum errors only occur at the upper left corner. The temperature goes down from $\theta = 1$ at the melt front to zero very fast. The maximum error of 2.2% occurs at the node on the upper boundary next to the left upper corner. There is a region with some larger errors at the upper part of the melt front, but the mean error is 0.0042%.

The contour plots for the parameter sets 2 and 3 must be omitted because of space limitations here, but they are presented in [2]. Table 2 also shows that the total number of iteration steps of all time steps is quite different. The more critical

the problem is, the higher is the number of iteration steps. For the most uncritical parameter set 2, the computation stops after 44 iteration steps, whereas we need 268 iteration steps for the most critical set 3 which is almost a factor of six. The form of the melt front at the end of each computation is completely different.

4 Modulation of the Gaussian Beam

Modulation of the Gaussian beam means that the intensity of the Gaussian beam is additionally varied spatially and temporally:

$$Q_a = \mu\, A(\mu)\, \gamma(\tau)\, (f(\mathbf{x}) + f_m(\tau, \mathbf{x})), \quad \mathbf{x} \in \Gamma_m(\tau) \tag{13}$$

with

$$f_m(\tau, \mathbf{x}) = \frac{1}{2} \frac{\widetilde{A}_m}{\tilde{w}^2(\zeta)} \left(1 + \sin\left(2\pi v_m \tau + \frac{3}{2}\pi\right)\right) \sin^2(\pi \xi_m) \tag{14}$$

where $\tilde{w} = \frac{w(\zeta)}{w_0}$ and v_m is the dimensionless modulation frequency. The spatial variation is restricted on the interval $\xi \in [\xi_1, \xi_1 + b_1]$, i.e., only there it holds $\xi_m > 0$ and the amplitude $\widetilde{A}_m > 0$. For the computation we choose parameter set 1, but we change the focal radius to $w_0 = 300\,\mu\text{m}$. First we compute the solution of the steady state one-phase problem. The maximum temperature is $\theta_{max} = 1.01$, the maximum and mean relative estimated errors are 1.3% and 0.0024%, respectively. After this computation we have initial values at $\tau = 0$ for the temperature θ in each node and an initial form of the melt front for the computation with the modulated Gaussian beam. Afterwards we want to compute 10 periods of modulation. We use the scale $t_0 = t/\tau = w_0^2/\kappa$ for the dimensionless time τ where κ is the thermal diffusivity. The dimensionless computation time period is $\tau = 11.\bar{1}$ which corresponds to $t = 0.1$ s.

We carried out some computations with different time step widths $\Delta\tau$. We show the position of the upper melt front node ($\zeta = 0$) for the time step widths 0.0001, 0.0002, 0.0004 and 0.001 in Fig. 3. We can clearly see that the choice of the time step width is quite important for the result, at least for the upper melt front node. The maximum and mean relative estimated error does not change if we change the time step width. So our error estimate does not detect the wrong movement of the melt

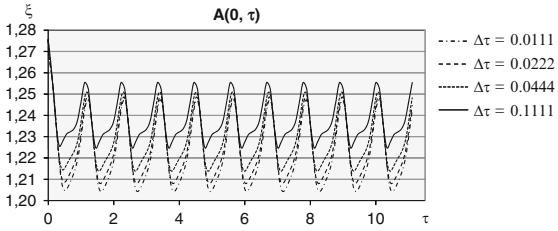


Fig. 3 Melt front position $A(0, \tau)$ for modulated Gaussian beam and different time step widths $\Delta\tau$

front. We can only compute the space discretization error of the elliptic problem on the given grid in each time step.

5 Concluding Remarks

We simulated numerically the cutting of a metal sheet by a laser beam. The metal sheet moves with constant velocity beneath a laser beam that is perpendicular to the sheet. For the first naive approach, our error estimate showed us that things go wrong, even before the solution became obviously wrong. So the numerical algorithm for the position of the melt front was unstable. We tried several strategies to overcome the problems but all attempts were in vain. Finally we found out how we can stabilize the melt front numerically: by the introduction of approximation polynomials that we compute by the method of least squares. We solved the laser cutting problem for three more or less critical parameter sets. Additionally, we modulated the laser beam spatially and temporally for a fourth parameter set. Here we saw that it is very important to choose a time step width small enough to get the correct movement of the melt front in time.

References

1. Adolph, T. (2005): The Parallelization of the mesh refinement algorithm in the finite difference element method. Doctoral Thesis. http://www.rz.uni-karlsruhe.de/rz/docs/FDEM/Literatur/par_mra_fdem.pdf
2. Adolph, T., Schönauer, W., Niessen, A. and Schulz, W. (2009): Numerical simulation of laser beam cutting: One-phase problem. Tech. report, Steinbuch Centre for Computing, Karlsruhe Institute of Technology
3. Elliot C.M., Ockendon J.R. (1982): Weak and variational methods for moving boundary problems. Pitman, Boston
4. Enß, V., Kostrykin, V., Schulz, W., Zimmermann, C., Zefferer, H., Petring, D. (1997): Thermal treatment using laser radiation: Laser beam fusion cutting. In: Hoffmann, K.-H., Jäger, W., Lohmann, Th., Schunk, H. (eds.) Mathematik - Schlüsseltechnologie für die Zukunft, pp. 161–174. Springer, Berlin
5. Fasano, A., Primicerio, M. (1977): General free-boundary value problems for the heat equation, I, II, III. J. Math. Anal. Appl. **57**, 694–723; **58**, 202–231; **59**, 1–14
6. LINSOL. http://www.rz.uni-karlsruhe.de/rd/linsol.php
7. Schönauer, W., Adolph, T. (2005): FDEM: The evolution and application of the finite difference element method (FDEM) program package for the solution of partial differential equations, Abschlussbericht des Verbundprojekts FDEM, Universität Karlsruhe. http://www.rz.uni-karlsruhe.de/rz/docs/FDEM/Literatur/fdem.pdf

Numerical Optimization of a Bioreactor for the Treatment of Eutrophicated Water

Lino J. Alvarez-Vázquez, Francisco J. Fernández, and Aurea Martínez

Abstract The fundamental idea of a bioreactor consists of holding up eutrophicated water (rich, for instance, in nitrogen) in large tanks where we add a certain quantity of phytoplankton, that we let grow in order to absorb nitrogen and purify water. Its optimal management can be formulated as an optimal control problem with state constraints, where the control can be the quantity of phytoplankton added at each tank or the permanence times, the state variables are the concentrations of the species involved (nutrient, phytoplankton, zooplankton and organic detritus), the objective function to be minimized is the phytoplankton concentration of water leaving the bioreactor, and the state constraints stand for the thresholds imposed on the nitrogen and detritus concentrations in each tank. We recall that this optimal control problem possesses a solution, which can be characterized by a first order optimality condition. After discretizing the control problem, we present a structured algorithm for solving the resulting nonlinear constrained optimization problem. Finally, we also give numerical results for a real-world example.

1 Introduction

Eutrophication is a process of nutrient enrichment (usually by nitrogen and/or phosphorus) in large waterbodies such that the productivity of the system ceases to be limited by the availability of nutrients. It occurs naturally over geological time, but may be accelerated by human activities (e.g., sewage or land drainage). In this work we deal with a model governing eutrophication processes (based on a system of

L.J. Alvarez-Vázquez (✉), A. Martínez
Departamento de Matemática Aplicada II, E.T.S.I. Telecomunicación, Universidad de Vigo, 36310 Vigo, Spain
e-mail: lino@dma.uvigo.es, aurea@dma.uvigo.es

F. J. Fernández
Departamento de Matemática Aplicada, Facultad de Matemáticas, Universidad de Santiago de Compostela, 15706 Santiago, Spain
e-mail: fjavier.fernandez@usc.es

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_7,

nonlinear parabolic partial differential equations with a great complexity) where a complete set of four species is analyzed: nutrient, phytoplankton, zooplankton and organic detritus.

The basis of the simplest bioreactors consists of holding up over-nitrogenated water in large tanks where we add a certain quantity of phytoplankton, that we let freely grow to absorb nitrogen from water. In the particular problem analyzed in this work we have considered only two large shallow tanks with the same capacities (but possibly different geometries). Water rich in nitrogen fills the first tank Ω_1, where we add an initial quantity ψ^1 and/or a distributed quantity ρ^1 of phytoplankton (which we let grow for a time period T^1) to drop nitrogen level down to a desired threshold. Additionally, we are also interested in obtaining a certain quantity of organic detritus (very valued as agricultural fertilizer) in this first tank. Once the desired levels of nitrogen and organic detritus have been reached (and the detritus have been reclaimed for agricultural use after settling in the bottom of the tank), we drain this first tank and transfer water to the second tank Ω_2, where the same operation is repeated, by adding new amounts ψ^2 and/or ρ^2 of phytoplankton. Water leaving this second fermentation tank after a permanence time T^2 will be usually poor in nitrogen, but rich in detritus (settled in the bottom) and phytoplankton (recovered from a final filtering). At this point, we are interested – for economic/ecological reasons - in minimizing this final quantity of phytoplankton. Thus, the optimal control problem consists of finding the minimal permanence times and the minimal quantities of phytoplankton that we must add to each tank, so that the nitrogen levels be lower that maximum thresholds and detritus levels be higher that minimum thresholds, and in such a way that the final phytoplankton concentration be as reduced as possible.

From a mathematical viewpoint, this problem can be formulated as an optimal control problem with state and control constraints, where the controls are $(\psi^1, \psi^2, \rho^1, \rho^2, T^1, T^2)$, the state variables are the concentrations of species inside the tanks, the objective function is related to final phytoplankton concentration, the state constraints stand for the thresholds imposed on the nitrogen and detritus concentrations in each tank, and the control constraints are technological bounds.

2 The Control Problem

Most accurate mathematical models for the simulation of eutrophication processes are based in systems of partial differential equations with a high complexity due to the large number of species involved and to the great variety of internal phenomena appearing in the processes. In this paper we consider a realistic model with four biological variables involved (the formulation of the biochemical interaction terms and their meaning can be found, for instance, in Drago et al. [2]). So, we consider the state $\mathbf{u} = (u^1, u^2, u^3, u^4)$, where u^1 stands for a generic nutrient concentration (for instance, nitrogen), u^2 for phytoplankton concentration, u^3 for zooplankton concentration, and u^4 for organic detritus concentration.

The interactions of these four species into a given still water domain $\Omega \subset \mathbf{R}^3$ and over a time interval $I = (0, T)$ can be described by the following system of coupled partial differential equations for reaction-diffusion with Michaelis-Menten kinetics:

$$\begin{cases} \frac{\partial u^1}{\partial t} - \nabla \cdot (\mu_1 \nabla u^1) + C_{nc} L \frac{u^1}{K_N + u^1} u^2 - C_{nc} K_r u^2 - C_{nc} K_{rd} \Theta^{\theta - 20} u^4 = g^1 \\ \frac{\partial u^2}{\partial t} - \nabla \cdot (\mu_2 \nabla u^2) - L \frac{u^1}{K_N + u^1} u^2 + K_r u^2 + K_{mf} u^2 + K_z \frac{u^2}{K_F + u^2} u^3 = g^2 \\ \frac{\partial u^3}{\partial t} - \nabla \cdot (\mu_3 \nabla u^3) - C_{fz} K_z \frac{u^2}{K_F + u^2} u^3 + K_{mz} u^3 = g^3 \\ \frac{\partial u^4}{\partial t} - \nabla \cdot (\mu_4 \nabla u^4) - K_{mf} u^2 - K_{mz} u^3 + K_{rd} \Theta^{\theta - 20} u^4 = g^4 \end{cases} \tag{1}$$

in $Q = I \times \Omega$, with suitable boundary conditions on $\Sigma = I \times \partial\Omega$ and initial conditions $\mathbf{u}_0$ in Ω, and where $\theta(t, x)$ is the water temperature (in Celsius), $L(t, x)$ is the luminosity function (related to incident light intensity and phytoplankton growth rate), μ_i, $i = 1, \ldots, 4$, are the diffusion coefficients of each species, C_{nc} is the nitrogen-carbon stoichiometric relation, Θ is the detritus regeneration thermic constant, K_N and K_F are the nitrogen and phytoplankton half-saturation constants, K_{mf} and K_{mz} are the phytoplankton and zooplankton death rates (including predation), K_{rd} is the detritus regeneration rate, K_r is the phytoplankton endogenous respiration rate, K_z is the zooplankton predation (grazing), and C_{fz} is the grazing efficiency factor. Existence and uniqueness of solution for above system (1) have been obtained by the authors in a recent paper [1].

To present a simpler expression for the state system (1), we consider the mapping $\mathbf{A} = (A^1, A^2, A^3, A^4) : \mathbf{R}_+ \times \Omega \times \mathbf{R}^4_+ \longrightarrow \mathbf{R}^4$, given by:

$$\mathbf{A}(t, x, \mathbf{u}) = \begin{bmatrix} -C_{nc} \left[L(t, x) \frac{u^1}{K_N + u^1} u^2 - K_r u^2 \right] + C_{nc} K_{rd} \Theta^{\theta(t,x) - 20} u^4 \\ \left[L(t, x) \frac{u^1}{K_N + u^1} u^2 - K_r u^2 \right] - K_{mf} u^2 - K_z \frac{u^2}{K_F + u^2} u^3 \\ C_{fz} K_z \frac{u^2}{K_F + u^2} u^3 - K_{mz} u^3 \\ K_{mf} u^2 + K_{mz} u^3 - K_{rd} \Theta^{\theta(t,x) - 20} u^4 \end{bmatrix} \tag{2}$$

Thus, the state system (1) can be rewritten in the following equivalent way:

$$\frac{\partial u^i}{\partial t} - \nabla \cdot (\mu_i \nabla u^i) = A^i(t, x, \mathbf{u}) + g^i \quad \text{in } Q, \quad \text{for } i = 1, \ldots, 4. \tag{3}$$

As proved by the authors in [1], the eutrophication system (3) admits a solution under non-smooth hypotheses: assuming the temperature $\theta \in L^2(Q)$ to be bounded, then the system (3) admits a unique solution $\mathbf{u} \in L^2(I; [H^1(\Omega)]^4) \cap [L^\infty(Q)]^4$. Moreover, solution $\mathbf{u}$ is nonnegative and bounded by a value only depending on time T, on second member $\mathbf{g} = (g^1, g^2, g^3, g^4)$, and on initial-boundary conditions.

With these notations we can formulate the bioreactor control problem with the following elements:

- **The controls:** We will control the system by means of three different types of design variables: the quantities $(\psi^1(x), \psi^2(x))$ of phytoplankton added in both tanks at initial times, the permanence times (T^1, T^2) of water inside both tanks, and the quantities $(\rho^1(t,x), \rho^2(t,x))$ of phytoplankton added in both tanks over the respective time intervals $I_1 = (0, T^1)$ and $I_2 = (0, T^2)$.
- **The state systems:** We consider two state systems giving the concentrations of nitrogen-phytoplankton-zooplankton-organic detritus in each tank. Since both tanks are isolated, no transference for any of the four species is considered through the boundaries (i.e., Neumann boundary conditions are assumed to be null). Both systems will be coupled by means of the initial/final conditions: when water is transferred from the first tank to the second one, it is natural to assume that water is mixed up, and this is the reason of considering the initial conditions for the concentrations inside the second tank as given by the corresponding averaged final concentrations in the first tank. These two state systems are given by:
  - *First tank Ω_1:* The state variables for the first tank will be denoted $\mathbf{u}^1 = (u^{1,1}, u^{2,1}, u^{3,1}, u^{4,1})$ with $u^{1,1}$ (nitrogen), $u^{2,1}$ (phytoplankton), $u^{3,1}$ (zooplankton), and $u^{4,1}$ (organic detritus). The permanence time of water inside this first tank will be T^1, and the initial concentrations will be given by $\mathbf{u}_0^1 = (u_0^{1,1}, u_0^{2,1}, u_0^{3,1}, u_0^{4,1})$. Thus, we have the system, for $i = 1, \ldots, 4$:

$$\begin{cases} \frac{\partial u^{i,1}}{\partial t} - \nabla \cdot (\mu_i \nabla u^{i,1}) = A^i(t, x, \mathbf{u}^1) + \delta_{2i}\rho^1 & \text{in } I_1 \times \Omega_1, \\ \frac{\partial u^{i,1}}{\partial n} = 0 & \text{on } I_1 \times \partial\Omega_1, \\ u^{i,1}(0) = u_0^{i,1} + \delta_{2i}\psi^1 & \text{in } \Omega_1, \end{cases} \tag{4}$$

    where δ_{2i} denotes the Kronecker's delta ($\delta_{2i} = 1$ if $i = 2, \delta_{2i} = 0$ otherwise).
  - *Second tank Ω_2 :* The state variables for the second tank will be now $\mathbf{u}^2 = (u^{1,2}, u^{2,2}, u^{3,2}, u^{4,2})$. The permanence time of water inside this second tank will be T^2. Thus, we have the similar system, for $i = 1, \ldots, 4$:

$$\begin{cases} \frac{\partial u^{i,2}}{\partial t} - \nabla \cdot (\mu_i \nabla u^{i,2}) = A^i(t, x, \mathbf{u}^2) + \delta_{2i}\rho^2 & \text{in } I_2 \times \Omega_2, \\ \frac{\partial u^{i,2}}{\partial n} = 0 & \text{on } I_2 \times \partial\Omega_2, \\ u^{i,2}(0) = M^i(\mathbf{u}^1(T^1)) + \delta_{2i}\psi^2 & \text{in } \Omega_2, \end{cases} \tag{5}$$

    where $\mathbf{M} = (M^1, M^2, M^3, M^4)$ is given by:

$$\mathbf{M}(\mathbf{v}) = \frac{1}{\text{meas}(\Omega_1)} \begin{bmatrix} \int_{\Omega_1} v^1 dx \\ \int_{\Omega_1} v^2 dx \\ \int_{\Omega_1} v^3 dx \\ 0 \end{bmatrix} \tag{6}$$

- **The objective function:** Since we are interested in reducing the total processing time $T^1 + T^2$ and the quantities $(\psi^1, \psi^2, \rho^1, \rho^2)$ of phytoplankton added to both tanks, and also in minimizing the final phytoplankton concentration of water leaving the second tank, we are led to consider the following quadratic cost functional J given by:

$$J(\psi^1, \psi^2, \rho^1, \rho^2, T^1, T^2) = \frac{N_1}{2}\Big[\int_{\Omega_1} (\psi^1)^2 dx + \int_{\Omega_2} (\psi^2)^2 dx\Big] \tag{7}$$
$$+\frac{N_2}{2}\Big[\int_0^{T^1}\int_{\Omega_1} (\rho^1)^2 dx dt + \int_0^{T^2}\int_{\Omega_2} (\rho^2)^2 dx dt\Big] + \frac{N_3}{2}[(T^1)^2 + (T^2)^2]$$
$$+\frac{N_4}{2}\frac{1}{\text{meas}(\Omega_2)}\int_{\Omega_2} (u^{2,2}(T^2))^2 dx$$

with $N_1, \ldots, N_4 \geq 0$ weight parameters.

- **The state constraints:** Final nitrogen concentration in each tank must be lower than a given threshold, and final organic detritus concentration in each tank must be greater than another threshold. These constraints translate into:

$$\begin{cases} C^1(\psi^1, \psi^2, \rho^1, \rho^2, T^1, T^2) = \frac{1}{\text{meas}(\Omega_1)}\int_{\Omega_1} u^{1,1}(T^1) dx \leq \sigma_1, \\ C^2(\psi^1, \psi^2, \rho^1, \rho^2, T^1, T^2) = \frac{1}{\text{meas}(\Omega_2)}\int_{\Omega_2} u^{1,2}(T^2) dx \leq \sigma_2, \\ C^3(\psi^1, \psi^2, \rho^1, \rho^2, T^1, T^2) = \frac{1}{\text{meas}(\Omega_1)}\int_{\Omega_1} u^{4,1}(T^1) dx \geq \theta_1, \\ C^4(\psi^1, \psi^2, \rho^1, \rho^2, T^1, T^2) = \frac{1}{\text{meas}(\Omega_2)}\int_{\Omega_2} u^{4,2}(T^2) dx \geq \theta_2, \end{cases} \tag{8}$$

for certain given values $\sigma_1,\ \sigma_2,\ \theta_1,\ \theta_2 > 0$.

- **The control constraints:** For technological reasons, permanence times (T^1, T^2) must range between two fixed bounds $0 < T_{min} < T_{max} < \infty$, and the quantities of added phytoplankton $(\psi^1, \psi^2, \rho^1, \rho^2)$ must be nonnegative and bounded by a maximal admissible value M. So, control $(\psi^1, \psi^2, \rho^1, \rho^2, T^1, T^2)$ must lie in

$$\begin{aligned} \mathcal{U}_{ad} = \{&(\psi^1, \psi^2, \rho^1, \rho^2, T^1, T^2) \\ &\in L^2(\Omega_1) \times L^2(\Omega_2) \times L^2((0, T_{max}) \times \Omega_1) \times L^2((0, T_{max}) \times \Omega_2) \times \mathbf{R}^2 : \\ &0 \leq \psi^j(x_j) \leq M \ \text{ a.e. } x_j \in \Omega_j, \quad T_{min} \leq T^j \leq T_{max},\ j = 1, 2, \\ &0 \leq \rho^j(t_j, x_j) \leq M \ \text{ a.e. } (t_j, x_j) \in (0, T_{max}) \times \Omega_j,\ j = 1, 2.\} \end{aligned}$$

which is a closed, bounded, convex, and nonempty subset of $L^2(\Omega_1) \times L^2(\Omega_2) \times L^2((0, T_{max}) \times \Omega_1) \times L^2((0, T_{max}) \times \Omega_2) \times \mathbf{R}^2$.

Thus, the formulation of the optimal control problem, denoted by $(\mathcal{P})$, will be:

$$(\mathcal{P}) \quad \inf\big\{J(\psi^1, \psi^2, \rho^1, \rho^2, T^1, T^2) \ \text{ such that } (\psi^1, \psi^2, \rho^1, \rho^2, T^1, T^2) \in \mathcal{U}_{ad}$$
$$\text{and } (\mathbf{u}^1, \mathbf{u}^2) \text{ satisfies (4)-(5) and (8)}\big\}$$

The existence of optimal solutions (and derivation of first order optimality conditions for their characterization) have been proved in [3].

3 The Discretized Problem

As a first step in our process, we need to numerically approximate the unique (nonnegative and bounded) solution of the general eutrophication system (3) in $Q = (0, T) \times \Omega$. In order to do this we use a first order implicit time discretization (based in a finite difference scheme), and a standard space discretization based in the Lagrange finite element method.

So, for the time semi-discretization we will consider a finite set of discrete times $\{t_n\}_{n=0}^{N_T} \subset [0, T]$ such that $t_0 = 0$, $t_{N_T} = T$, and $t_n - t_{n-1} = \Delta t$ for all $n = 1, \ldots, N_T$, with a time step $\Delta t > 0$. Associated to above set we construct the following time semi-discretization of the state system (3), with $\alpha = \frac{1}{\Delta t} > 0$:

$$\begin{cases} \alpha \mathbf{u}_n - \nabla \cdot (\Lambda_\mu \nabla \mathbf{u}_n) = \mathbf{A}(t_n, x, \mathbf{u}_n) + \alpha \mathbf{u}_{n-1} + \mathbf{g}(t_n) & \text{in } \Omega, \\ \frac{\partial \mathbf{u}_n}{\partial n} = 0 & \text{on } \partial\Omega, \end{cases} \tag{9}$$

where Λ_μ is a diagonal matrix with diagonal elements $(\mu_1, \mu_2, \mu_3, \mu_4)$.

We can easily prove - by standard fixed point techniques - that under assumption:

$$\alpha > \max\{\|L\|_{L^\infty(Q)} - K_r - K_{mf}, C_{fz} K_z - K_{mz}\} \tag{10}$$

there exists a constant $C(\alpha, M)$ – only depending on α and M – such that the unique solution $\mathbf{u}_n \in [H^1(\Omega)]^4$ of (9) is nonnegative and bounded by $C(\alpha, M)$.

To deal with the nonlinear part $\mathbf{A}(t_n, x, \mathbf{u}_n)$ of the semi-discretized system (9) we propose for each discrete time $n = 1, \ldots, N_T$ a fixed point scheme that, again under assumption (10) on α, will be monotone and convergent for any initial iterate being nonnegative and bounded by $C(\alpha, M)$.

For the fully discretized formulation of the eutrophication system (3) – and due to the fact that we have used a first order time semi-discretization – we propose a standard P_1-Lagrange finite element method. Then, for the domain Ω (supposed to be polygonal), we consider a regular mesh $\mathcal{T}_h$ and the finite dimensional vector subspace $V_h = \{u \in \mathcal{C}(\overline{\Omega}) : u_{|_T} \in P_1(T), \forall T \in \mathcal{T}_h\} \subset H^1(\Omega)$, where $P_1(T)$ stands for the space of degree one polynomials on T.

If we denote N_h the number of nodes in the mesh $\mathcal{T}_h$, $\{b_j\}_{j=1}^{N_h}$ the set of nodes of the mesh $\mathcal{T}_h$, and $\{\phi_i\}_{i=1}^{N_h}$ the standard basis of the space V_h (i.e., $\phi_i(b_j) = \delta_{ij}$, $\forall i, j = 1, \ldots, N_h$), we have that any element $u_h \in V_h$ admits a unique representation $u_h = \sum_{i=1}^{N_h} u_h(b_i)\phi_i$. Thus, we can obtain the fully-discretized solution $\mathbf{u}_{h,n}$ of above system, again under assumption (10) on α (assuring now the positive definiteness of the matrices involved in linear systems giving the fully-discretized solution).

Once we can solve the state system, we proceed to discretize the controls, the cost function and the constraints in a direct way (using trapezoidal rule for integrals), arriving to a nonlinear constrained optimization problem ($\mathscr{P}_h$). In order to solve problem ($\mathscr{P}_h$) (providing us with a discrete approximation of the solution of our original optimal control problem ($\mathscr{P}$)) we propose an interior-point type method, due to its effectiveness in large size problems. In particular, we use the IPOPT code, an interior-point filter line-search algorithm for large-scale nonlinear programming recently developed by Wächter and Biegler [4]. To apply the IPOPT code to our problem, we have needed to write an interface linking the optimization code with our own C++ code for computing the states and their derivatives.

4 Numerical Results

In this final section we present one of the several numerical results obtained for a realistic example consisting of two tanks of equal dimensions $20\,m \times 20\,m \times 16\,m$ with the following physical coefficients $\mu_i = 2\ 10^{-3}\,m^2/s$, $i = 1, \ldots, 4$, $K_N = 2.8\ 10^{-2}\,mg/l$, $K_F = 2\ 10^{-1}\,mg/l$, $K_{mf} = 3.8\ 10^{-7}\,s^{-1}$, $K_{mz} = 3.78\ 10^{-7}\,s^{-1}$, $K_r = 3.8\ 10^{-7}\,s^{-1}$, $K_z = 2.3\ 10^{-6}\,s^{-1}$, $K_{rd} = 2.3\ 10^{-5}\,s^{-1}$, $C_{fz} = 6\ 10^{-1}$, $\Theta = 1.05$, $\theta = 19°C$. Moreover, the initial conditions for the first tank will be given by $u_0^{1,1} = 0.28\,mg/l$, $u_0^{2,1} = 0.01\,mg/l$, $u_0^{3,1} = 0.02\,mg/l$ and $u_0^{4,1} = 4.50\,mg/l$. For the other values related to state and control constraints, we have taken the thresholds $\sigma_1 = 0.23$, $\sigma_2 = 0.18$, $\theta_1 = 0.015$ and $\theta_2 = 0.04$ (i.e., we are imposing a nitrogen reduction of the 82% in the first tank and of the 64% in the second one), a phytoplankton upper bound $M = 100$, and time bounds $T_{min} = T_0/2$ and $T_{max} = 2T_0$ (that is, we are restricting our search to time variations among the half and the double of the starting time period).

Finally, for the numerical solution of the eutrophication systems, we have considered a tetrahedral regular mesh of both tanks formed by P_1 finite elements with a characteristic size of $1\,m$, and – for a starting time period of $T_0 = 135$ hours – a time step length of 3 hours (with a total number of 45 time steps).

In the present example we are only interested in controlling the initial quantities of phytoplankton added to both tanks and the permanence times into them, taking the distributed controls as fixed to zero. So, in the objective function J we have taken $N_1 = 1$, $N_2 = 0$, $N_3 = 1$ and $N_4 = 1$. Then applying our algorithm we have passed, in 22 iterations, from the initial controls $\Psi_0^1 = 2$, $\Psi_0^2 = 1$ and $T_0^1 = T_0^2 = 135$, to the optimal controls $\Psi^1 = 0.91$, $\Psi^2 = 0.93$, $T^1 = 77.05$ and $T^2 = 76.57$. Then, as a first consequence, we have obtained a significant time reduction for the whole process of about a 42%. The values showing the reduction of the objective function and the enforcement of the state constraints can be seen in the following table:

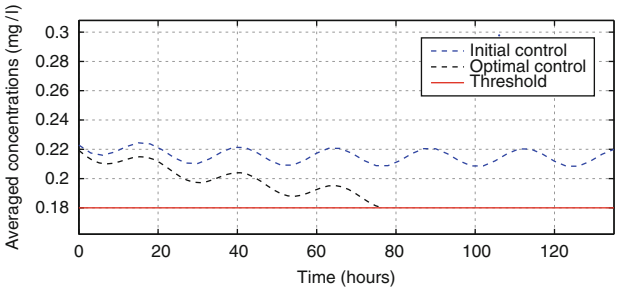


Fig. 1 Averaged concentrations of nitrogen in the second tank Ω_2

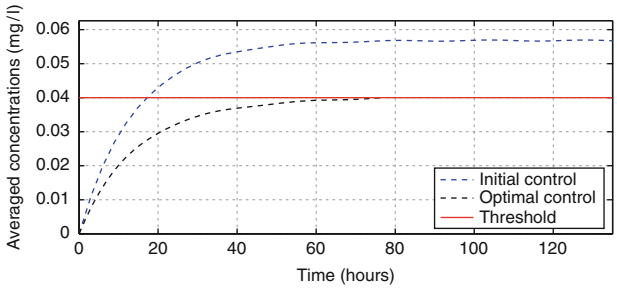


Fig. 2 Averaged concentrations of organic detritus in the second tank Ω_2

Element	Initial value	Optimal value	Threshold
J	$4.86007\,10^{5}$	$1.57356\,10^{5}$	
C^1	$2.22887\,10^{-1}$	$2.19282\,10^{-1}$	$2.3\,10^{-1}$
C^2	$2.20001\,10^{-1}$	$1.80000\,10^{-1}$	$1.8\,10^{-1}$
C^3	$3.98949\,10^{-2}$	$2.05899\,10^{-2}$	$1.5\,10^{-2}$
C^4	$5.67390\,10^{-2}$	$3.99999\,10^{-2}$	$4.0\,10^{-2}$

In Fig. 1 we show the averaged concentrations of nitrogen for the second tank, corresponding to the initial and the optimal controls. We can observe how, in the second tank, after the optimal permanence time $T^2 = 76.57$, the second constraint C^2 reaches the exact threshold σ_2. Moreover, daily oscillations due to night/day luminosity variations can be clearly identified.

In Fig. 2 we can see the averaged concentrations of organic detritus for the second tank, corresponding to the initial and the optimal controls. We can observe how, after the optimal permanence time T^2, the fourth constraint C^4 is also active (showing the optimality of the achieved solution).

Finally, in Fig. 3 we show the averaged concentrations of phytoplankton for the second tank, corresponding to the initial and the optimal controls. We must outstand the significant decrease of the value obtained for the controlled case, which stays always under the value obtained for the uncontrolled one.

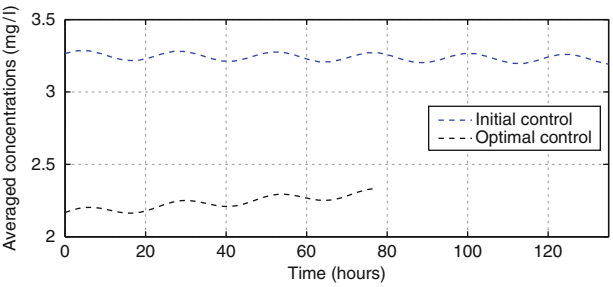


Fig. 3 Averaged concentrations of phytoplankton in the second tank Ω_2

Acknowledgements The financial support provided by Projects MTM2009-07749 of MICIIN (Spain), and INCITE09-291-083-PR of Xunta de Galicia is gratefully acknowledged.

References

1. Alvarez-Vázquez, L.J., Fernández, F.J., Muñoz-Sola, R.: Mathematical analysis of a three-dimensional eutrophication model. J. Math. Anal. Appl. **349**, 135–155 (2009)
2. Drago, M., Cescon, B., Iovenitti, L.: A three-dimensional numerical model for eutrophication and pollutant transport. Ecological Modelling **145**, 17–34 (2001)
3. Fernández, F.J.: Algunos problemas de control en procesos de eutrofizacion. PhD Thesis, Universidad de Santiago de Compostela (2008) (in Spanish)
4. Wächter, A., Biegler, L.T.: On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. Math. Program. **106**, 25–57 (2006)

Finite Element Approximation of a Quasi-3D Model for Estuarian River Flows

Mohamed Amara, Agnès Pétrau, and David Trujillo

Abstract We present here the main ideas and results concerning the derivation of a new quasi-3D hydrodynamical model, also called 2.5D model, within the framework of nonlinear weak formulations. The idea is to work in the sum of spaces concerning 2D models, one in the horizontal plane, the other in the vertical one. The new model takes into account the river's geometry and provides a three-dimensional velocity and pressure. We present the finite element approximation of the model and some numerical results.

1 Introduction

We are interested by the modeling and the numerical simulation of a quasi-3D river flow, within the context of hydrodynamical multidimensional modeling and simulation of estuarian river flows. The ideal model to be employed is a 3D one, but due to the huge computational cost, it cannot be used on the whole length of the river. Therefore, it is interesting to use different lower-dimensional models on adequated regions of the river.

In a previous work [1], new hydrodynamical models were proposed. One started from the physical time-discretized 3D problem, based on the instationary Navier–Stokes equations with physical boundary conditions, which was written in a nonlinear weak form. Then simpler models were derived by means of a projection method. In particular two bidimensional models were obtained, called 2D-horizontal and 2D-vertical models, either they are written on the free surface or on the median longitudinal surface of the river. A 1D model was also derived on the median curve of the river. All these models take into account the geometry of the river and provide

M. Amara (✉), A. Pétrau, and D. Trujillo
Laboratoire de Mathématiques Appliquées & CNRS UMR 5142, BP 1155, IPRA, Université de Pau et des Pays de l'Adour, 64013 Pau Cedex, France
e-mail: mohamed.amara@univ-pau.fr, agnes.petrau@etud.univ-pau.fr, david.trujillo@univ-pau.fr

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_8,

3D velocity and pressure (which is an unknown of the problem and not supposed to be hydrostatic).

In this paper we couple the 2D-horizontal and 2D-vertical models in order to build a quasi-3D model. This coupling allows us to get an intermediate model between the 2D and the 3D one, with good compromise when the 3D is too expensive.

The paper is organized as follows: in Sect. 2, the 3D physical problem is introduced and written in weak form, from which we next derive the two 2D models as conforming approximations. The choice of the projection subspaces is detailed in Sect. 3 with a brief mathematical study. Then we present in Sect. 4 the derivation of the quasi-3D model and its finite element approximation. Finally, the last section is devoted to the numerical results provided by this new model.

2 The 3D Physical Problem

In what follows, we agree to write the vector functions in bold letters and to denote the vector product by $\wedge$. The physical problem is described by the instationary incompressible Navier–Stokes equations, in a moving domain $\Omega_F(t) \subset \mathbf{R}^3$:

$$\begin{cases} \operatorname{div} \mathbf{u} = 0 \\ \rho \dfrac{\partial \mathbf{u}}{\partial t} + \rho\, \mathbf{curlu} \wedge \mathbf{u} + \mu\, \mathbf{curl}(\mathbf{curlu}) + \nabla p - \rho \mathbf{f} \wedge \mathbf{u} = \rho \mathbf{g} \end{cases}$$

The unknowns are the velocity $\mathbf{u}$ and the dynamic pressure $p = \tilde{p} + \frac{\rho}{2}|\mathbf{u}|^2$ where $\tilde{p}$ is the pressure of the Navier–Stokes problem. The density ρ, the viscosity μ, the gravity force $\mathbf{g} = (0, 0, -g)$, the earth's rotation velocity $\mathbf{f} = (0, 0, f)$ are given constants. We add the initial condition $\mathbf{u}(0) = \mathbf{u_0}$ to the previous system, as well as boundary conditions. For that, we decompose the boundary into three parts: $\partial\Omega_F(t) = \Gamma_B(t) \cup \Gamma_S(t) \cup \Gamma_I(t)$, where $\Gamma_B(t)$ denotes the riverbed, $\Gamma_S(t)$ the free surface and $\Gamma_I(t)$ the inflow and outflow boundaries. We impose the atmospheric pressure p_S and the tangential stresses corresponding to the wind force on the surface, the friction and impermeability conditions on the bottom, while on Γ_I we suppose that the flux and the tangential forces are known. These physical conditions translate into the following relations:

$$\begin{cases} \mathbf{u} \cdot \mathbf{n} = 0\,, & \mu\, \mathbf{curlu} \wedge \mathbf{n} = -c_B \mathbf{u} & \text{on } \Gamma_B(t), \\ p = p_S, & \mu\, \mathbf{curlu} \wedge \mathbf{n} = \mathbf{w} & \text{on } \Gamma_S(t), \\ \mathbf{u} \cdot \mathbf{n} = k, & \mu\, \mathbf{curlu} \wedge \mathbf{n} = \mathbf{w} & \text{on } \Gamma_I(t), \end{cases}$$

where $p_S, k, \mathbf{w}$ are given functions and the friction coefficient $c_B > 0$ is a given constant. We denote by $Z_B(x, y)$ the elevation of the bottom, given by the bathymetry and defined on a maximal domain $\Sigma \subset \mathbf{R}^2$. We also introduce the height of the

water $h(x, y, t)$ and the 2D domain $\Sigma_F(t) \subset \Sigma$, defined at each instant by $h > 0$:

$$\Sigma_F(t) = \{(x, y);\ h(x, y, t) > 0\}, \quad \Sigma_F(t) \subset \Sigma,\ \forall t > 0. \tag{1}$$

Then we have:

$$\Omega_F(t) = \{(x, y, z);\ (x, y) \in \Sigma_F(t),\ Z_B(x, y) \le z \le Z_B(x, y) + h(x, y, t)\}.$$

We close the system by adding the free surface equation (cf. for instance [3]):

$$\frac{\partial h}{\partial t} + \sum_{i=1}^{2} u_i \partial_i (h + Z_B) - u_3 = 0 \quad \text{on } \Gamma_S(t),$$

with an initial condition $h(0) = h_0$.

We thus get a 3D problem in the unknowns $(h, \mathbf{u}, p)$, which is next discretized with respect to time by means of an implicit scheme:

$$\frac{h - h^n}{\Delta t} + \sum_{i=1}^{2} u_i^n \partial_i (h + Z_B) - u_3^n = 0, \qquad \text{on } \Gamma_S(t^n), \tag{2}$$

$$\begin{cases} \text{div}\mathbf{u} = 0, \\ \dfrac{\rho}{\Delta t}(\mathbf{u} - \mathbf{u}^n) + \rho\, \mathbf{curlu} \wedge \mathbf{u} + \mu\, \mathbf{curl}(\mathbf{curlu}) \\ \quad + \nabla p - \rho \mathbf{f} \wedge \mathbf{u} = \rho \mathbf{g}, \end{cases} \quad \text{in } \Omega_F\left(t^{n+1}\right), \tag{3}$$

where $u = u^{n+1}$ and $p = p^{n+1}$. We agree to denote in what follows the domain occupied by the fluid at t^{n+1} by Ω_F. Let us next introduce the Hilbert spaces:

$$\begin{aligned} &\mathrm{M} = \mathrm{L}^2(\Omega_F), \\ &\mathbf{X} = \left\{\mathbf{v} \in \mathbf{H}(\text{div}, \mathbf{curl}; \Omega_F);\ \mathbf{v}_{|\Gamma_B} \in \mathbf{L}^2(\Gamma_B)\right\}, \\ &\mathbf{X}^0 = \{\mathbf{v} \in \mathbf{X};\ \mathbf{v} \cdot \mathbf{n} = 0 \text{ on } \Gamma_B \cup \Gamma_I\}, \\ &\mathbf{X}^* = \{\mathbf{v} \in \mathbf{X};\ \mathbf{v} \cdot \mathbf{n} = 0 \text{ on } \Gamma_B,\ \mathbf{v} \cdot \mathbf{n} = k \text{ on } \Gamma_I\}, \end{aligned}$$

which are endowed with the following norms:

$$\begin{aligned} \|\mathrm{q}\|_{\mathrm{M}} &= \|\mathrm{q}\|_{0,\Omega_F}, \\ \|\mathbf{v}\|_{\mathbf{X}}^2 &= \frac{1}{\Delta t}\|\mathbf{v}\|_{0,\Omega_F}^2 + \|\text{div}\mathbf{v}\|_{0,\Omega_F}^2 + \|\mathbf{curlv}\|_{0,\Omega_F}^2 + \|\mathbf{v}\|_{0,\Gamma_B}^2. \end{aligned}$$

Problem (3) can be written in weak form:

$$\begin{cases} \textit{Find } (\mathbf{u}, p) \in \mathbf{X}^* \times \mathrm{M} \\ \forall \mathbf{v} \in \mathbf{X}^0, \quad A(\mathbf{u}; \mathbf{u}, \mathbf{v}) + B(p, \mathbf{v}) = F^n(\mathbf{v}) \\ \forall q \in \mathrm{M}, \quad B(q, \mathbf{u}) = 0, \end{cases} \tag{4}$$

with:

$$
\begin{aligned}
A\left(\mathbf{w};\mathbf{u},\mathbf{v}\right) &= A_0\left(\mathbf{u},\mathbf{v}\right) + A_1\left(\mathbf{w};\mathbf{u},\mathbf{v}\right),\\
A_0\left(\mathbf{u},\mathbf{v}\right) &= \int_{\Omega_F} \frac{\rho}{\Delta t}\mathbf{u}\cdot\mathbf{v}\,d\Omega + \int_{\Omega_F} \mu\mathbf{curlu}\cdot\mathbf{curlv}\,d\Omega\\
&\quad + \int_{\Gamma_B} c_B\mathbf{u}\cdot\mathbf{v}\,d\gamma - \int_{\Omega_F} \rho\left(\mathbf{f}\wedge\mathbf{u}\right)\cdot\mathbf{v}\,d\Omega,\\
A_1\left(\mathbf{w};\mathbf{u},\mathbf{v}\right) &= \int_{\Omega_F} \rho\left(\mathbf{curlu}\wedge\mathbf{w}\right)\cdot\mathbf{v}\,d\Omega,\\
B\left(p,\mathbf{v}\right) &= -\int_{\Omega_F} p\ \mathrm{div}\mathbf{v}\,d\Omega,
\end{aligned}
$$

$$
F^n\left(\mathbf{v}\right) = \int_{\Omega_F} \rho\left(\frac{1}{\Delta t}\mathbf{u}^n + \mathbf{g}\right)\cdot\mathbf{v}\,d\Omega + \left\langle \mathbf{v}\wedge\mathbf{n}, \mathbf{w}\wedge\mathbf{n}\right\rangle_{\Gamma_S\cup\Gamma_I} - \left\langle \mathbf{v}\cdot\mathbf{n}, p_S\right\rangle_{\partial\Omega_F},
$$

and where $< \ \cdot,\cdot\ >_{\Gamma}$ stands for the duality product between $H_{00}^{1/2}(\Gamma)$ and $H^{-1/2}(\Gamma)$.

Theorem 1. *Assuming some classical regularities on the boundary data and that* $\mathbf{X}^0$ *is continuously embedded in* $\mathbf{L}^4(\Omega_F)$ *and compactly embedded in* $\mathbf{L}^2(\Omega_F)$*, problem (4) has at least a solution at any* t^{n+1}*. If the data is small enough, the solution is unique.*

Proof. We present here the main steps and we refer to [2] for the details of the proof. We apply a variant of Brouwer's theorem (cf. for instance [4], p. 280) to show existence and, under the usual hypothesis of small data, uniqueness of the solution. Let us note that due to the dependence on Δt of the continuity constants, the uniqueness holds for instance if Δt is sufficiently small. The main ingredients of the proof are the inf − sup condition for $B(\cdot,\cdot)$, the coercivity of $A_0(\cdot,\cdot)$ on the kernel $\mathbf{V}$ of $B(\cdot,\cdot)$, the sequentially weak-continuity of $A_1(\cdot;\cdot,\mathbf{v})$ on $\mathbf{V}$ and the fact that:

$$
A_1(\mathbf{v};\mathbf{v},\mathbf{v}) = 0, \quad \forall\mathbf{v}\in\mathbf{X}. \tag{5}
$$

Note that the inf − sup condition is written as follows, with c a positive constant independent of both the time and space discretizations:

$$
\inf_{q\in\mathrm{M}}\sup_{\mathbf{v}\in\mathbf{X}^0}\frac{B\left(q,\mathbf{v}\right)}{\|\mathbf{v}\|_{\mathbf{X}}\,\|q\|_M} \geq c\sqrt{\Delta t},
$$

□

3 The Two 2D Models: Derivation and Mathematical Analysis

We can now derive several semi-discretized models as conforming approximations of (4) on convenient subspaces $\mathbf{X}_a^0\times\mathrm{M}_a$ of $\mathbf{X}^0\times\mathrm{M}$. We obtain a 2D-horizontal model written on the free surface and a 2D-vertical one written on the median longitudinal

plane of the river. We do not explicit here the 1D model since it does not take place in the construction of the quasi-3D model.

2D Horizontal Model

The 2D-horizontal model is written on the 2D domain $\Sigma_F(t) \subset \Sigma$, defined in (1). One can see $\Sigma_F(t)$ as the horizontal projection of the free surface $\Gamma_S(t)$ on Σ. Then we look for the pressure and the velocity given by:

$$p(x,y,z) = p_S + (Z_B(x,y) + h(x,y,t) - z)\, P_H(x,y),$$
$$\mathbf{u}(x,y,z) = (\mathbf{u}_H(x,y), u_{3H}(x,y,z))^t,$$

where $\mathbf{u}_H = (u_{1H}(x,y), u_{2H}(x,y))^t$ and $u_{3H} = \mathbf{u}_H \cdot \nabla Z_B + (z - Z_B)\, U_{3H}(x,y)$. This choice guarantees a conforming approximation with respect to the semi-discretized 3D problem. The unknowns are h, P_H, $\mathbf{u}_H$ and U_{3H}, all functions of (x,y) defined on $\Sigma_F(t)$.

2D Vertical Model

The 2D vertical model is written in curvilinear coordinates, in order to better take into account the river's geometry. Let us first present the geometrical and physical frameworks in which the model holds. For this purpose, we introduce the median curve $C(t)$ of the free surface $\Gamma_S(t)$. Let us denote the projection of $C(t) \subset \mathbf{R}^3$ on the fixed plane Σ by $C \subset \mathbf{R}^2$ and admit that the curve C is independent of time, smooth and described by $\varphi : I = [s_0, s_1] \to C$ where s is the curvilinear abscissa. In the sequel, we shall employ the orthonormal basis $\{\boldsymbol{\tau}(s), \boldsymbol{\nu}(s), \mathbf{e}_3\}$ where $\{\boldsymbol{\tau}(s), \boldsymbol{\nu}(s)\}$ is the Frenet local basis in each point $\varphi(s) \in C$ and we shall denote the associated curvilinear coordinates by $\{s, l, z\}$. Then the 2D vertical model is written on the *plane* domain:

$$\omega_F(t) = \{(s,z);\ s \in I,\ Z_B \le z \le Z_B + h\}$$

Let us also introduce the curvature $r = r(s)$ of C and the mid-width of the river L.

The subspaces $\mathbf{X}_a \times M_a$, on which the 3D formulation (4) is now approximated, are built by taking the dynamic pressure and the velocity as:

$$p(s,z) = p_V(s,z),$$
$$\mathbf{u}(s,l,z) = \left((1 - l\, r(s))\, u_{1V}(s,z), \frac{l L'(s)}{L(s)} u_{1V}(s,z), u_{3V}(s,z) \right)^t.$$

This choice guarantees a conforming approximation with respect to the semi-discretized 3D problem. The unknowns are h, p_V, u_{1V} and u_{3V}, all functions of (s,z) and t, and defined on $\omega_F(t)$.

Analysis of Both Models

Any of the two previous models can be written in a weak form:

$$\begin{cases} \textit{Find } (\mathbf{u}_a, p_a) \in \mathbf{X}_a^* \times \mathrm{M}_a \\ \forall \mathbf{v} \in \mathbf{X}_a^0, \quad A(\mathbf{u}_a; \mathbf{u}_a, \mathbf{v}) + B(p_a, \mathbf{v}) = F_a^n(\mathbf{v}) \\ \forall q \in \mathrm{M}_a, \quad B(q, \mathbf{u}_a) = 0, \end{cases} \tag{6}$$

where the spaces $\mathbf{X}_a$ and M_a were previously described and where $F_a^n(\cdot)$ is obtained from $F^n(\cdot)$ by replacing $\mathbf{u}^n$ by $\mathbf{u}_a^n$. Under some regularity assumptions on the data Z_B for the 2D-horizontal model and the following hypotheses:

(H1) the riverbanks are stiff, *i.e* $h = h(s,t),\ Z_B = Z_B(s)$,

(H2) the data $r,\ Z_B,\ L$ satisfy: $0 < L_0 \leq L \leq L_1,\ r \in W^{1,\infty}(I),\ Z_B \in W^{1,\infty}(I),\ L \in W^{2,\infty}(I)$,

for the 2D-vertical one, the choice of the previous spaces allows us to establish the well-posedness of both models.

Theorem 2. *For the two previous choices of* $\mathbf{X}_a^0 \times \mathrm{M}_a$*, problem (6) has at least one solution. The uniqueness of the solution holds under the same hypotheses as in the Theorem 1.*

The tools for the proof are the same as in the 3D case.

4 The Quasi-3D Model

Derivation of the Quasi-3D Model

We define the quasi-3D model by coupling the two previous 2D approaches. The idea is to search the approximation of the velocity and the pressure in the sum of both spaces. The model is written in curvilinear coordinates under the previous hypotheses *(H1)* and *(H2)*. The quasi-3D model is defined on the following 3D domain:

$$\Omega_F(t) = \{(s,l,z);\ s \in I,\ -L \leq l \leq L,\ Z_B \leq z \leq Z_B + h\}.$$

Then we look for the pressure and the velocity as follows:

$$p(s,l,z) = p_V(s,z) + (Z_B + h - z)\, P_H(s,l),$$

$$\mathbf{u} = \begin{pmatrix} (1 - lr)\,(u_{1V}(s,z) + u_{1H}(s,l)) \\ \frac{lL'}{L} u_{1V}(s,z) + u_{2H}(s,l) \\ u_{3V}(s,z) + u_{1H}(s,l)\partial_s Z_B + (z - Z_B)\, U_{3H}(s,l) \end{pmatrix},$$

with $u_{1V}(s, Z_B + h) = u_{1H}(s, 0)$ and $u_{3V}(s, Z_B + h) = hU_{3H}(s, 0)$. Thus the unknowns of this model are h, u_{1H}, u_{2H}, U_{3H}, P_H on $\Sigma_F(t)$, all functions of (s, l), and u_{1V}, u_{3V} and p_V on $\omega_F(t)$, all functions of (s, z). The subspaces $\mathbf{X}_{Q3D} \times \mathrm{M}_{Q3Dd}$ of the quasi-3D model, on which the 3D formulation (4) is now approximated, are built by taking the pressure and the velocity as above.

Finite Element Approximation of the Quasi-3D Model

We are now interested in the finite element approximation of the quasi-3D model. Let $\mathscr{T}_{Hd}$ a 2D mesh of the domain Σ_F, consisting of quadrangles K_H such that $\overline{\Sigma}_F = \bigcup_{K_H \in \mathscr{T}_{Hd}} \overline{K}_H$, and $\mathscr{T}_{Vd}$ a 2D mesh of the domain ω_F, consisting of quadrangles K_V such that $\overline{\omega}_F = \bigcup_{K_V \in \mathscr{T}_{Vd}} \overline{K}_V$. Next we can build a 3D mesh of the domain Ω_F from these two 2D meshes. We first write the free surface equation (2) by taking the 3D velocity in $\mathbf{X}_{Q3D}$. This leads to:

$$\frac{h - h^n}{\Delta t} + u_{1H}^n \partial_s h + u_{1V}^n \partial_s (Z_B + h) - hU_{3H}^n - u_{3V}^n = 0.$$

The space discretization is achieved by means of a vertex-centered finite volume method. The height of water h is approximated by piecewise linear elements on each $K \in \mathscr{T}_{Hd}$, continuous on Σ_F.

Next we have to choose compatible spaces $\mathbf{X}_d \times \mathrm{M}_d$ such that the discrete inf − sup condition is satisfied. We propose the following subspaces:

$$\begin{aligned}\mathbf{X}_d = \big\{ & \mathbf{u} \in \mathbf{X}_{Q3D};\ (u_{1V}, u_{3V})^t \in \mathbf{H}^1(\omega_F),\ (u_{1H}, u_{2H}, U_{3H})^t \in \mathbf{H}^1(\Sigma_F);\\ & (u_{1V})_{|K_V},\ (u_{3V})_{|K_V} \in Q_1(K_V),\ \forall K_V \in \mathscr{T}_{Vd},\\ & (u_{1H})_{|K_H},\ (u_{2H})_{|K_H},\ (U_{3H})_{K_H} \in Q_1(K_H),\ \forall K_H \in \mathscr{T}_{Hd} \big\},\end{aligned}$$

$$\begin{aligned}\mathrm{M}_d = \big\{ & p \in \mathrm{M}_{Q3D};\ p_V \in H^1(\omega_F),\ P_H \in H^1(\Sigma_F);\\ & (p_V)_{|K_V} \in Q_1(K_V),\ \forall K_V \in \mathscr{T}_{V_d},\ \text{and}\ (P_H)_{|K_H} \in Q_1(K_H),\\ & \forall K_H \in \mathscr{T}_{H_d} \big\}.\end{aligned}$$

The discrete inf − sup condition is satisfied with this choice of subspaces. We note that the couple $Q_1 - Q_1$ which is not suitable for the Stokes problem is sufficient for our problem. This comes from the form $B(\cdot, \cdot)$ of the quasi-3D problem which is more complex than the bilinear form of the Stokes problem.

5 Numerical Results

We present here the case of a channel with irregular width and bottom of 500 m length. We impose the velocity upstream and downstream and an initial water height of 4 m. We represent the streamlines on the 3D domain (Fig. 1) at a given time step

Fig. 1 Streamlines on the 3D domain at $t = 40s$

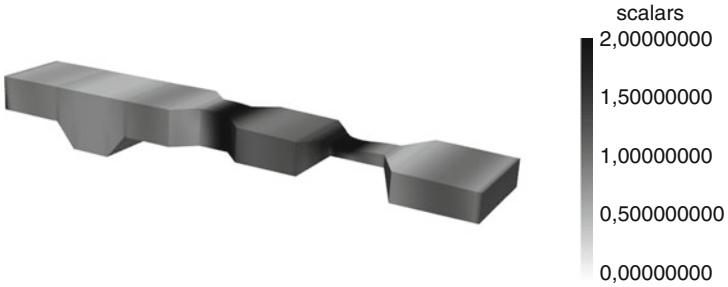


Fig. 2 Error estimators of the 2D-vertical model at $t = 40s$

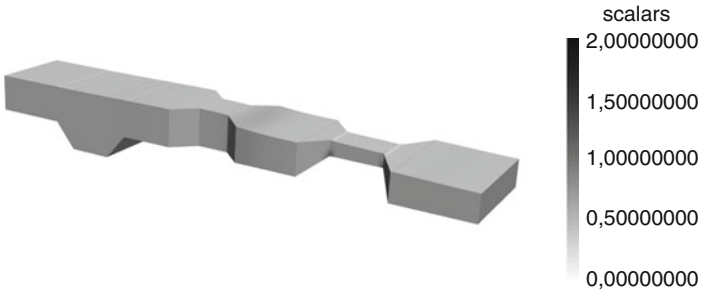


Fig. 3 Error estimators of the quasi-3D model at $t = 40s$

$t = 40s$. On all the following graphics, the scale of the height is dilated by 8 and the one of the width by 2.
The comparison of these results to those of both 2D models leads to the conclusions:

- If only the bottom is irregular, the 2D-vertical model is sufficiently accurate.
- If only the width is irregular, the 2D-horizontal model is sufficiently accurate.
- If both the bottom and the width are irregular, having a good accuracy requires the use of the quasi-3D model.

We confirm these results with error estimators defined between the 3D model and any of its lower-dimensional approximations [1]. These estimators measure the error between the 3D model and the lower-dimensional models. In example, we present the error estimators of the 2D-vertical model (Fig. 2) and those of the quasi-3D one (Fig. 3). The scale beside the graphics represents the values of the estimator.

References

1. Amara M., Capatina-Papaghiuc D., Trujillo D. (2004): Hydrodynamical modelling and multidimensional approximation of estuarian river flows, Comput. Vis. Sci. 6, 39–46
2. Amara M., Capatina-Papaghiuc D., Trujillo D. (2008): Variational approach for the multiscale modeling of an estuarian river. Part 1: Derivation and numerical approximation of a 2D horizontal model, Preprint LMA, University of Pau
3. Gerbeau J.-F., Perthame B. (2001): Derivation of Viscous Saint-Venant System for laminar shallow water. Numerical validation, discrete and continuous dynamical systems, Ser. B I(1), 89–102
4. Girault V., Raviart P.A. (1986): Finite element methods for Navier-Stokes equations. Theory and algorithms, Springer, Berlin

Convergence of a Mixed Discontinuous Galerkin and Finite Volume Scheme for the 3 Dimensional Vlasov–Poisson–Fokker–Planck System

Mohammad Asadzadeh and Piotr Kowalczyk

Abstract We construct a numerical scheme for the multi-dimensional Vlasov–Poisson–Fokker–Planck system based on a combined finite volume method for the Poisson equation in spatial domain and streamline-diffusion/ discontinuous Galerkin finite element methods in phase-space-time variables for the Vlasov–Fokker–Planck part. We derive error estimates with optimal convergence rates.

1 Introduction

In this note we study the approximate solution for the deterministic multi-dimensional Vlasov–Poisson–Fokker–Planck (VPFP) system described below: given the parameters $\beta \geq 0$, $\sigma \geq 0$ and the initial distribution of particle density $f_0(x, v)$, $(x, v) \in \Omega_x \times \mathbf{R}^d \subset \mathbf{R}^d \times \mathbf{R}^d$, $d = 1, 2, 3$; we seek the evolution of charged plasma particles (ions and electrons), at time t, with a phase space density $f(x, v, t)$ satisfying

$$\begin{cases} \partial_t f + v \cdot \nabla_x f - \nabla_x \varphi \cdot \nabla_v f - \operatorname{div}_v(\beta v f) - \sigma \Delta_v f = S, & \text{in } \Omega \times [0, T], \\ f(x, v, 0) = f_0(x, v), & \text{in } \Omega = \mathbf{R}^d \times \mathbf{R}^d, \\ -\Delta_x \varphi = \displaystyle\int_{\mathbf{R}^d} f(x, v, t)\, dv, & \text{in } \mathbf{R}^d \times [0, T], \end{cases} \tag{1}$$

M. Asadzadeh (✉)
Department of Mathematics, Chalmers University of Technology and the University of Gothenburg, SE–412 96, Göteborg, Sweden
e-mail: mohammad@chalmers.se

P. Kowalczyk
Department of Mathematics, Informatics and Mechanics, Warsaw University, Banacha 2, 02-097 Warszawa, Poland
e-mail: pkowal@mimuw.edu.pl

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_9,

where $\cdot$ denotes the scalar product and S is a source. To construct numerical methods we shall restrict both space and velocity variables x and v to be in bounded domains Ω_x and Ω_v; and provide the equation with a Dirichlet type inflow boundary condition. To solve problem (1) the idea is to split the equation system and separate Poisson equation from the Vlasov–Fokker–Planck equation. The two parts are coupled by the potential φ. Thus we reformulate the problem (1) as follows: Given the initial data $f_0(x,v)$, $(x,v) \in \Omega := \Omega_x \times \Omega_v \subset \mathbf{R}^d \times \mathbf{R}^d$, $d = 1,2,3$; find the density function $f(x,v,t)$ in the Dirichlet initial-boundary value problem for the Vlasov–Fokker–Planck equation

$$\text{(P1)} \quad \begin{cases} \partial_t f + v \cdot \nabla_x f - \nabla_x \varphi \cdot \nabla_v f - \operatorname{div}_v(\beta v f) - \sigma \Delta_v f = S, & \text{in } \Omega \times [0,T], \\ f(x,v,0) = f_0(x,v), & \text{in } \Omega_x \times \Omega_v, \\ f(x,v,t) = 0, & \text{on } \Gamma_G^- \times [0,T], \end{cases} \tag{2}$$

where $G := (v, -\nabla_x \varphi)$, $\Gamma_G^- := \{(x,v) \in \Gamma := \partial\Omega | G \cdot \mathbf{n} < 0\}$, is the inflow boundary and the potential φ satisfies the following problem for the Poisson equation:

$$\text{(P2)} \quad \begin{cases} -\Delta_x \varphi = \displaystyle\int_{\Omega_v} f(x,v,t)\, dv, & \text{in } \Omega_x \times [0,T], \\ |\nabla_x \varphi(x,t)| = 0, & \text{on } \partial\Omega_x \times [0,T]. \end{cases} \tag{3}$$

Now we can solve problem (P2) replacing f by a given function g. Then inserting the corresponding solution φ in (P1) we obtain an equation for f, viz (2). In this way we link the solution f to the given data function g as, say, $f = \Lambda[g]$. Now a solution f for the Vlasov–Poisson–Fokker–Planck system is a fixed point of the operator Λ, i.e., $f = \Lambda[f]$, which is obtained by a procedure using Schauder fixed point theorem. For the discrete version this step can, roughly speaking, be repeated using a Brouwer type fixed point argument, see, e.g., [1] and the references therein. Positivity, existence, uniqueness and regularity of the solution for the continuous problem are given in [5]. These results rely on the positivity and boundedness requirement for the second phase-space moment of the initial data: $f_0 \in L_\infty(\mathbf{R}^6) \geq 0$ and $\int_{\mathbf{R}^6}(1 + |x|^2 + |v|^2) f_0\, dx dv < \infty$. Further analytic approaches are given , e.g., by Horst in [11]. For the general mathematical study of equations of this type we refer to studies by Baouendi and Grisvard [4] and Lions [14].

Conventional numerical methods for the Vlasov–Poisson and related equations have been dominated by the particle-in-cell method studied, e.g., by Cottet and Raviart [7]; Ganguly, Lee, and Victory [9]; and Wollman, Ozizmir, and Narasimhan [16]. Filbet has studied a 1-dimensional finite volume scheme for the Vlasov–Poisson [8].

Our study of the VPFP system is, mainly, devoted (see also [1–3]) to the construction and analysis of finite element schemes. In this note, however, we study the Poisson part using a finite volume approach. To this end we consider the study of a three dimensional VPFP model ($\Omega_x \subset \mathbf{R}^3$, $\Omega_v \subset \mathbf{R}^3$). As for the discontinuous Galerkin approximation relevant in the VPFP estimates we also refer to the articles

by Brezzi, Manzini, Marini, and Russo for elliptic problem in [6], and Johnson and Saranen for the Euler and Navier–Stokes equations in [12].

In this note, we give only sketch of the proofs. They can be completed following the techniques in [15] for finite volume, and [1–3] and [12] in the finite element cases.

2 The Finite Volume Method for Poisson Equation in 3D

We consider the cell-center finite volume (FV) scheme for the problem (P2):

$$-\nabla_{\mathbf{x}}^2\varphi = \rho, \quad \text{in} \quad \Omega_{\mathbf{x}} = (0,1)\times(0,1)\times(0,1) \quad |\nabla_x\varphi| = 0, \quad \text{on} \quad \partial\Omega_{\mathbf{x}}, \quad (4)$$

where $\rho = \int_{\Omega_v} f(x,v,t)\,dv$. Existence, uniqueness and regularity studies for this problem are extensions of two-dimensional results in [10]: $\rho \in H^{-1}(\Omega_{\mathbf{x}})$ implies that $\exists!\ \varphi \in H_0^1(\Omega_{\mathbf{x}})$, and for $\rho \in H^s(\Omega_{\mathbf{x}})$, with $-1 \le s < r$, $r \ne \pm 1/2$, $\varphi \in H^{s+2}(\Omega_{\mathbf{x}})$. The finite volume scheme can be described as: exploiting divergence from the differential equation, integrating over disjoint "volumes" and using Gauss divergent theorem to convert volume-integrals to counter-integrals, and then discretizing to obtain the approximate solution φ_h with the mesh size h. Here, the finite volume method is defined on Cartesian product of non-uniform meshes as Petrov–Galerkin method using piecewise trilinear trial functions on *finite element* mesh and piecewise constant test functions on the dual box mesh. The main result of this section reads as follows:

Theorem 1. *For $1/2 < s \le 2$, the optimal finite volume error estimates for general non-uniform and quasi-uniform meshes are given by*

$$\|\varphi - \varphi_h\|_{1,h} \le C h^s |\varphi|_{H^{s+1}}, \quad \textit{and} \quad \|\varphi - \varphi_h\|_{\infty} \le C h^s |\log h||\varphi|_{H^{s+1}}. \quad (5)$$

The corresponding finite element estimates is given by the following result

Theorem 2. *a) For the finite element solution of the Poisson problem (4) with a quasiuniform triangulation we have the error estimate:*

$$\|\varphi - \varphi_h\|_{1,\infty} \le C h^r |\log h| \times \|\varphi\|_{r+1,\infty}, \qquad \textit{for} \quad r \le 2$$

b) $\forall \varepsilon \in (0,1)$ small, $\exists C_\varepsilon$ such that $\|\varphi - \varphi_h\|_{1,\infty} \ge C_\varepsilon h^{r-\varepsilon}|\log h|$, cf [13].

Note that, for the L_∞ estimate, $s = 2$ in Theorem 1 corresponds to $r = 1$ in Theorem 2. To derive the finite volume formula we consider the Cartesian mesh

$$\begin{aligned} I_x^h &:= \{x_i :\ i = 0,1,\ldots,I;\ \ x_0 = 0,\ x_i - x_{i-1} = h_i;\ \ x_I = 1\},\\ I_y^h &:= \{y_j :\ j = 0,1,\ldots,J;\ y_0 = 0,\ y_j - y_{j-1} = k_j;\ y_J = 1\},\\ I_z^\ell &:= \{z_n :\ n = 0,1,\ldots,N;\ z_0 = 0,\ z_n - z_{n-1} = \ell_n;\ \ z_N = 1\}. \end{aligned}$$

With each (x_i, y_j, z_n) we associate the finite volume box:

$$\omega_{ijn} = \Big(x_{i-1/2}, x_{i+1/2}\Big) \times \Big(y_{j-1/2}, y_{j+1/2}\Big) \times \Big(z_{n-1/2}, z_{n+1/2}\Big),$$

where we choose *central finite volume boxes* inside each 27-points stencil element:

$$\begin{cases} x_{i-1/2} = x_i - h_i/2, \;\; x_{i+1/2} = x_i + h_{(i+1)}/2, & \hbar_i = \frac{h_i + h_{i+1}}{2} \\ y_{j-1/2} = y_j - k_j/2, \; y_{j+1/2} = y_j + k_{(j+1)}/2, \text{ and let } & \bar{k}_j = \frac{k_j + k_{j+1}}{2}, \\ z_{n-1/2} = z_n - \ell_n/2, \;\; z_{n+1/2} = z_n + \ell_{(n+1)}/2, & \bar{\ell}_n = \frac{\ell_n + \ell_{n+1}}{2}. \end{cases}$$

Further, $\forall\, \tau < 1/2$, we define the characteristic function:

$$\chi_{ijn} = \text{Char}\Big[\Big(-\frac{h_{i+1}}{2}, \frac{h_i}{2}\Big) \times \Big(-\frac{k_{j+1}}{2}, \frac{k_j}{2}\Big)\Big] \times \Big(-\frac{\ell_{n+1}}{2}, \frac{\ell_n}{2}\Big)\Big] \in H^{\tau}(\mathbf{R}^3).$$

For finite volume approximation we let $\rho \in H^s(\Omega_{\mathbf{x}})$, $r > -1/2$ and extend ρ to $\mathbf{R^3}$ preserving its Sobolev class. Thus, we may define using three dimensional convolutions, $\chi_{ijn} * \rho$, which is continuous in $\mathbf{R^3}$, and Gauss divergence theorem that

$$\frac{1}{|\omega_{ijn}|} \int_{\partial\omega_{ijn}} \frac{\partial\varphi}{\partial \mathbf{n}}\, ds = \frac{1}{|\omega_{ijn}|} \Big(\chi_{ijn} * \rho\Big)(x_i, y_j, z_n). \tag{6}$$

Further, recalling that $\rho \in L^1_{loc}(\Omega_{\mathbf{x}})$ we may write

$$\frac{1}{|\omega_{ijn}|} \int_{\partial\omega_{ijn}} \frac{\partial\varphi}{\partial \mathbf{n}}\, ds = \frac{1}{\hbar_i \bar{k}_j \bar{\ell}_n} \int_{x_{i-1/2}}^{x_i+1/2} \int_{y_{j-1/2}}^{y_j+1/2} \int_{z_{n-1/2}}^{z_n+1/2} \rho(x, y, z)\, dx\, dy\, dz. \tag{7}$$

Now we let $\mathscr{V}_h$ be the set of piecewise trilinear functions defined on the box $\Omega_{\mathbf{x}}$ induced by $\bar{\Omega}^h_{\mathbf{x}}$, i.e., $\mathscr{V}_h^{\circ} = \{F \in \mathscr{V}_h \,\big|\, F = 0 \quad \text{on} \quad \partial\Omega_{\mathbf{x}}\}$.

Definition 1. The finite volume approximation of the solution φ for the Poisson equation (4); $\varphi_h \in \mathscr{V}_h^{\circ}$ is defined (implicitly) through the following algorithm:

$$-\frac{1}{\hbar_i \bar{k}_j \bar{\ell}_n} \int_{\partial\omega_{ijn}} \frac{\partial\varphi_h}{\partial \mathbf{n}}\, ds = \frac{1}{\hbar_i \bar{k}_j \bar{\ell}_n} \Big(\chi_{ijn} * \rho\Big)(x_i, y_j, z_n), \qquad (x_i, y_j, z_n) \in \Omega^h_{\mathbf{x}}.$$

Stability and convergence of this method are generalization of Süli's [15] results in two dimensions for the Dirichlet problem. $|\nabla_x \varphi| = 0$ on $\partial\Omega_x$ with extended $\varphi(\infty) = 0$ yields $\varphi = 0$ on $\partial\Omega_x$. The first assertion in Theorem 1, may be proved repeating the arguments in [15] (we skip) for the 3d case in discrete $H^1(\Omega^h_{\mathbf{x}})$ and $L_2(\Omega^h_{\mathbf{x}})$ norms:

$$\|\psi\|_{1,h} = \Big(\|\psi\|^2 + |\psi|^2_{1,h}\Big)^{1/2}, \quad \text{and} \quad \|\psi\| = (\psi, \psi)^{1/2}, \qquad \text{where}$$

$$(\phi,\psi)=\sum_{i=1}^{I-1}\sum_{j=1}^{J-1}\sum_{n=1}^{N-1}\hbar_i\bar{k}_j\bar{\ell}_n\phi_{ijn}\psi_{ijn},\qquad\text{and}$$

$$|\psi|_{1,h}=\left(||\Delta_x^-\psi||_x^2+||\Delta_y^-\psi||_y^2+||\Delta_z^-\psi||_z^2\right)^{1/2},\qquad\text{with}$$

divided differences $\Delta_x^-\psi_{ijn}=(\psi_{ijn}-\psi_{i-1,j,n})/\hbar_i$, $\Delta_y^-\psi_{ijn}=(\psi_{ijn}-\psi_{i,j-1,n})/\bar{k}_j$ and $\Delta_z^-\psi_{ijn}=(\psi_{ijn}-\psi_{i,j,n-1})/\bar{\ell}_n$, and the, *one-sided discrete L_2-norms*

$$||\Delta_x^-\psi||_x^2=(\psi,\psi]_x,\qquad(\phi,\psi]_x=\sum_{i=1}^{I}\sum_{j=1}^{J-1}\sum_{n=1}^{N-1}\hbar_i\bar{k}_j\bar{\ell}_n\phi_{ijn}\psi_{ijn},$$

with the similar notations corresponding to the y and z directions.

3 Streamline Diffusion and Discontinuous Galerkin Approaches

For a finite element scheme over the phase–space–time domain $\Omega_T:=[0,T]\times\Omega$ we start with a phase–space subdivision of Ω, into the product of triangular elements τ_x and τ_v as $\mathscr{T}_h:=\{\tau=\tau_x\times\tau_v\}$ combined with a partition of the time interval $(0,T)$: $0=t_0<t_1<\ldots<t_M=T$, and let $I_m:=(t_m,t_{m+1})$; $m=0,1,\ldots,M-1$. Then the corresponding partition of Ω_T is given by the *prism-type triangulation*

$$\mathscr{C}_h:=\{K|K:=\tau\times I_m,\ \tau\in\mathscr{T}_h\}.$$

We seek piecewise polynomial approximations for the solution of problem (1) in a finite dimensional space

$$V_h:=\{f\in\mathscr{H}:f|_K\in\mathscr{P}_k(\tau)\times\mathscr{P}_k(I_m);\ \ \forall K=\tau\times I_m\in\mathscr{C}_h\},$$

with V_h being continuous in x and v, possibly discontinuous in t across time levels t_m and $\mathscr{H}:=\prod_{m=0}^{M-1}H^1(\Omega_m)$; $\Omega_m=\Omega\times I_m$. We shall also use the standard notation

$$(f,g)_m=(f,g)_{\Omega_m}=\int_{\Omega_m}fg\,dx\,dv\,dt,\qquad\|g\|_m=(g,g)_m^{1/2},$$

$$\langle f,g\rangle_m=\int_{\Omega}f(\cdot,\cdot,t_m)g(\cdot,\cdot,t_m)\,dx\,dv,\qquad|g|_m=<g,g>_m^{1/2},$$

$$<f^{\mp},g^{\mp}>_{\Gamma^{\pm}}=\int_{\Gamma^{\pm}}f^{\mp}g^{\mp}|G^h\cdot n|dv,\qquad\text{and the jumps}\quad[g]=g^+-g^-$$

$$<f^{\mp},g^{\mp}>_{\lambda\pm}=\int_{I_m}<f^{\mp},g^{\mp}>_{\Gamma^{\pm}}\,dt,\qquad g^{\pm}=\lim_{s\to0\pm}g(x,v,t+s).$$

Using notation $\nabla f := (\nabla_x f, \nabla_v f) = (\partial f/\partial x_1, \ldots, \partial f/\partial x_d, \partial f/\partial v_1, \ldots, \partial f/\partial v_d)$ and $G := \Big(v_1, \ldots, v_d, -\partial\phi/\partial x_1, \ldots, -\partial\phi/\partial x_d\Big)$ we get $\mathrm{div} G(f) = 0$. For our finite element procedure (both in the streamline diffusion and the discontinuous Galerkin case) we let $\mathscr{F}$ to be a certain (linear) function space, $\tilde{f} \in \mathscr{F}$ an approximation of f and $\Pi f \in \mathscr{F}$ a projection of f into $\mathscr{F}$, then to estimate the approximation error

$$f - \tilde{f} = (f - \Pi f) + (\Pi f - \tilde{f}) \equiv \eta + \xi; \quad \xi \in \mathscr{F},$$

(i) we use interpolation theory to give sharp error bounds for a certain $|||\eta|||$-norm
(ii) we establish $|||\xi||| \le C|||\eta|||$, $(||| \cdot ||| := \|\cdot\|_{\Xi}$, Ξ=SD or Ξ=DG, below).

Now we consider the streamline diffusion (SD) method for (P1) with test functions of the form $u + \delta\Big(u_t + G(\tilde{f})\cdot\nabla u\Big)$ with $\delta \sim h$, the mesh size. For convenience we use the notation $\mathscr{D}w := w_t + G(f_h)\cdot\nabla w$ and formulate the SD method for problem (I) as follows: given $f_h^-(\cdot,\cdot,t_m)$ find $f_h \in V_h$ such that for $m = 0, \ldots, M-1$,

$$(P_m^h) \qquad B_m^\delta(G(f_h); f_h, u) - J_m^\delta(f_h, u) = L_m^\delta(u), \qquad \forall u \in V_h. \tag{8}$$

$$B_m^\delta := (\mathscr{D} f_h, u + \delta\mathscr{D}u)_m + \sigma(\nabla_v f_h, \nabla_v u)_m + \langle [f_h], u\rangle_m - \delta\sigma(\Delta_v f_h, \mathscr{D}u)_m, \tag{9}$$

$$J_m^\delta := (\nabla_v\cdot(\beta v f_h), u + \delta\mathscr{D}u)_m, \tag{10}$$

and

$$L_m^\delta := (S, u + \delta\mathscr{D}u)_m + \langle f^+, u^+\rangle_{\lambda_m^-} + \langle f^-, u^-\rangle_{\lambda_m^+}. \tag{11}$$

Problem P_m^h is a linear system of equations leading to an implicit scheme. Therefore to solve P_1 by the SD method is equivalent to find $f_h \in V_h$ such that

$$B^\delta(G(f_h); f_h, u) - J^\delta(f_h, u) = L^\delta(u), \qquad \forall u \in V_h, \tag{12}$$

$$B^\delta := \sum_{m=0}^{M-1} B_m^\delta, \quad J^\delta := \sum_{m=0}^{M-1} J_m^\delta, \quad L^\delta := \sum_{m=0}^{M-1} L_m^\delta. \tag{13}$$

3.1 Stability and Error Estimates

Lemma 1. *For the SD method we have the coercivity and stability estimates*

$$\forall g \in \mathscr{H}, \quad B^\delta\Big(G(f^h); g, g\Big) \ge \frac{1}{2}||g||_{SD}^2, \qquad \textit{with}$$

$$||g||_{SD}^2 = \frac{1}{2}\Big[2\sigma\|\nabla_v g\|_{\Omega_T}^2 + |g|_M^2 + |g|_0^2 + \sum_{m=1}^{M-1}|[g]|_m^2 + 2\|\mathscr{D}g\|_{\Omega_T}^2 + \int_{\Gamma\times I} g^2|G^h\cdot n|\Big],$$

$$||g||^2_{L_2(\Omega_T,SD)} \le \Big[\frac{1}{C_1}||\mathscr{D}g||^2 + \sum_{m=1}^{M-1}|[g]|^2_m + \int_{\partial\Omega\times I} g^2|G^h\cdot n|\Big]\delta e^{C_1\delta}, \quad \forall C_1 \ge 0.$$

Remark 1. In the discontinuous Galerkin case $||g||_{DG}$ and $||g||_{L_2(\Omega_T,DG)}$ are defined by replacing the $\int$-term, in the SD case, by $\sum\int_{\partial K_-(G'')}[g]^2|G^h\cdot n|\,ds$ where

$$\partial K_-(G'') = \{(x,v,t)\in\partial K_-(G') : n_t(x,v,t) = 0\}.$$

Theorem 3. *Assume that there is a constant C such that*

$$||\nabla f||_\infty + ||G(f)||_\infty + ||\nabla\eta||_\infty \le C. \tag{14}$$

Then we have the following error estimate for the SD method for (P1):

$$\|f - f_{SD}\|_{SD} \le C h^{k+1/2}\|f\|_{H^{k+1}(\Omega_T)},$$

where $f_{SD}\in V_h$ is the SD-approximation for f, and we have assumed $f\in H^{k+1}(\Omega_T)$.

Proof. (sketch of the main ideas) Let $\tilde{f}^h$ be an interpolant of f, and split the error as

$$e = f - f_{SD} = f - \tilde{f}^h + \tilde{f}^h - f_{SD} := \eta - \xi.$$

Then, by the above coercivity estimate and Galerkin orthogonality, we may write

$$\frac{1}{2}||\xi||^2_{SD} \le B(G(f^h);\xi,\xi) = B(G(f);f,\xi) - B(G(f^h);\tilde{f}^h,\xi) + J(f^h,\xi) - J(f,\xi)$$
$$:= \Delta B + \Delta J \le \frac{1}{8}||\xi||^2_{SD} + C_B||\eta||^2_{SD} + \frac{1}{8}||\xi||^2_{SD} + C_J||\eta||^2_{SD},$$

where to estimate J-term, we have used the inverse estimate. Further interpolation estimates give $||\eta||^2_{SD} \le C_i h^{k+1/2}\|f\|_{H^{k+1}(\Omega_T)}$, which yields the desired result.

In the discontinuous Galerkin (DG) case we assume also discontinuities in x and v over the interelement boundaries. Here, we shall use the discrete function spaces

$$W_h = \Big\{g\in L_2(Q_T) : g|_K \in P_k(K) \quad \forall K\in\mathscr{C}_h\Big\}, \quad \text{and}$$
$$W_h^d = \Big\{w\in [L_2(Q_T)]^d : w|_K \in [P_k(K)]^d \quad \forall K\in\mathscr{C}_h\Big\}.$$

Then, the corresponding final error estimate for the DG case reads as follows:

Theorem 4. *Under the assumptions (14) of Theorem 3 and regularity assumption for the exact solution as $f\in H^{k+1}(\Omega_T)\cap W^{k+1,\infty}(\Omega_T)$; we have that the discontinuous Galerkin approximation $f_{DG}\in W_h^d$ for f in (P1) satisfies the error estimate*

$$\|f - f_{DG}\|_{DG} \leq Ch^{k+1/2}\Big(\|f\|_{H^{k+1}(\Omega_T)} + \|f\|_{W^{k+1,\infty}(\Omega_T)}\Big).$$

Proof. (Sketchy) Here we demonstrate only the terms that are involved in estimations of the interelement jump terms, which are additional to those in the SD-case. To this end, let E_v be the set of all interior edges of the triangulation T_h^v and χ^{ext} the value of χ in the element τ_v^{ext} which has $e \in E_v$ as a common edge with τ_v. Define $(\chi)^0 := (\chi + \chi^{ext})/2$ and $[[\chi]] := \chi - \chi^{ext}$. Now we define $R: W_h \to W^d$, cf [6], by

$$R(g)w = -\sum_{\tau_x \times I_m} \int_{\tau_x \times I_m} \sum_{e \in E_v} \int_e [[g]] n_v \cdot (w)^0 \, dv, \qquad \forall\, w \in W_h^d, \tag{15}$$

and let r_e be the restriction of R to the elements sharing the edge $e \in E_v$, then

$$r_e(g)w = -\sum_{\tau_x \times I_m} \int_{\tau_x \times I_m} \int_e [[g]] n_v \cdot (w)^0 \, dv, \qquad \forall\, w \in W_h^d. \tag{16}$$

Hence, we may easily verify that

$$\sum_{e \subset \partial\tau_v \cap E_v} r_e = R \quad \text{on } \tau_v \Longrightarrow \|R(g)\|_K^2 \leq \gamma \sum_{e \subset \partial\tau_v \cap E_v} \|r_e(g)\|_K^2, \tag{17}$$

where τ_v corresponds to the element K and $\gamma = \gamma(d) > 0$ is a constant. Furthermore, since the support of each r_e is the union of elements sharing the edge e, we get

$$\sum_{e \in Ev} \|r_e(g)\|^2 = \sum_{K \in C} \sum_{e \subset \partial\tau_v \cap Ev} \|r_e(g)\|_K^2. \tag{18}$$

The corresponding discontinuous Galerkin method reads as: find $f_h \in W_h$ such that

$$B_{DG}(G(f_h); f_h, g) - \Big(\nabla_v(\beta v f), g + h\mathscr{D}g\Big) = L(g), \qquad \forall g \in W_h, \tag{19}$$

Proving a coercivity which, compared to B_{SD}, contains also interelement jumps;

$$(B_{DG}G(f^h); g, g) \geq \alpha |||g|||^2, \qquad \forall g \in W_h,$$

and following the same procedure as in the SD case yields the DG error estimate.

Acknowledgements The first author was supported by the Swedish Foundation of Strategic Redearch (SSF) in Gothenburg Mathematical Modeling Center (GMMC).

References

1. M. Asadzadeh, Streamline Diffusion Methods for the Vlasov–Poisson equation, Math. Model. Numer. Anal., **24** (1990), no. 2, 177–196
2. M. Asadzadeh and P. Kowalczyk, Convergence of Streamline Diffusion Methods for the Vlasov–Poisson–Fokker–Planck System Numer Methods Partial Differential Eqs., **21** (2005), 472–495
3. M. Asadzadeh and A. Sopasakis, Convergence of a hp Streamline Diffusion Scheme for Vlasov–Fokker–Planck system Math. Mod. Meth. Appl. Sci., **17** (2007), 1159–1182
4. M. S. Baouendi and P. Grisvard, Sur une équation d' évolution changeant de type, J. Funct. Anal., (1968), 352–367
5. F. Bouchut, Smoothing Effect for the Non-linear Vlasov–Poisson–Fokker–Planck System, J. Part. Diff. Equ., **122** (1995), 225–238
6. F. Brezzi, G. Manzini, D. Marini, P. Pietra and A. Russo, Discontinuous Galerkin Approximations for Elliptic Problems, Numer. Meth. Part. Diff. Equ., **16** (2000), no. 4, 365–378
7. G. H. Cottet and P. A. Raviart, On Particle-in-Cell Methods for the Vlasov–Poisson equations, Trans. Theory Stat. Phys., **15** (1986), 1–31
8. F. Filbet, Convergence of a Finite Volume Scheme for the Vlasov–Poisson System, SIAM J. Numer. Anal., **39** (2001), 1146–1169
9. K. Ganguly, J. Todd Lee, and H. D. Victory, Jr., On Simulation Methods for Vlasov-Poisson Systems with Particles Initially Asymptotically Distributed, SIAM J. Numer. Anal., **28** (1991), no. 6, 1547–1609
10. P. Grisvard, Elliptic Problems in Non-Smooth Domains, Pitman, London (1965)
11. E. Horst, On the Asymptotic Growth of the Solutions of the Vlasov–Poisson System, Math. Meth. Appl. Sci., **16** (1993), no. 2, 75–78
12. C. Johnson and J. Saranen, Streamline Diffusion Methods for the Incompressible Euler and Navier–Stokes Equations, Math. Comp., **47** (1986), 1–18
13. Y. Lin, V. Thomee and L. Wahlbin, Ritz-Volterra Projections to Finite-Element Spaces and Applications to Integrodifferential and Related Equations. SIAM J. Numer. Anal., **28** (1991), 1047–1070
14. J. L. Lions, Equations différentielles opérationnelle et problm̀es aux limites, Springer, Berlin (1961)
15. E. Süli, Convergence of Finite Volume Schemes for Poisson's Equation on Nonuniform Meshes, SIAM, J. Numer. Anal., **26** (1991), no. 5, 14191–1430
16. S. Wollman, E. Ozizmir and R. Narasimhan, The Convergence of the Particle Method for the Vlasov–Poisson System with Equally Spaced Initial Data Points, Transport Theory Stat. Phys., **30** (2001), no. 1, 1–62

Infrastructure for the Coupling of Dune Grids

Peter Bastian, Gerrit Buse, and Oliver Sander

Abstract We describe an abstract interface for the geometric coupling of finite element grids. The scope of the interface encompasses a wide range of domain decomposition techniques in use today, including nonconforming grids and grids of different dimensions. The couplings are described as sets of remote intersections, which encapsulate the relationships between pairs of elements on the coupling interface.

The abstract interface is realized in a module `dune-grid-glue` for the software framework DUNE. Several implementations of this interface exist, including one for general nonconforming couplings and a special efficient implementation for conforming interfaces. We present two numerical examples to show the flexibility of the approach.

1 Introduction

Domain decomposition methods are a standard tool for a wide range of multiphysics problems. Whenever the application involves subdomains with different equations, discretizations, or grid types, coupling conditions and domain decomposition algorithms need to be employed. We refer to [7] for a general introduction.

Even though domain decomposition methods have found widespread use, the software support available is generally not satisfactory. Implementing domain decomposition methods can be tedious and error prone, especially when nonmatch-

P. Bastian
Universität Heidelberg, Germany
e-mail: peter.bastian@iwr.uni-heidelberg.de

G. Buse
Technische Universität München, Germany
e-mail: buse@in.tum.de

O. Sander (✉)
Freie Universität Berlin, Germany
e-mail: sander@mi.fu-berlin.de

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_10,

ing grids are involved. A central problem is finding the geometric correspondences between the grids. Today, there still exist mainly ad hoc solutions geared towards specific purposes, with little chance of code reuse.

In this article, we propose a general implementation as part of the DUNE framework [1]. DUNE is a set of C++ libraries providing support for various aspects of grid-based PDE solution methods such as grids, linear algebra, or shape functions. DUNE's main goal is flexibility, achieved by defining abstract interfaces to such things as grids and shape functions, and allowing the user to select the appropriate implementation according to his or her needs. DUNE also promotes code reuse by a modular architecture and by allowing legacy implementations to be used with the interface.

For our domain decomposition infrastructure we have tried to follow the same philosophy:

- We propose abstract interfaces to general grid coupling mechanisms, allowing to implement most existing domain decomposition algorithms.
- We allow and encourage the use of existing coupling implementations as legacy backends.
- We strive to make the code efficient, using generic programming where appropriate.

Adhering to the modular structure of DUNE, our code is available as a DUNE module, termed `dune-grid-glue`.

2 General Grid Coupling

We begin by describing the concept of the abstract grid coupling interface. For simplicity we focus on the case of nonoverlapping coupling. Consider two domains Ω_1, Ω_2 that meet at a common interface Γ (Fig. 1). Both domains are assumed to be discretized by grids, not necessarily simplicial. The restrictions of the grids to the coupling boundary, denoted by $\mathcal{G}_{\Gamma_1}$ and $\mathcal{G}_{\Gamma_2}$, are not related to each other in any way.

Overlaying these two boundary grids results in a set of intersections of the elements of $\mathcal{G}_{\Gamma_1}$ and $\mathcal{G}_{\Gamma_2}$, which we call $\mathcal{G}_M$. Together with the embeddings into $\mathcal{G}_{\Gamma_1}$

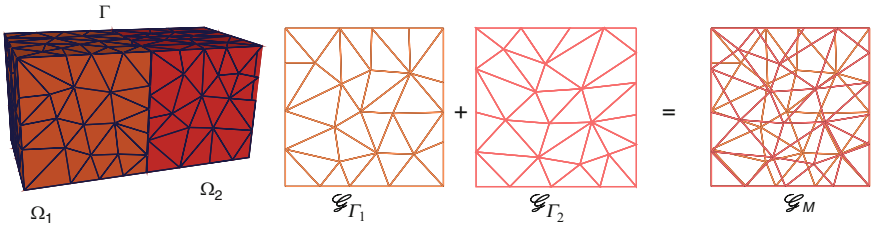


Fig. 1 *Left*: two domains Ω_1 and Ω_2 that meet at a common interface Γ. *Center*: the restrictions of the two grids on Γ. *Right*: together they form the set of remote intersections $\mathcal{G}_M$

and $\mathcal{G}_{\Gamma_2}$, the intersections constitute the information necessary to implement most nonoverlapping domain decomposition algorithms.

As an example, consider the mortar method. There, the coupling is effected through a mass matrix

$$M \in \mathbb{R}^{n \times m}, \qquad M_{ij} = \int_{\Gamma} \phi_i \psi_j \, ds, \tag{1}$$

where the ϕ_i, ψ_j, $0 \leq i < n$, $0 \leq j < m$, are finite element basis functions on $\mathcal{G}_{\Gamma_1}$ and $\mathcal{G}_{\Gamma_2}$, respectively. The matrix can be computed by splitting the integral in (1) into a sum of integrals over individual elements of $\mathcal{G}_M$. By construction, to each element $e \in \mathcal{G}_M$ correspond unique elements of $\mathcal{G}_{\Gamma_1}$ and $\mathcal{G}_{\Gamma_2}$, and associated shape functions there. If a quadrature rule is available for e, then $\int_e \phi_i \psi_j \, ds$ can be computed directly. Otherwise, e needs to be triangulated and $\int \phi_i \psi_j \, ds$ computed for each triangle.

The approach covers more than just mortar methods. If the two grids on Ω_1 and Ω_2 match, the set $\mathcal{G}_M$ degenerates and we have $\mathcal{G}_M = \mathcal{G}_{\Gamma_1} = \mathcal{G}_{\Gamma_2}$. In this case, the set of intersections e together with their embeddings into the elements of $\mathcal{G}_1$ and $\mathcal{G}_2$ allows to identify the grid vertices, or, more generally, edge and face degrees of freedom. Overlapping couplings can be handled by letting $\mathcal{G}_M$ have the same dimension as the computational grids $\mathcal{G}_1$ and $\mathcal{G}_2$. Finally, consider a d-dimensional grid attached in parallel to the boundary of a $d+1$-dimensional one (cf. Sect. 5.2). The grids may or may not be conforming on Γ. This time coupling is between the surface grid $\mathcal{G}_{\Gamma_2}$ and the grid $\mathcal{G}_1$ itself. As the dimensions are the same, a set of intersections just as in Fig. 1 is obtained.

3 Implementation: Remote Intersections

The intersections described in the previous section bear close resemblance to the intersections that are part of the DUNE grid interface [2, Sect. 4]. Within a single grid, DUNE intersections describe the coupling between neighboring elements. An intersection between two elements e_1 and e_2 is the (set-theoretic) intersection between θ_{e_1} and θ_{e_2}, where θ_{e_1} and θ_{e_2} are the subsets of the world space occupied by e_1 and e_2, respectively. The `Intersection` class of the DUNE grid interface provides information about these set intersections, e.g., their geometry in the world space, the geometry in coordinates of e_1 and e_2, normal vectors, and whether an intersection is conforming.

In the case of domain decomposition methods, the elements e_1 and e_2 are elements of different grids $\mathcal{G}_1$ and $\mathcal{G}_2$. However, the relevant information remains largely the same. We will call such intersections *remote intersections*, to distinguish them from the intersections of the DUNE grid interface. Remote intersections may be set-theoretic intersections if $\mathcal{G}_1$ and $\mathcal{G}_2$ meet at a common interface Γ. In case of

contact problems, where there may be a positive distance between $\mathcal{G}_{\Gamma_1}$ and $\mathcal{G}_{\Gamma_2}$, the remote intersections can be defined via a contact mapping $\Phi : \mathcal{G}_{\Gamma_1} \to \mathcal{G}_{\Gamma_2}$ (cf. [9]).

Due to the conceptual similarity between remote intersections and grid intersections it is natural to make the implementation of remote intersections resemble DUNE intersections as well.

The `dune-grid-glue` module provides the class `RemoteIntersection`, which again has methods for the geometry of the intersection in world space, geometries in local coordinates of e_1 and e_2, normal vectors, etc. The main differences concern methods that deal with global coordinates. Since θ_{e_1} and θ_{e_2} may actually be disjoint (e.g., in a contact problem), there are two embeddings of the remote intersection in the world space. For the same reason, there are two methods for the normal vectors. Please see the class documentation provided with the module for details.

Access to the remote intersections is provided via three types of DUNE-style iterators.

The `RemoteIntersectionIterator` iterates over the entire set of remote intersections and can be used to, e.g., assemble mortar mass matrices.

The `DomainIntersectionIterators` and `TargetIntersectionIterators` iterate over all remote intersections of a given element of $\mathcal{G}_1$ or $\mathcal{G}_2$, respectively. This can be useful to assemble element-wise contributions in DG methods.

4 Constructing Couplings

The construction of sets of remote intersections proceeds in two steps. First, the grid interface boundaries or coupling parts are extracted and transformed to an intermediate representation. Then, two such extracted grids are combined to yield the set of remote intersections.

4.1 Extractors

`Extractor` classes select the subsets of grid entities that are involved in the coupling. They are classified according to the codimension (with respect to the grids) of the objects they extract. The most common one, `Codim1Extractor`, extracts boundary faces, and will be used for nonoverlapping couplings. The faces are marked using predicate classes provided by the user. The `Codim0Extractor` extracts actual elements. Such extractors will be needed for an overlapping coupling. A `Codim2Extractor` has not been implemented yet, but may be useful to couple, e.g., 1d partial differential equations to sequences of edges in a 3d mesh.

The extracted grid entities can be manipulated with a geometric transformation $\mu : \mathbb{R}^{n_1} \to \mathbb{R}^{n_2}$, $n_1 \leq n_2$. This may be a deformation or an embedding into

a higher-dimensional space. There are various uses for such a feature. For example, you may want to consider coupled problems on deformed meshes, such as the finite-strain contact problem described in [8]. Also, when coupling a 1d grid to the boundary of a 2d grid, then most likely the 1d grid implementation will live in a 1d world. A transformation can then be used to place the 1d grid in the 2d world and deform it, if necessary (see Sect. 5.2 for an example).

4.2 Computing Remote Intersections

With the two interacting grid parts extracted, they can be combined to obtain the set of remote intersections. How this should be implemented differs considerably depending on the actual scenario. A general implementation computing remote intersections would have to handle nonmatching grids and geometries, grids of arbitrary dimensions and element types. Besides being very difficult to write and debug, such a program would be inefficient in more regular situations such as when the grids match.

To resolve this dilemma we follow the DUNE philosophy. We prescribe an abstract interface that algorithms computing remote intersections should conform to. We then provide different implementations of the interface for different cases such as contact problems, conforming meshes, or overlapping grids. Also in accordance with the DUNE philosophy, legacy implementations can be used through the interface.

The current default implementation uses the PSURFACE library. This library was originally written to manage boundary parametrizations [6], and extended to also handle mappings for contact problems [9]. It manages piecewise affine mappings between simplicial hypersurfaces in 2d and 3d. The surfaces are identified by a normal projection $\Phi : \Gamma_1 \to \Gamma_2$. PSURFACE is free software and can be downloaded from http://numerik.mi.fu-berlin.de/dune/psurface.

Also, a special efficient implementation `ConformingMerge` for conforming couplings is available.

5 Numerical Examples

In this last chapter we demonstrate some of the possibilities of `dune-grid-glue` with two example applications. The first one, a two-body contact problem, has already appeared in [1], where the coupling was implemented using PSURFACE directly.

5.1 Contact Between a Structured and an Unstructured Grid

In this first example we compute mechanical contact between a human femur bone and an elastic foundation. Consider two disjoint domains Ω_1, Ω_2 in $\mathbb{R}^3$. The boundary $\Gamma_i = \partial\Omega_i$, $i = 1, 2$, of each domain is decomposed in three disjoint parts $\Gamma_i = \Gamma_{i,D} \cup \Gamma_{i,N} \cup \Gamma_{i,C}$. With $\mathbf{f}_i \in (L_2(\Omega_i))^3$ two body force density fields we look for functions $\mathbf{u}_i \in (H^1(\Omega_i))^3$ which fulfill

$$-\operatorname{div}\sigma(\mathbf{u}_i) = \mathbf{f}_i,$$

and suitable boundary conditions. The stress tensor σ is defined as $\sigma = \frac{E}{1+\nu}(\epsilon + \frac{\nu}{1-2\nu}\operatorname{tr}\epsilon I)$, and $\epsilon(\mathbf{u}) = \frac{1}{2}(\nabla\mathbf{u} + \nabla\mathbf{u}^T)$ is the linear strain tensor. For the contact condition, assume that the areas where contact occurs will be subsets of $\Gamma_{1,C}$ and $\Gamma_{2,C}$. These two contact boundaries are identified using a homeomorphism $\Phi : \Gamma_{1,C} \to \Gamma_{2,C}$, and this identification is used to define an initial distance function $g : \Gamma_{1,C} \to \mathbb{R}$, $g(x) = \|\Phi(x) - x\|$. The contact condition then states that the relative normal displacement of any two points x, $\Phi(x)$, $x \in \Gamma_{1,C}$, should not exceed this normal distance, in formulas

$$\mathbf{u}_1|_{\Gamma_{1,C}} \cdot \mathbf{n}_1 + (\mathbf{u}_2 \circ \Phi)|_{\Gamma_{2,C}} \cdot \mathbf{n}_2 \leq g, \tag{2}$$

where $\mathbf{n}_i$, $i = 1, 2$, is the unit outward normal of $\Gamma_{i,C}$. Condition (2) can be derived as a linearization of the actual nonpenetration condition and is reasonable to use in the context of linear elasticity [4].

For the discretization of the problem we use first-order Lagrangian elements for the interior and dual mortar elements for the contact condition. That is, (2) is discretized in a weak form requiring

$$\int_{\Gamma_{1,C}} \left[\mathbf{u}_1|_{\Gamma_{1,C}} \cdot \mathbf{n}_1 + (\mathbf{u}_2 \circ \Phi)|_{\Gamma_{2,C}} \cdot \mathbf{n}_2\right]\theta \, ds \leq \int_{\Gamma_{1,C}} g\theta \, ds \tag{3}$$

for all θ from a cone of dual mortar test functions defined on $\Gamma_{1,C}$ [10]. The resulting discrete obstacle problem is solved with a truncated nonsmooth Newton multigrid method as described by [5].

As the femur geometry we choose the distal part of the Visible Human femur data set. As grid implementations we use `UGGrid` for the femur and the structured hexahedral `SGrid` for the foundation. Material parameters are $E = 17\,\text{GPa}$, $\nu = 0.3$ for the bone and softer $E = 250\,\text{MPa}$, $\nu = 0.3$ for the obstacle. The latter is clamped at its base, whereas a uniform displacement of 3 mm downward is prescribed on the top section of the bone (see Fig. 2). The bone serves as the nonmortar domain. The computation of (3) involves a mortar mass matrix similar to (1). Two `Codim1Extractors` are used to mark the contact boundaries and the remote intersections are computed using the PSURFACE backend. The result can be seen in Fig. 2, right.

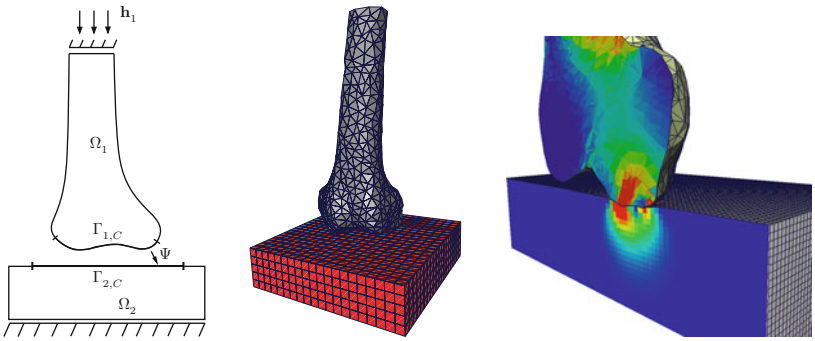


Fig. 2 Two-body contact problem. *Left*: schematic view. *Center*: coarse grids. *Right*: close-up view of the deformed solution

5.2 Coupling a 2d Richards Equation and a 1d Shallow-Water Equation

In the second example we show how `dune-grid-glue` can be used to couple two domains of differing dimensions.[1] Consider a domain Ω as in Fig. 3. It is supposed to represent a vertical section of ground. We assume unsaturated subsurface flow modeled by the Richards equation

$$\theta(p)_t + \operatorname{div} \mathbf{v}(p) = 0, \qquad \mathbf{v}(p) = -K \operatorname{kr}(\theta(p)) \nabla(p - \rho g z),$$

for the water pressure p in Ω. We denote the upper horizontal boundary of Ω by Γ and assume surface water there modeled by the shallow water equations

$$\begin{aligned} h_t + \operatorname{div} \mathbf{q} &= F \\ \mathbf{q}_t + \operatorname{div}(\mathbf{q}^2/h + 0.5 g h^2) &= -g h \nabla f, \end{aligned} \tag{4}$$

for the surface water height h and the horizontal water flux $\mathbf{q}$.

The two equations are coupled by assuming that the pressure p of the ground water on Γ equals the hydrostatic pressure induced by the surface water

$$p = \rho g h,$$

and that the flow $\mathbf{v} \cdot \mathbf{n}$ across Γ enters the surface water balance as an additive term in (4).

The coupled problem is solved with a Dirichlet–Neumann-type solver. At each iteration i, a Richards problem is solved on Ω with Dirichlet boundary conditions $p_i = \rho g h_i$ on Γ using a multigrid solver as described in [3]. Then 1,000 steps of

[1] The authors would like to thank C. Grümme and H. Berninger for their help with this example.

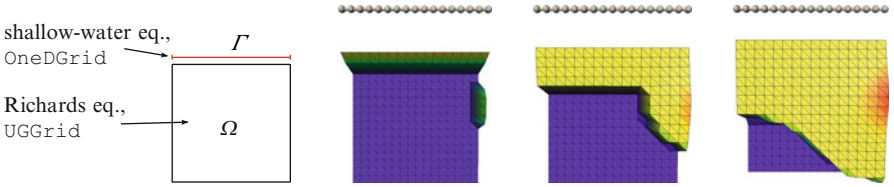


Fig. 3 Coupling the Richards equation to the shallow-water equation

the shallow-water equation are computed using a Lax–Friedrichs scheme. The flow $\mathbf{v}_i \cdot \mathbf{n}$ of subsurface water across Γ is interpolated in time and used as the source term in (4).

The Richards equation is discretized on a uniform triangle grid using the `UGGrid` grid manager. For the shallow water equation a `OneDGrid` is used. From the `UGGrid`, the interface Γ is extracted using a `Codim1Extractor` and the entire `OneDGrid` is extracted with a `Codim0Extractor`. A transformation $\tau : \mathbb{R} \to \mathbb{R}^2$ is given to the `Codim0Extractor` that places the 1d grid on the coupling boundary Γ such that the grids match. The `ConformingMerge` backend is used to generate the remote intersections. Figure 3 shows several steps in the evolution of the problem.

References

1. Bastian, P., Blatt, M., Dedner, A., Engwer, C., Klöfkorn, R., Kornhuber, R., Ohlberger, M., Sander, O.: A generic interface for parallel and adaptive scientific computing. Part II: Implementation and tests in DUNE. Computing **82**(2–3), 121–138 (2008)
2. Bastian, P., Blatt, M., Dedner, A., Engwer, C., Klöfkorn, R., Ohlberger, M., Sander, O.: A generic interface for parallel and adaptive scientific computing. Part I: Abstract framework. Computing **82**(2–3), 103–119 (2008)
3. Berninger, H., Kornhuber, R., Sander, O.: Fast and robust numerical solution of the Richards equation in homogeneous soil. SIAM Journal on Numerical Analysis, (2010), submitted
4. Eck, C.: Existenz und Regularität der Lösungen für Kontaktprobleme mit Reibung. Ph.D. thesis, Universität Stuttgart (1996)
5. Gräser, C., Sack, U., Sander, O.: Truncated nonsmooth Newton multigrid methods for convex minimization problems. In: M. Bercovier, M. Gander, R. Kornhuber, O. Widlund (eds.) Domain Decomposition Methods in Science and Engineering XVIII, LNCSE. Springer, Berlin (2009)
6. Krause, R., Sander, O.: Automatic construction of boundary parametrizations for geometric multigrid solvers. Comp. Vis. Sci. **9**, 11–22 (2006)
7. Quarteroni, A., Valli, A.: Domain Decomposition Methods for Partial Differential Equations. Oxford Science Publications (1999)
8. Sander, O.: A fast solver for finite deformation contact problems. Tech. Rep. 319, DFG Research Center Matheon (2006)
9. Sander, O.: Multidimensional coupling in a human knee model. Ph.D. thesis, Freie Universität Berlin (2008)
10. Wohlmuth, B., Krause, R.: Monotone methods on nonmatching grids for nonlinear contact problems. SIAM J. Sci. Comp. **25**(1), 324–347 (2003)

FEM for Flow and Pollution Transport in a Street Canyon

Petr Bauer, Atsushi Suzuki, and Zbyněk Jaňour

Abstract We develop a mathematical model of air flow and pollution transport in a 2D street canyon. The model is based on Navier–Stokes equations for viscous incompressible flow and convection–diffusion equation describing pollution transport. The solution is obtained by means of finite element method (FEM). We use the non-conforming Crouzeix–Raviart elements for velocity, the piecewise constant elements for pressure, and the piecewise linear elements for concentration. The resulting linear systems are solved by multigrid methods. We present computational studies of air flow and pollutant dispersion.

1 Introduction

We consider a polygonal domain $\Omega \subset \mathrm{R}^2$ which represents a vertical cut through a street canyon (Fig. 1). The domain is derived from a rectangle by substitution of the bottom edge by a piecewise linear line representing the terrain. The boundary of the domain consists of "inlet," "terrain," "outlet" and "upper" parts.

Combining the incompressible Navier–Stokes equations for air flow and the convection–diffusion equation for concentration, we obtain the following system of equations for pollution transport in $[0, T] \times \Omega$:

$$\frac{\partial c(t,x)}{\partial t} + \mathbf{u}(t,x) \cdot \nabla c(t,x) - D \triangle c(t,x) = f(t,x)$$

P. Bauer (✉)
FNSPE, CTU Prague
e-mail: bauerp@kmlinux.fjfi.cvut.cz

A. Suzuki
FNSPE, CTU Prague; Faculty of Mathematics, Kyushu University
e-mail: asuzuki@math.kyushu-u.ac.jp

Z. Jaňour
Institute of Thermomechanics, CAS Prague
e-mail: janour@it.cas.cz

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_11,

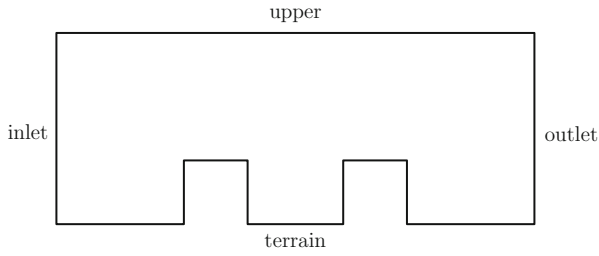


Fig. 1 Street canyon – 2D cut

$$\frac{\partial \mathbf{u}(t,x)}{\partial t} + \mathbf{u}(t,x)\cdot\nabla\mathbf{u}(t,x) - \nu\triangle\mathbf{u}(t,x) + \nabla p(t,x) = 0$$
$$\nabla\cdot\mathbf{u}(t,x) = 0$$

$$c(0,x) = c_0(x) \quad x\in\Omega$$
$$\mathbf{u}(0,x) = \mathbf{u_0}(x) \quad x\in\Omega$$

For velocity $\mathbf{u}$, we set homogeneous Dirichlet boundary condition on the terrain, Poiseuille profile on the inlet, Neumann condition on the outlet, and slip condition on the upper boundary.

The question of an appropriate boundary condition for concentration c is non-trivial, see [1]. For simplicity, we consider homogeneous Dirichlet boundary condition on the inlet, and Neumann boundary condition on the terrain, outlet and upper boundary. The term $f(t,x)$ represents the pollution source.

2 Numerical Scheme

In case of low concentrations, we can neglect the pollutant overall momentum, and solve the systems for velocity and concentration separately using a passive transport model.

2.1 Weak Formulation of Navier–Stokes Equations

Let $X = (H^{(1)}(\Omega))^2$, $V(\mathbf{u}_{\text{in}}) = \{\mathbf{u}\in X : \mathbf{u}|_{\text{terrain}} = \mathbf{0}, \mathbf{u}|_{\text{inlet}} = \mathbf{u}_{\text{in}}, \mathbf{u}|_{\text{upper}}\cdot\mathbf{n} = 0\}$, $Q = L^2(\Omega)$. We set the following forms:

$$(\nabla\mathbf{u},\nabla\mathbf{v}) = \int_\Omega \sum_{i,j=1}^{2} \frac{\partial u_i}{\partial x_j}\frac{\partial v_i}{\partial x_j}, \quad b(\mathbf{u},\mathbf{v},\mathbf{w}) = \frac{1}{2}\int_\Omega \sum_{i,j=1}^{2}\left(u_j\frac{\partial v_i}{\partial x_j}w_i - u_j v_i\frac{\partial w_i}{\partial x_j}\right).$$

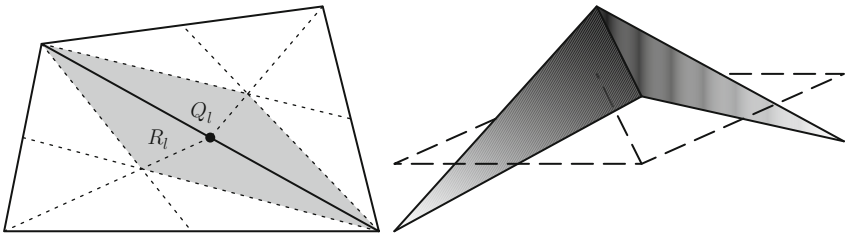


Fig. 2 (**a**) Lumped regions (**b**) Crouzeix–Raviart element

We use the backward Euler difference for the time derivative $\frac{\partial \mathbf{u}(t^n,x)}{\partial t} \approx \frac{\mathbf{u}^n - \mathbf{u}^{n-1}}{\tau}$ where $t^n = n\tau$. For each timestep t^n, we seek $\mathbf{u}^n \in V(\mathbf{u}_{\text{in}})$ and $p^n \in Q$, such that $\forall \mathbf{v} \in V(\mathbf{0})$, $\forall q \in Q$:

$$(\mathbf{u}^n, \mathbf{v}) + \tau b(\mathbf{u}^{n-1}, \mathbf{u}^n, \mathbf{v}) + \tau(\nabla \mathbf{u}^n, \nabla \mathbf{v}) - \tau(p^n, \nabla \cdot \mathbf{v}) = (\mathbf{u}^{n-1}, \mathbf{v})$$
$$(q, \nabla \cdot \mathbf{u}^n) = 0$$

Let index h denote the respective finite-dimensional spaces $V^h(\mathbf{u}_{\text{in}})$, Q^h, and the corresponding functions $\mathbf{u}_h^n$, p_h^n. We use the upwinding technique proposed by [5], based on dual elements R_l given by the barycentric nodes of the original mesh (Fig. 2).

We introduce $\mathbf{w}_h \in V^h(\mathbf{u}_{\text{in}})$ to respresent inhomogeneous Dirichlet data. Taking $\mathbf{v}_h = \mathbf{u}_h - \mathbf{w}_h \in V^h(\mathbf{0})$, the discrete problem for each timestep t^n rewritten in the matrix form stands:

$$\mathbf{M}\mathbf{v}_h^n + \tau \mathbf{N}(\mathbf{u}_h^{n-1})\mathbf{v}_h^n + \tau \mathbf{A}\mathbf{v}_h^n + \tau \mathbf{B}^{\mathbf{T}} p_h^n = \tilde{\mathbf{f}},$$
$$\mathbf{B}\mathbf{v}_h^n = \tilde{\mathbf{g}},$$

where

$$\tilde{\mathbf{f}} = \mathbf{M}(\mathbf{v}_h^{n-1} + \mathbf{w}_h^{n-1} - \mathbf{w}_h^n) - \tau \mathbf{N}(\mathbf{u}_h^{n-1})\mathbf{w}_h^n - \tau \mathbf{A}\mathbf{w}_h^n,$$
$$\tilde{\mathbf{g}} = -\mathbf{B}\mathbf{w}_h^n.$$

2.2 Weak Formulation of Convection–Diffusion Equation

Let $X = H^{(1)}(\Omega)$, $V(c_{\text{in}}) = \{c \in X : c\,|_{\text{inlet}} = c_{\text{in}}\}$. To avoid difficulties with oscillating schemes, we employ the Characteristic Galerkin method [4]. By introducing the mapping $\varphi^n(x) = x - \tau \mathbf{u}^n(x)$ representing convection, and setting $\frac{\partial c(t^n,x)}{\partial t} + \mathbf{u} \cdot \nabla c(t^n, x) \approx \frac{c^n - c^{n-1} \circ \varphi^n}{\tau}$, we get the elliptic problem [2] for each timestep $t^n = n\tau$:

$$(c^n, v) + \tau D(\nabla c^n, \nabla v) = (\tau f^n + c^{n-1} \circ \varphi^n, v) \quad \forall v \in V(0).$$

3 Numerical Solution Using FEM

We choose the non-conforming Crouzeix–Raviart elements (Fig. 2) to approximate the components of velocity, the piecewise constant elements for pressure, and the piecewise linear elements for concentration.

We use multigrid solvers based on Vanka-type and Gauss–Seidel smoothers to solve the respective linear systems. An extension for higher order elements can be found in [3].

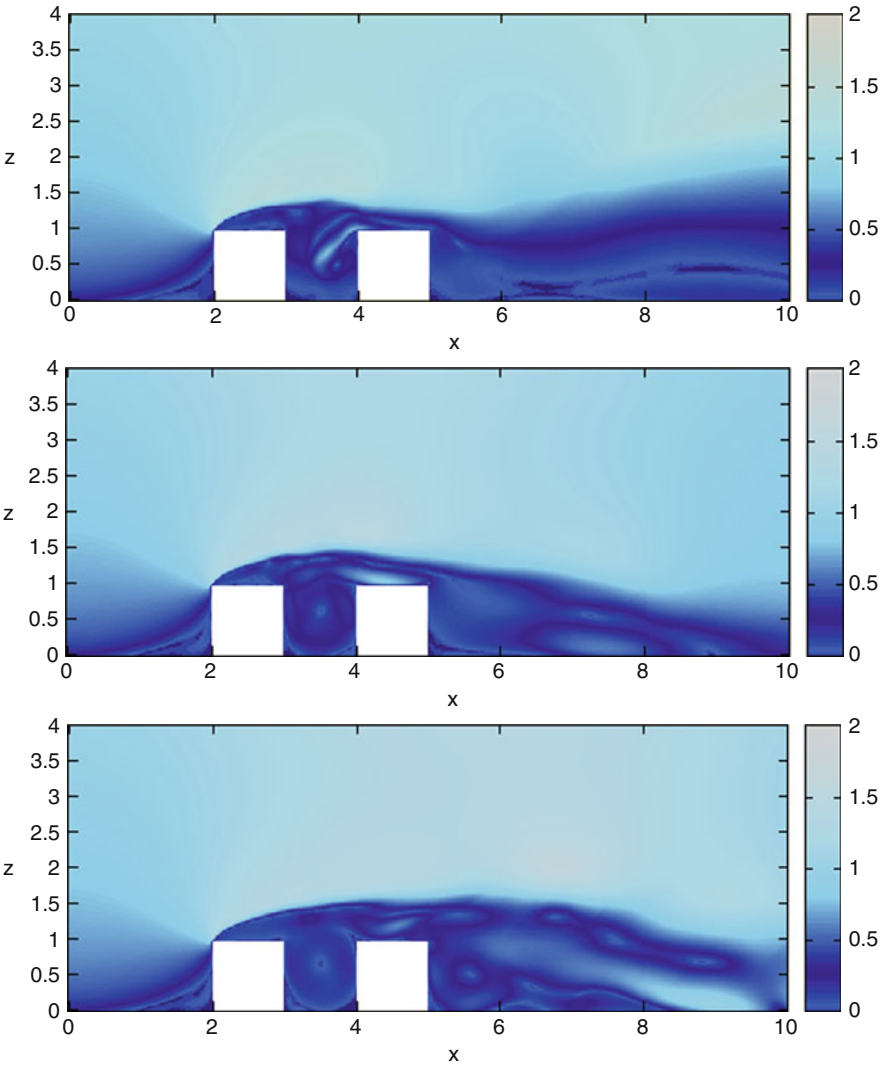


Fig. 3 $|u(t)|$ at time $t = 8, 16, 32$

4 Numerical Results

We consider three different configurations of a street canyon. Two with the canyon of the same size as the buildings, and the last one twice as wide. The Reynolds number is $Re = 10^4$ for all cases.

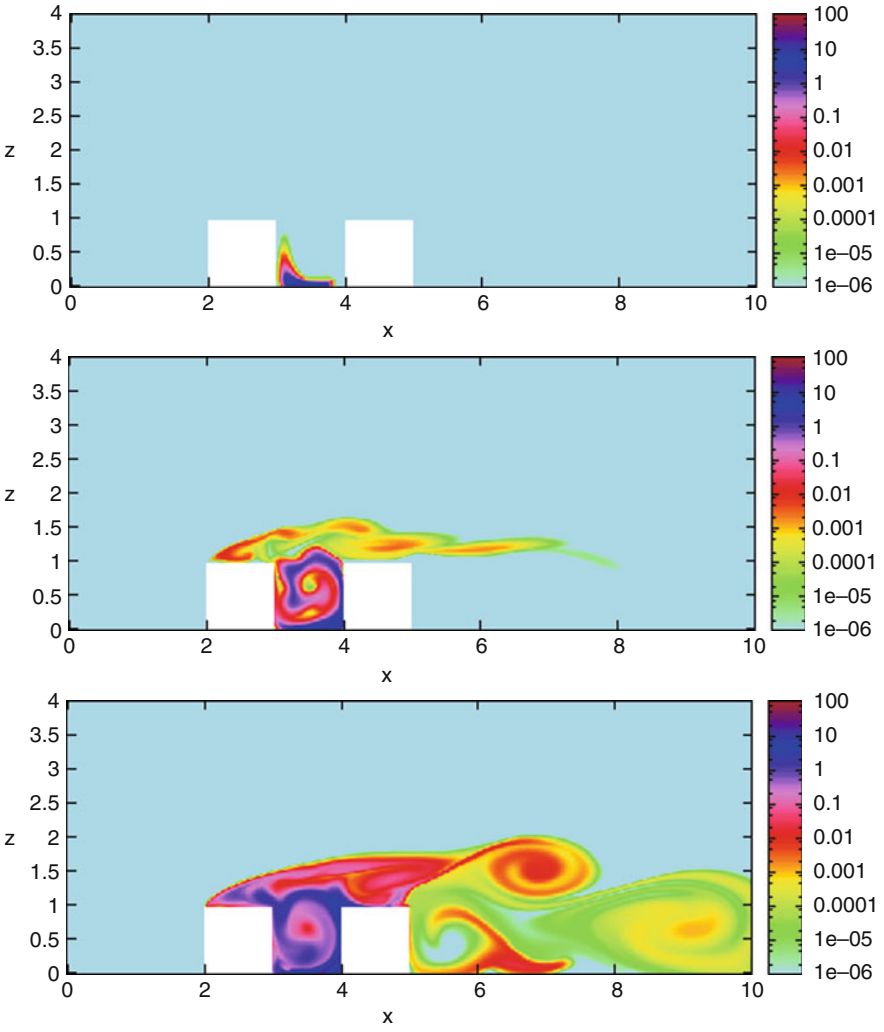


Fig. 4 $c(t)$ at time $t = 8, 16, 32$

4.1 Example: Square Canyon

This is the basic configuration with two buildings forming a square canyon. We place a constant source of pollution at the bottom of the canyon. The absolute values of velocity are displayed in linear scale (Fig. 3), whereas the concentration levels are displayed in logarithmic scale (Fig. 4).

4.2 Example: Wide Canyon

We consider a rectangular canyon with 2 : 1 ratio. The other settings remain the same as in the previous case.

4.3 Example: Two Consecutive Canyons

This example demonstrates the difference between the flow in the first and the second canyon (Fig. 5).

5 Conclusion

We obtained the computational results of pollution transport in a 2D street canyon, which can be compared with experimental data from the environmental wind tunnel at the Institute of Thermomechanics of the Czech Academy of Sciences. The current choice of stationary inlet profile is inadequate, and we need to consider more realistic, fluctuating velocity profiles to catch the turbulent properties of the atmospheric boundary layer.

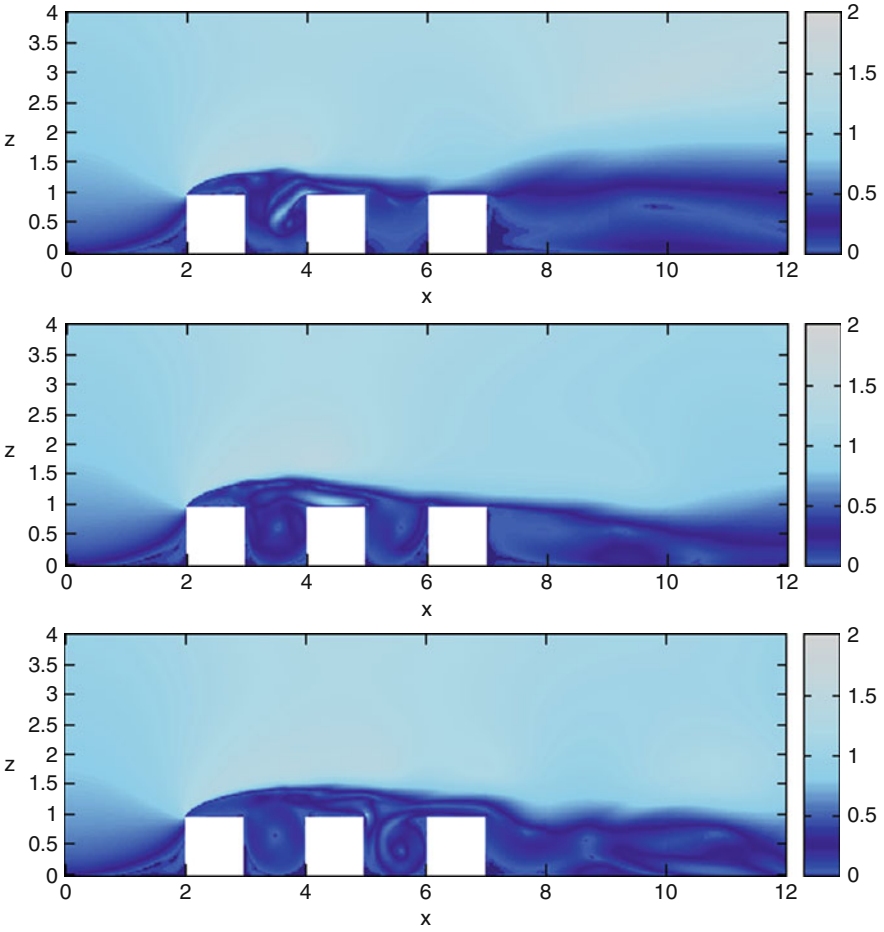
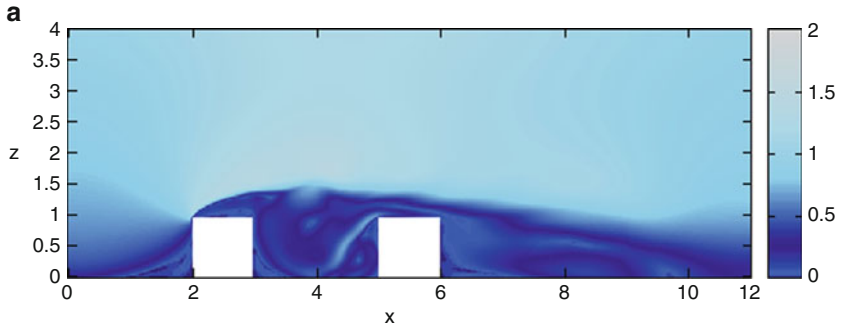
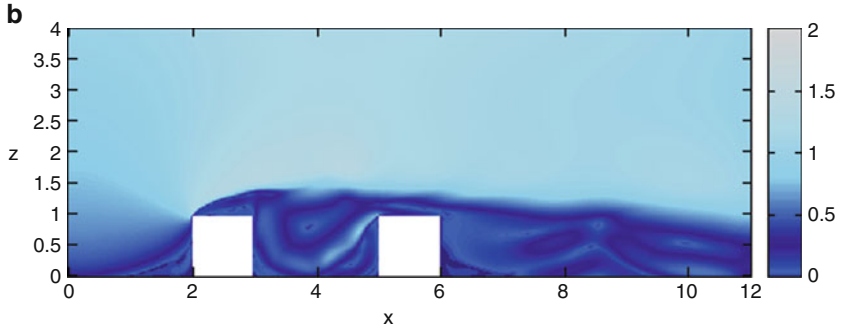


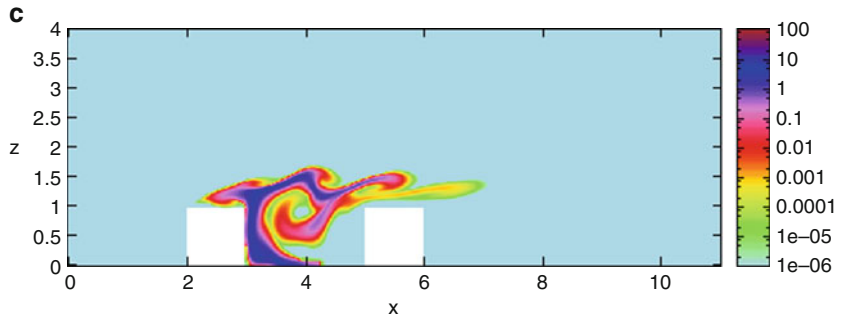
Fig. 5 $|u(t)|$ at time $t = 8, 16, 32$



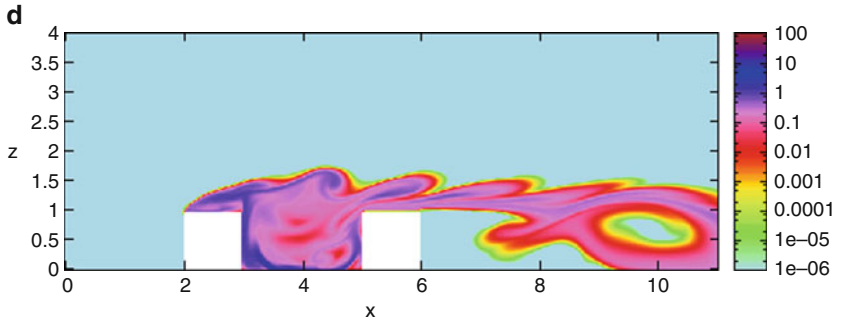
(a) $|u(t)|$ at time $t = 16$



(b) $|u(t)|$ at time $t = 32$



(c) $c(t)$ at time $t = 16$



(d) $c(t)$ at time $t = 32$

Acknowledgements This work was carried out within the project No. MSM 6840770010 "Applied Mathematics in Technical and Physical Sciences" of the Ministry of Education, Youth and Sports of the Czech Republic.

References

1. M. Beneš, R.F. Holub, *Aerosol wall deposition in enclosures investigated by means of a stagnant layer,* Environment International, 22 Suppl. 1, 883–889 (1996)
2. P.G. Ciarlet, *The finite-element method for elliptic problems,* North-Holland, Amsterdam (1978)
3. V. John, P. Knobloch, G. Matthies, L. Tobiska, *Non-nested multi-level solvers for finite element discretisations of mixed problems,* Computing, 68, 313–341 (2002)
4. J. Kačur, *Solution of degenerate convection-diffusion problems by the method of characteristics,* SIAM Journal on Numerical Analysis, 39, 858–879 (2001)
5. F. Schieweck, L. Tobiska, *An optimal order error estimate for upwind discretization of the Navier–Stokes equation,* Numerical methods in partial differential equations, 12(4), 407–421 (1996)

Stabilized Finite Element Methods with Shock-Capturing for Nonlinear Convection–Diffusion-Reaction Models

Markus Bause

Abstract In this work stabilized higher-order finite element approximations of convection-diffusion-reactions models with nonlinear reaction mechanisms are studied. Streamline upwind Petrov–Galerkin (SUPG) stabilization together with anisotropic shock-capturing as an additional stabilization in crosswind-direction is used. The parameter design of the scheme is described precisely and error estimates are provided. Theoretical results are illustrated by numerical computations. The work extends former investigations for linear problems to more realistic nonlinear models.

1 Introduction

Time-dependent nonlinear convection-diffusion-reaction problems are often studied in various technical and environmental applications. The accurate and reliabel numerical simulation of such processes is still a challenging task. The model equations are strongly coupled such that inaccuracies in one unknown directly affect all other unknowns. In large chemical systems with complex reaction mechanisms and interactions these numerical artifacts can lead to completely wrong predictions; cf. [2]. Typically, the transport systems are convection- and/or reaction-dominated and characteristic solutions have sharp layers. In these cases standard finite element methods cannot be applied. Modified finite element aproaches are required that are able to handle sharp layers and prevent the occurence of spurious oscillations.

The streamline upwind Petrov–Galerkin (SUPG) method (cf. [4]) is capable to stabilize most of the unphysical oscillations of finite element discretizations which are caused by dominating convection. Nevertheless, spurious localized oscillations in particular in crosswind-direction may still be present. As a remedy, discontinuity- or shock-capturing variants of SUPG stabilized schemes as an additional consistent

M. Bause
Department of Mechanical Engineering, Helmut Schmidt University, University of the Federal Armed Forces Hamburg, Holstenhofweg 85, 22043 Hamburg, Germany
e-mail: bause@hsu-hh.de

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_12,

stabilization have been proposed in the literature. For the efficiency of these methods the design of the inherent parameter is of importance. Various linear and nonlinear realizations of shock-capturing methods have been considered. For an overview of these techniques we refer to a recent work of John and Schmeyer [6]. Most of these methods were derived and studied for steady-state linear boundary value problems only; cf., e.g., [7, 9]. Even convection-diffusion problems without reaction are assumed often. However, mathematical models describing reactive transport phenomena lead to systems of instationary convections-diffusions-reaction equations with nonlinear reactive terms coupling the set of equations. Therefore, the convergence analysis and parameter design of SUPG stabilized schemes with shock-capturing need to be generalized. This has been done by the author in a recent work [1]. For the lack of space, we consider a steady-state nonlinear model problem here. Error estimates that are proved in [1] are summarized. Then these estimates and the capability of an anisotropic shock-capturing stabilization technique to further reduce oscillations in crosswind direction are illustrated by numerical experiments.

Whereas in [6] stabilized finite element discretizations are studied in the context of linear finite element methods, we consider using higher-order finite element approaches in this work. In particular for reactive multicomponent transport systems, higher-order methods have demontrated to be superior to linear finite element approximations; cf. [2]. They have shown to be less diffusive and help to prevent an artifical mixing of chemical species and to increase the accuracy of simulations. For linear finite element discretizations, flux-corrected transport methods (cf. [8]) that work on an algebraic level and not on the weak formulation of the partial differential equation as the SUPG approach offer an alternative. In [6], the most accurate results of all considered schemes were obtained for the flux-corrected transport methods.

As a model problem for our investigations we consider solving

$$L(u) := \alpha u + b \cdot \nabla u - \nabla \cdot (a \nabla u) + r(u) = f \quad \text{in } \Omega\,, \qquad u = 0 \quad \text{on } \partial\Omega\,. \tag{1}$$

In (1), let $\alpha \in \mathbb{R}$, $a \in L^\infty(\Omega)$, $\mathbf{b} \in \mathbf{H}^1(\Omega) \cap \mathbf{L}^\infty(\Omega)$ and $f \in L^2(\Omega)$ with $\alpha > 0$, $(\nabla \cdot \mathbf{b})(x) = 0$, $a(x) \geq a_0 > 0$ almost everywhere in $\Omega \subset \mathbb{R}^2$ or $\mathbb{R}^3$. Equations (1) can be considered as the semidiscrete problem arising from the discretization in time of an evolution equation of convection-diffusion-reaction type. The parameter α mimics the temporal discretization parameter $1/\Delta t$ where Δt is the time step size. Therefore, we assume that $\alpha > 0$ can be prescibed and be chosen arbitrarily large. We make the following assumption about the parametrization $r(\cdot)$:

$$r \in C^1(\mathbb{R}_0^+)\,, \qquad r(0) = 0\,, \qquad r'(s) \geq r_0 \geq 0 \quad \text{for } s \geq 0\,,\ s \in \mathbb{R}\,. \tag{2}$$

Conditions (2) can be further weakened. For the sake of brevity, this is not done here. Further, we suppose that the solution u of (1) is non-negative and bounded,

$$0 =: u_0 \leq u \leq u_1 \quad \text{a.e. in } \Omega\,, \tag{3}$$

which is reasonable from the point of view of physics, for instance, if u denotes the concentration of a chemical species. Throughout this work we use standard notation.

2 Stabilized Finite Element Approximation with Shock-Capturing

In this section we briefly introduce the higher-order finite element scheme with anisotropic shock-capturing that we will use to approximate solutions of (1) and study further. A standard framework of an hp-version of the finite element method is assumed; cf. [9]. In particular, for a family of admissible and sharpe-regular triangulations $\mathscr{T}_h = \{T\}$ of the polyhedral domain $\Omega \subset \mathbb{R}^d$, with $d = 2$ or 3, let

$$V_h^{\mathbf{p}} = X_h^{\mathbf{p}} \cap H_0^1(\Omega)\,, \quad \text{with} \quad X_h^{\mathbf{p}} = \{v \in C(\overline{\Omega}) \mid v_{|T} \circ F_T \in \mathscr{P}_{p_T}(\widehat{T})\ \forall T \in \mathscr{T}_h\}\,,$$

denote the underlying finite element space of piecewise polynomials of local order p_T for all $T \in \mathscr{T}_h$. Here, $\widehat{T}$ is the (open) unit simplex or the (open) unit hypercube in $\mathbb{R}^d$ and $\mathscr{P}_n(\widehat{T})$, with $n \geq 1$, is the set of all polynomials of degree at most n on $\widehat{T}$. We assume that each $T \in \mathscr{T}_h$ is a smooth bijective image of $\widehat{T}$, i.e., $T = F_T(\widehat{T})$. The vector $\mathbf{p}$ is defined by $\mathbf{p} = \{p_T \mid T \in \mathscr{T}_h\}$.

Then, the SUPG-stabilized approximation of (1) reads as: *Find* $u_h \in V_h^{\mathbf{p}}$ *such that*

$$A_s(u_h, v_h) = L_s(v_h) \tag{4}$$

for all $v_h \in V_h^{\mathbf{p}}$, *where*

$$A_s(u, v) = \widehat{A}(u, v) + \sum_{T \in \mathscr{T}_h} \delta_T \langle \widehat{L}u, \mathbf{b} \cdot \nabla v \rangle_{L^2(T)}\,, \tag{5}$$

$$L_s(v) = \langle f, v \rangle + \sum_{T \in \mathscr{T}_h} \delta_T \langle f, \mathbf{b} \cdot \nabla v \rangle_{L^2(T)}\,, \tag{6}$$

$$\widehat{A}(u, v) = A_{\text{lin}}(u, v) + \langle \widehat{r}(u), v \rangle\,, \quad \widehat{L}u = \widehat{L}_{\text{lin}}u + \widehat{r}(u)\,, \tag{7}$$

$$\widehat{L}_{\text{lin}}u_{|T} = \alpha u + \mathbf{b} \cdot \nabla u - \nabla \cdot \Pi_T(a \nabla u)\,. \tag{8}$$

If shock-capturing is applied, the discrete problem reads as: *Find* $u_h \in V_h^{\mathbf{p}}$ *such that*

$$A_s(u_h, v_h) + A_{\text{sc}}(u_h; u_h, v_h) = L_s(v_h) \tag{9}$$

for all $v_h \in V_h^{\mathbf{p}}$, *where*

$$A_{\text{sc}}(w; u, v) := \sum_{T \in \mathscr{T}_h} \langle \tau_T(w) \mathbf{D}_{\text{sc}} \nabla u, \nabla v \rangle\,. \tag{10}$$

Together, the second terms on the right-hand sides of (5) and (6), respectively, represent the SUPG-stabilization. The choice of the stabilization parameter δ_T is given in (15) below. In (7) we changed $r(\cdot)$ to $\widehat{r}(\cdot)$ where

$$\widehat{r}(u) = \begin{cases} r(u_0) + r'(u_0)(u-u_0) & \text{for } u \le u_0\,, \\ r(u) & \text{for } u_0 \le u \le u_1\,, \\ r(u_1) + r'(u_1)(u-u_1) & \text{for } u \ge u_1\,. \end{cases} \tag{11}$$

This modification is necessary to prove an error estimates when r' grows with $|u|$ or even stronger. Since $r'(u)$ is bounded above compact intervals of u, the function $\widehat{r}$ is Lipschitz continuous, i.e., there exists some constant $L_r > 0$ such that

$$|\widehat{r}(u) - \widehat{r}(v)| \le L_r |u-v| \quad \forall u, v \in \mathbb{R}\,. \tag{12}$$

In (8), the mapping $\Pi_T : \mathbf{L}^2(\Omega) \mapsto (\mathscr{P}_{p_T}(T))^d$ denotes the (elementwise) orthogonal projection onto $(\mathscr{P}_{p_T}(T))^d$. This modifcation is necessary in order to allow variable diffusion coefficients and to apply an inverse inequality; cf. (18). We use an anisotropic variant of shock-capturing that is proposed in [3] by choosing

$$\left.\begin{aligned} \mathbf{D}_{\mathrm{sc}} := \begin{cases} \mathbf{I} - \dfrac{\mathbf{b}\otimes\mathbf{b}}{|\mathbf{b}|^2}\,, & \mathbf{b} \neq \mathbf{0} \\ \mathbf{0}\,, & \mathbf{b} = \mathbf{0} \end{cases}, \quad \tau_T(w) := l_T(w) R_T^*(w) \equiv \frac{l_T(w) R_T(w)}{|w|_{H^1(T)} + \kappa}\,, \\ R_T(w) := \|\widehat{L}w - f\|_{L^2(T)}\,, \qquad l_T(w) := l_0 h_T \max\left\{0, \beta - \frac{2\|a\|_{L^\infty(T)}}{h_T R_T^*(w)}\right\} \end{aligned}\right\} \tag{13}$$

in (10). The non-negtive limiter function $\tau_T(w)$ aims to restrict the effect of shock-capturing to subregions where the residual $\widehat{L}w - f$ is too large. The term $\frac{h_T R_T^*(w)}{2\|a\|_{L^\infty(T)}}$ can be seen as a pseudo mesh Peclet number. The choice of l_0, κ and β is given in Sect. 4. We note that $\tau_T(u_h)$ depends nonlinearly on the discrete solution u_h. Since the reaction rate $r(\cdot)$ is assumed to be nonlinear, the shock-capturing technique (9), (10), (13) does not change the type of the discrete problem. This is in contrast to linear convection-diffusion-reaction models that become nonlinear by adding the shock-capturing term (10), (13) which increases strongly the cost for solving the discrete system. Under the above-made assumptions problems (4) and (9) admit solutions $u_h \in V_h^{\mathbf{p}}$. This can be shown by Brouwer's fixed point theorem; cf. [1]. To solve (9), we use an inexact variant of Newton's method.

3 Error Estimates

Next, we recall some error estimates that are proved in [1] for the given numerical schemes (4) and (10). Moreover, we describe the choice of the stabilization parameter δ_T. An appropriate norm for analyzing the scheme (4)–(8) is given by

$$|||v||| := \left(\sum_{T \in \mathscr{T}_h} \left(\|\sqrt{a}\nabla v\|^2_{\mathbf{L}^2(T)} + (\alpha + r_0)\|v\|^2_{L^2(T)} + \delta_T \|\mathbf{b} \cdot \nabla v\|^2_{L^2(T)} \right) \right)^{1/2}, \tag{14}$$

where r_0 has to be chosen as in assumption (2). For the shock-capturing approach (9) an additional error control in crosswind direction is obtained; cf. Theorem 2.

First, for the SUPG-stabilized finite element method we have the following result.

Theorem 1. *Let $u \in H^1_0(\Omega)$ be the solution of (1) with $u \in H^{k_T}(T)$, $k_T > d/2$, and $a \in W^{k_T-1,\infty}(T)$ for all $T \in \mathscr{T}_h$. Suppose that $\alpha + r_0 > 96L_r$ is satisfied and that the stabilization parameter δ_T in (5), (6) is chosen of the order of magnitude*

$$\delta_T \sim \min \left\{ \frac{h_T}{p_T \|\mathbf{b}\|_{\mathbf{L}^\infty(T)}}; \frac{h_T^2}{p_T^4 \mu_{\text{inv}}^2 \|a\|_{L^\infty(\Omega)}}; \frac{1}{\alpha + r_0}; \frac{\alpha + r_0}{L_r^2} \right\}. \tag{15}$$

Then, for the SUPG-stabilized finite element approximation (4)–(8) it holds that

$$|||u - u_h|||^2 \leq C_{\text{SUPG}} \sum_{T \in \mathscr{T}_h} \frac{h_T^{2(l_T-1)}}{p_T^{2(k_T-1)}} M_T^{\text{opt}} \|u\|^2_{H^{k_T}(T)} \tag{16}$$

with

$$\begin{aligned} M_T^{\text{opt}} := \|a\|_{L^\infty(T)} \Big(& 1 + \frac{\|a\|^2_{W^{k_T-1}(T)}}{\|a\|^2_{L^\infty(T)}} + Pe_T + \Gamma_T^{(1)} \\ & + \min \left\{ \frac{\|a\|_{L^\infty(T)}}{\alpha_T} Pe_T^2; \max\{Pe_T; \Gamma_T^{(1)}; \Gamma_T^{(2)}; p_T^2 \mu_{\text{inv}}^2\} \right\} \Big) \end{aligned} \tag{17}$$

and the characteristic numbers

$$Pe_T := \frac{h_T \|\mathbf{b}\|_{\mathbf{L}^\infty(T)}}{p_T \|a\|_{L^\infty(T)}}, \quad \Gamma_T^{(1)} := \frac{\alpha h_T^2}{p_T^2 \|a\|_{L^\infty(T)}}, \quad \Gamma_T^{(2)} := \frac{L_r^2 h_T^2}{(\alpha + r_0) p_T^2 \|a\|_{L^\infty(T)}}.$$

In (17), the parameter μ_{inv} denotes the constant of the local inverse inequality

$$\|\nabla w\|_{\mathbf{L}^2(T)} \leq \mu_{inv} p_T^2 h_T^{-1} \|w\|_{L^2(T)} \quad \forall w \in X_h^p \tag{18}$$

on $T \in \mathscr{T}_h$ and depends on the shape-regularity parameter of the triangulation. The parameter r_0 and L_r in (15) are defined by (2) and (12), respectively. The error estimate (16) is of quasi-optimal order. The L^2-part in inequality (16) is only suboptimal. The condition $\alpha + r_0 > 96L_r$ comes from using absorption arguments along with Cauchy–Young's inequality. We do not believe that this condition is really sharp. Nevertheless, increasing α has in impact on the numerical performance properties. It improves the convergence behavior of the Newton iteration for solving the nonlinear equation (9).

Now, for the shock-capturing approach the following result is obtained.

Theorem 2. *Let the assumptions of Theorem 1 be satisfied. Suppose that* $\alpha + r_0 > \frac{1280}{11} L_r \approx 116.36 L_r$ *is satisfied and that the stabilization parameter* δ_T *in (9) is chosen of the same order of magnitude as in (15). Then, for the SUPG-scheme with shock-capturing (9), (10), (13) the error estimate*

$$|||u-u_h|||^2 + \sum_{T\in\mathscr{T}_h} \tau_T(u_h) \left\| \mathbf{D}_{\mathrm{sc}}^{1/2} \nabla u_h \right\|_{L^2(T)}^2 \leq C_{\mathrm{SC}} \sum_{T\in\mathscr{T}_h} \frac{h_T^{2(l_T-1)}}{p_T^{2(k_T-1)}} M_T^{\mathrm{opt}} \|u\|_{H^{k_T}(T)}^2$$

with the same parameter M_T^{opt} *as in (17) is satisfied.*

Theorem 2 shows that our anisotropic shock-capturing technique provides an additional error control in crosswind-direction. Asymptotically, the same rate of convergence as for the SUPG-scheme without shock-capturing is obtained. The difference comes only through the error constant. However, a slightly severer condition is imposed on the parameter α. We do not know if this condition is really sharp.

4 Numerical Experiments

In this section we shall illustrate the error estimates given in Sect. 3 by numerical computations. In particular, we show that shock-capturing reduces spurious oscillations in crosswind-direction. Moreover, we illustrate the positive impact of using higher-order finite element methods on the accuracy of the numerical results.

Example 4.1. Our first test problem is an adaption of Example 4.2 from [9] for the linear convection–diffusion equation. We consider problem (1) on $\Omega = (0,1)^2$ with $\alpha = 1.0$, $a = 10^{-6}$, $\mathbf{b}(\mathbf{x}) = \frac{1}{\sqrt{5}}(1,2)^\top$ and $r(u) = u^2$. The source f is chosen in such a way that $u(\mathbf{x}) = \frac{1}{2}\left(1 - \tanh \frac{2x_1 - x_2 - \frac{1}{4}}{\sqrt{5a}}\right)$ is the exact solution. It is characterized by an interior layer of thickness $\mathscr{O}(\sqrt{a}|\ln a|)$ around $2x_1 - x_2 = \frac{1}{4}$. We study the solutions of (4) and (9). In (13), we put $l_0 = 0.2$, $\kappa = 10^{-4}$ and $\beta = 0.7$.

For our computations we used the finite element toolbox ALBERTA [10]. Table 1 and 2 summarize the calculated errors for the L^2-norm and the streamline diffusion norm (14) and different P_p-elements for $p \in \{1,2,3,4\}$. Although the SUPG-scheme shows a slightly smaller error, we observe that the errors of the either schemes are of the same magnitude, as claimed in Theorem 2. The larger errors of the shock-capturing approach can be explained by its additional artificial crosswind-diffusion that however reduces the spurious oscillations. In Table 1 we do not observe the optimal uniform convergence rates given in Theorem 1 and 2, respectively. The reason for this is that the solution u depends on the diffusion parameter a. In such cases the optimal convergence rates are observed for very small step sizes only.

To study the effects of crosswind-diffusion more precisely, the crosswind-diffusion parameter $\tau_T(u_h)$ is presented in Fig. 1. We nicely observe, that the additional diffusion is located around the layer. As expected, no additional diffusion is

Table 1 Example 4.1: Mesh size, number of degrees of freedom, errors in $\|\cdot\|_{L^2(\Omega)}$ (*left*) and $|||\cdot|||$ (*right*) and convergence rates for the SUPG-scheme without (SUPG) and with (SC-CD) shock-capturing and h-refinement

h	d.o.f.	$p=2$, $\|\cdot\|_{L^2(\Omega)}$			
		SUPG		SC-CD	
1.77e-1	145	1.34e-1	–	1.43e-1	–
8.84e-2	545	9.58e-2	0.48	1.02e-1	0.49
4.42e-2	2113	6.88e-2	0.48	7.38e-2	0.46
2.21e-2	8321	4.79e-2	0.52	5.22e-2	0.50
1.10e-2	33025	3.24e-2	0.56	3.61e-2	0.53
5.52e-3	131585	2.10e-2	0.63	2.42e-2	0.58

h	d.o.f.	$p=2$, $\|\|\|\cdot\|\|\|$			
		SUPG		SC-CD	
1.77e-1	145	1.51e-1	–	1.53e-1	–
8.84e-2	545	1.12e-1	0.43	1.15e-1	0.41
4.42e-2	2113	8.50e-2	0.40	8.71e-2	0.39
2.21e-2	8321	6.54e-2	0.38	6.70e-2	0.38
1.10e-2	33025	5.07e-2	0.37	5.18e-2	0.37
5.52e-3	131585	3.87e-2	0.39	3.95e-2	0.39

h	d.o.f.	$p=4$, $\|\cdot\|_{L^2(\Omega)}$			
		SUPG		SC-CD	
1.77e-1	545	8.57e-2	–	1.01e-1	–
8.84e-2	2113	6.22e-2	0.46	7.43e-2	0.44
4.42e-2	8321	4.47e-2	0.48	5.38e-2	0.46
2.21e-2	33025	3.11e-2	0.53	3.78e-2	0.51
1.10e-2	131585	1.86e-2	0.74	2.38e-2	0.67
5.52e-3	525313	8.63e-3	1.11	1.26e-2	0.92

h	d.o.f.	$p=4$, $\|\|\|\cdot\|\|\|$			
		SUPG		SC-CD	
1.77e-1	545	1.02e-1	–	1.13e-1	–
8.84e-2	2113	7.67e-2	0.41	8.45e-2	0.42
4.42e-2	8321	5.77e-2	0.41	6.35e-2	0.41
2.21e-2	33025	4.29e-2	0.43	4.74e-2	0.42
1.10e-2	131585	2.95e-2	0.54	3.36e-2	0.49
5.52e-3	525313	1.83e-2	0.69	2.19e-2	0.62

Table 2 Example 4.1: Errors in $\|\cdot\|_{L^2(\Omega)}$ and $|||\cdot|||$ and convergence rates for the SUPG-scheme without (SUPG) and with (SC-CD) shock-capturing and p-refinement; $h = 1.10\text{e-}2$

p	$\|\cdot\|_{L^2(\Omega)}$	
	SUPG	SC-CD
1	5.98611034e-2	6.29579483e-2
2	3.24058212e-2	3.61115581e-2
3	1.88279842e-2	2.41896732e-2
4	1.85794405e-2	2.37970677e-2

p	$\|\|\|\cdot\|\|\|$	
	SUPG	SC-CD
1	9.09154371e-2	9.17590477e-2
2	5.06827331e-2	5.18041693e-2
3	3.36901794e-2	3.57942179e-2
4	2.95139317e-2	3.36195918e-2

added away from the layer. Further, cross-section plots of the SUPG-method without and with shock-capturing in the crosswind-direction at $x_1 + 2x_2 = 1$ are also given in Fig. 1. Significant over- and undershoots of the SUPG-solution without shock-capturing in the neighborhood of the layer are observed. These unphysical oscillations are clearly damped with shock-capturing. The strong gradient of the SUPG solution in the layer is preserved. This underlines the proper construction of the limiter function of the shock-capturing approach. Moreover, for fixed $h > 0$ the resolution of the steep gradient is improved with increasing approximation order.

Example 4.2. Our second more sophisticated test problem (cf. Fig. 2) is an adaption of an example from [5, Sect. 4] for the linear convection-diffusion equation. We consider problem (1) on $\Omega = (0,1)^2$ with $\alpha = 1.0$, $a = 10^{-6}$, $\mathbf{b}(\mathbf{x}) = (-y, x)^\top$, $f \equiv 0$ and a Monod-type reaction rate $r(u) = -u/(1+u)$; cf. [2]. We prescribe the Dirichlet condition $u(x,y) = 1$ for $1/3 \le x \le 2/3$, $y = 0$ and $u(x,y) = 0$ on the remaining parts of the lower boundary as well as on the right and upper boundary. We put $\frac{\partial u(x,y)}{\partial \mathbf{n}} = 0$ for $x = 0$, $0 \le y \le 1$ where $\mathbf{n}$ is the

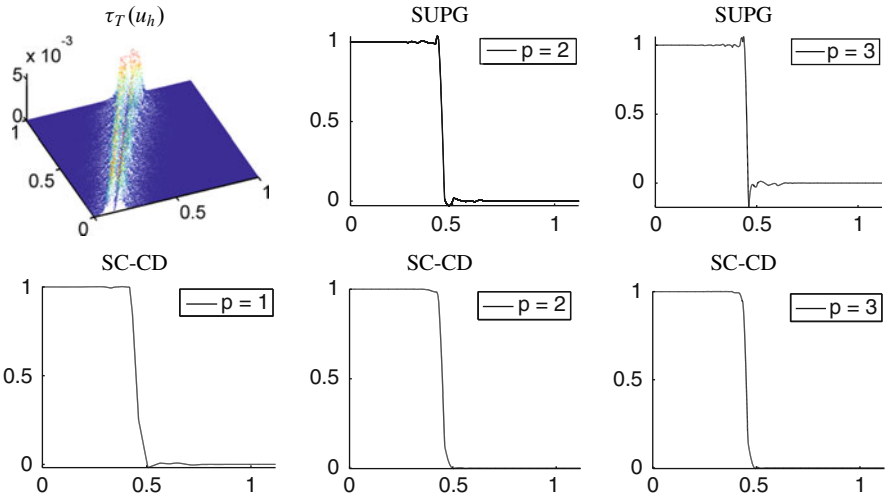


Fig. 1 Crosswind-diffusion parameter $\tau_T(u_h)$ for $p = 2$ and $h = 1/128$ (*top left*) and cross-section plots of the SUPG-solution without (*top*) and with (*bottom*) shock-capturing in crosswind-direction at $x_1 + 2x_2 = 1$ for $p = 2, 3$ (*top*) and $p = 1, 2, 3$ (*bottom*) and $h = 1/20$ for Example 4.1

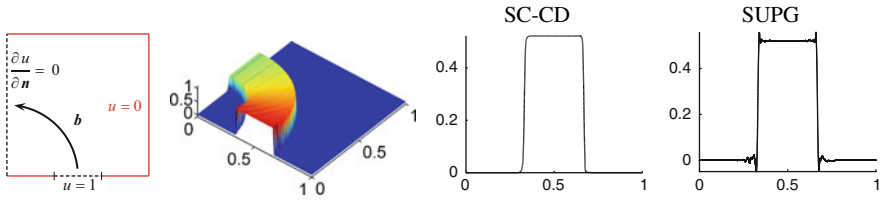


Fig. 2 Geometry, discrete solution u_h, cross-section plot of u_h at the left outflow boundary with (*SC-CD*) and without (*SUPG*) shock-capturing for Example 4.2; $p = 2$ and $h = 0.01$

unit outer normal. Figure 2 shows the calculated solution and its pofile at the outflow boundary. The solution has two small interior layers that are resolved by the SUPG method with shock-capturing. Again, by anisotropic shock-capturing stabilization over- and undershoots of the numerical approximation close to the layers are damped.

Summarizing, our numerical studies have shown that higher order finite element methods along with anisotropic shock-capturing techniques help to reduce spurious oscillations in crosswind-direction in the numerical approximation of convection-dominated transport equations with reaction. Shock-capturing stabilization increases the accuracy and reliability of simulations which is of great importance for coupled systems of equations modelling multicomponent reactive flow.

References

1. M. Bause, *Analysis of stabilized higher-order finite element approximation of nonstationary and nonlinear convection-diffusion-reaction equations*, to appear (2010), 1–32
2. M. Bause, P. Knabner, *Numerical simulation of contaminant biodegradation by higher order methods and adaptive time stepping*, Comput. Visual. Sci., **7** (2004), 61–78
3. R. Codina, O. Soto, *Finite element implementation of two-equation and algebraic stress turbulence models for steady incompressible flows*, Int. J. Numer. Meth. Fluids, **90** (1999), 309–343
4. T. Huge, M. Mallet, A. Mizukami, *A new finite element formulation for computational fluid dynamics: II. Beyond SUPG*, Comput. Meth. Appl. Mech. Eng., **54** (1986), 341–355
5. V. John, P. Knobloch, *On the performance of SOLD methods for convection-diffusion problems with interior layers*, Int. J. Comp. Sci. Math., **1** (2007), 245–258
6. V. John, E. Schmeyer, *Finite element methods for time-dependent convection-diffusion-reaction equations with small diffusion*, Comput. Methods Appl. Mech. Eng., **198** (2008), 475–494
7. T. Knopp, *Stabilized finite element methods with shock capturing for advection-diffusion problems*, Comput. Meth. Appl. Mech. Eng., **191** (2002), 2997–3013
8. D. Kuzmin, S. Turek, *Flux correction tools for finite elements*, J. Comput. Phys., **175** (2002), 525–558
9. G. Lube, G. Rapin, *Residual-based stabilized higher order FEM for advection-dominated problems*, Comput. Methods Appl. Mech. Eng., **195** (2006), 4124–4138
10. A. Schmidt, K. G. Siebert, *Design of Adaptive Finite Element Software*, Springer, Berlin, 2005

Finite Element Discretization of the Giesekus Model for Polymer Flows

Roland Becker and Daniela Capatina

Abstract We consider the Giesekus model for steady flows of polymeric liquids. This model, characterized by the presence in the constitutive law of a quadratic term in the stress tensor, yields a realistic behavior for shear, elongational and mixed flows. Its numerical approximation is achieved by means of Crouzeix–Raviart nonconforming finite elements for the velocity and the pressure, respectively piecewise constant elements for the stress tensor. Appropriate upwind schemes are employed for the convective terms, and the nonlinear discrete problem is solved by Newton's method. We next investigate the positive definiteness of the discrete conformation tensor and show that under certain hypotheses, this property is preserved by Newton's method. This allows us to attain the convergence of the algorithm for rather large Weissenberg numbers. Numerical tests validating the code are presented.

1 Introduction

We are interested in the numerical simulation of polymeric liquids which are, from a rheological point of view, viscoelastic non-Newtonian fluids.

The rheological behavior of polymers is so complex that many different constitutive equations have been proposed in the literature in order to describe it, see for instance [12]. We choose here to study the Giesekus model (cf. [4]) which presents two main advantages. First, it yields a realistic behavior for shear flows, elongational flows and mixed flows. Second, only two material parameters are needed to describe the model: the viscosity η and the relaxation time λ. However, the Giesekus constitutive law is strongly nonlinear since it involves, besides the objective derivative, a quadratic term in the stress tensor.

R. Becker (✉) and D. Capatina
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

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_13,

Despite numerous efforts, the numerical approximation of polymer flows is still a challenging research area, due to the internal coupling between the viscoelasticity of the liquid and the flow, which is quantified by the adimensional Weissenberg number $\mathrm{We} = \lambda\dot{\gamma}$ with $\dot{\gamma}$ the shear rate. A major issue to be addressed is the breakdown in convergence of the algorithms at critical values of *We*. The existing commercial codes are generally only able to deal with *We* up to 10, which is insufficient to describe polymer flows in a processing machine.

Our aim is to obtain realistic simulations of polymer flows (for $We > 10$), by using stabilized finite element methods and, in the future, mesh adaptivity. An outline of the paper is as follows. We describe the physical model in Sect. 2 and its discretization in Sect. 3. In Sect. 4, we investigate the positive definiteness of the discrete conformation tensor, which is a crucial point in computational rheology. It is generally believed that the loss of this property is responsible for the high Weissenberg number problem. We show, under certain hypotheses, that Newton's method yields a symmetric positive definite conformation tensor. Finally, we present some numerical tests illustrating the theoretical results.

2 The Giesekus Model

Giesekus introduced in [4] the following constitutive law

$$\lambda \overset{\nabla}{\underline{\tau}} + \frac{1}{2G}\underline{\tau}^2 + \underline{\tau} = 2\eta \underline{D}(\mathbf{u}) \tag{1}$$

with $\underline{\tau}$ the viscous stress tensor, $\underline{D}(\mathbf{u}) = \frac{1}{2}(\nabla\mathbf{u}^T + \nabla\mathbf{u})$ the shear rate tensor and G the elastic modulus given by the formula $\lambda = \eta/G$. Here above, $\overset{\nabla}{\underline{\tau}}$ is the upper convective derivative:

$$\overset{\nabla}{\underline{\tau}} = \partial_t \underline{\tau} + (\mathbf{u}\cdot\nabla)\underline{\tau} - (\underline{\tau}\nabla\mathbf{u}^T + \nabla\mathbf{u}\underline{\tau}).$$

The complete model is obtained by adding the momentum and mass conservation equations, where the density ρ is supposed to be constant:

$$\rho\partial_t\mathbf{u} + \rho(\mathbf{u}\cdot\nabla)\mathbf{u} - \mathrm{div}\underline{\tau} + \nabla p = \mathbf{f},$$
$$\mathrm{div}\,\mathbf{u} = 0$$

as well as initial conditions $\mathbf{u} = \mathbf{u}_0$, $\underline{\tau} = \underline{\tau}_0$ and boundary conditions $\mathbf{u} = \mathbf{g}$ on $\partial\Omega$, $\underline{\tau} = \underline{\tau}^{in}$ on the inflow boundary $\partial\Omega^- = \{x \in \partial\Omega; \mathbf{u}(x)\cdot\mathbf{n}(x) < 0\}$, where Ω is a polygonal domain of $\mathbf{R}^2$. In what follows, we consider the stationary case and we denote by $\mathbf{u_g}$ a lifting of the Dirichlet boundary condition on the velocity. The unknowns of the problem are the pressure p, the velocity $\mathbf{u}$ and the symmetric

viscous stress tensor $\underline{\tau}$. The nonlinear weak formulation can be formally written as follows:

$$\begin{cases} (\mathbf{u}, p, \underline{\tau}) \in (\mathbf{u_g} + \mathbf{H}_0^1(\Omega)) \times L_0^2(\Omega) \times \underline{X} & \\ a(\mathbf{u}, \mathbf{u}; \mathbf{v}) + b(p, \mathbf{v}) + c_0(\underline{\tau}, \mathbf{v}) = l(\mathbf{v}) & \forall \mathbf{v} \in \mathbf{H}_0^1(\Omega) \\ b(q, \mathbf{u}) = 0 & \forall q \in L_0^2(\Omega) \\ c(\mathbf{u}, \underline{\tau}; \underline{\sigma}) + d(\underline{\tau}, \underline{\tau}; \underline{\sigma}) = 0 & \forall \underline{\sigma} \in \underline{X} \end{cases} \tag{2}$$

where $\underline{X} = \underline{L}^2_{sym}(\Omega) \cap \underline{H}^1(\Omega)$. The forms are given by

$$a(\mathbf{u}, \mathbf{u}; \mathbf{v}) = \int_\Omega \rho \mathbf{u} \nabla \mathbf{u} \cdot \mathbf{v} dx, \;\; b(q, \mathbf{v}) = -\int_\Omega q div \mathbf{v} dx, \;\; c_0(\underline{\tau}, \mathbf{v}) = \int_\Omega \underline{\tau} : \underline{D}(\mathbf{v}) dx,$$
$$c(\mathbf{u}, \underline{\tau}; \underline{\sigma}) = 2\eta c_0(\underline{\sigma}, \mathbf{u}) + c_1(\mathbf{u}, \underline{\tau}; \underline{\sigma}) + c_2(\mathbf{u}, \underline{\tau}; \underline{\sigma}),$$
$$d(\underline{\tau}, \underline{\tau}; \underline{\sigma}) = d_0(\underline{\tau}, \underline{\sigma}) + d_1(\underline{\tau}, \underline{\tau}; \underline{\sigma})$$

where

$$c_1(\mathbf{u}, \underline{\tau}; \underline{\sigma}) = \lambda \int_\Omega \mathbf{u} \cdot \nabla \underline{\tau} : \underline{\sigma} dx, \qquad c_2(\mathbf{u}, \underline{\tau}; \underline{\sigma}) = -\lambda \int_\Omega (\underline{\tau} \nabla \mathbf{u}^T + \nabla \mathbf{u} \underline{\tau}) : \underline{\sigma} dx,$$
$$d_0(\underline{\tau}, \underline{\sigma}) = \int_\Omega \underline{\tau} : \underline{\sigma} dx \;, \qquad d_1(\underline{\tau}, \underline{\tau}; \underline{\sigma}) = \tfrac{1}{2G} \int_\Omega \underline{\tau}^2 : \underline{\sigma} dx.$$

All the previous forms are well-defined, thanks to the Sobolev embedding theorem which states that $H^1(\Omega) \subset L^4(\Omega)$.

3 Discretization of the Three-Fields Formulation

Let $(\mathscr{T}_h)_{h>0}$ be a regular family of triangulations of Ω consisting of triangles. We agree to denote by ε_h, respectively ε_h^{int} the set of edges, respectively internal edges of $\mathscr{T}_h$. On every internal edge e such that $\{e\} = \partial T_1 \cap \partial T_2$, we define once for all the unit normal $\mathbf{n}$; for a given function φ with $\varphi_{/T_i} \in \mathscr{C}(T_i)$ $(1 \leq i \leq 2)$, we define on e: $\varphi^{ext}(\mathbf{x}) = lim_{\varepsilon \to 0} \varphi(\mathbf{x} - \varepsilon \mathbf{n})$, $\varphi^{in}(\mathbf{x}) = lim_{\varepsilon \to 0} \varphi(\mathbf{x} + \varepsilon \mathbf{n})$ as well as the jump $[\varphi] = \varphi^{ext} - \varphi^{in}$. If the edge belongs to $\partial\Omega$, then $\mathbf{n}$ is the outward normal and the jump coincides with the trace.

Let us introduce the nonconforming, respectively piecewise constant spaces:

$$\mathbf{V}_h = \left\{ \mathbf{v} \in \mathbf{L}^2(\Omega);\; \mathbf{v}_{/K} \in \mathbf{P}_1,\; \forall K \in \mathscr{T}_h \text{ and } \int_e [\mathbf{v}] ds = 0,\; \forall e \in \varepsilon_h \right\}$$
$$Q_h = \left\{ q \in L_0^2(\Omega);\; q_{/K} \in P_0,\; \forall K \in \mathscr{T}_h \right\}$$
$$\underline{X}_h = \left\{ \underline{\sigma} = (\sigma_{ij})_{1 \leq i,j \leq 2} \in \underline{L}_2(\Omega);\; \underline{\sigma} = \underline{\sigma}^T \text{ and } \underline{\sigma}_{/K} \in \underline{P}_0,\; \forall K \in \mathscr{T}_h \right\}.$$

The discrete formulation can be written in the following form:

$$\begin{cases} (\mathbf{u}_h, p_h, \underline{\tau}_h) \in \mathbf{V}_h \times Q_h \times \underline{X}_h & & \\ (a_h + \gamma J)(\mathbf{u}_h, \mathbf{v}_h) + b(p_h, \mathbf{v}_h) + c_0(\underline{\tau}_h, \mathbf{v}_h) = l(\mathbf{v}_h) & \forall \mathbf{v}_h \in \mathbf{V}_h \\ b(q_h, \mathbf{u}_h) = 0 & \forall q_h \in Q_h \\ c_h(\mathbf{u}_h, \underline{\tau}_h; \underline{\sigma}_h) + d(\underline{\tau}_h, \underline{\tau}_h; \underline{\sigma}_h) = 0 & \forall \underline{\sigma}_h \in \underline{X}_h \end{cases} . \tag{3}$$

The convection term $\mathbf{u}\cdot\nabla\underline{\tau}$ in the constitutive law is treated by an upwind scheme, similarly to Lesaint–Raviart [9]. More precisely, $c_1(\cdot,\cdot;\cdot)$ is approximated by

$$c_{1h}(\mathbf{u}_h, \underline{\tau}_h; \underline{\sigma}_h) = \lambda \sum_{e \in \varepsilon_h} \int_e F(\underline{\tau}_h, \mathbf{u}_h, \mathbf{n}) : [\underline{\sigma}_h] ds$$

where $F(\underline{\tau}_h, \mathbf{u}_h, \mathbf{n}) = (\pi_0 \mathbf{u}_h \cdot \mathbf{n})^+ \underline{\tau}_h^{ext} + (\pi_0 \mathbf{u}_h \cdot \mathbf{n})^- \underline{\tau}_h^{in}$ and $\pi_0 \mathbf{v} = \dfrac{1}{|e|} \displaystyle\int_e \mathbf{v} ds$, $\forall e \in \varepsilon_h$.

As regards the momentum conservation law, the nonlinear form $a(\cdot,\cdot;\cdot)$ is replaced by $a_h(\cdot,\cdot;\cdot) + \gamma J(\cdot,\cdot)$. The first term takes into account the stabilization of the convective term $\mathbf{u} \cdot \nabla \mathbf{u}$, $\gamma > 0$ is a stabilization parameter whereas the term $J(\cdot,\cdot)$ is added in order to retrieve a Korn inequality on $\mathbf{V}_h$ and is defined (cf. [2], [10]) by

$$J(\mathbf{u}_h, \mathbf{v}_h) = \sum_{e \in \varepsilon_h^{\text{int}}} \frac{\eta}{|e|} \int_e [\mathbf{u}_h \cdot \mathbf{n}][\mathbf{v}_h \cdot \mathbf{n}] ds.$$

The nonlinear problem (3) is solved by Newton's method; this implies the computation of the following Jacobian matrix:

$$\begin{pmatrix} \partial_u a_h + \gamma J & b & c_0 \\ b^T & 0 & 0 \\ c_0^T + \partial_u(c_{1h} + c_2) & 0 & d_0 + \partial_\tau(c_{1h} + c_2 + d_1) \end{pmatrix}.$$

The stability of this mixed matrix holds under usual inf-sup conditions. The Newtonian case follows from [1]; the analysis of the general case is in progress, based on the results of Sect. 4 (see also Remark 1). Note that $\partial_u C_2$, $\partial_\tau C_2$ and $\partial_\tau D_1$ are defined locally on each triangle, whereas the stencils of $\partial_u C_1$ and $\partial_\tau C_1$ are reduced to the element itself and its neighbours. The development of a specially designed Newton algorithm is undergoing work.

4 Positive Definiteness of the Conformation Tensor

We define the conformation tensor by the relation

$$\underline{C} = \frac{\lambda}{\eta} \underline{\tau} + \underline{I}. \tag{4}$$

Note that in [4], Giesekus introduced his model by means of a configuration tensor which actually coincides with $\underline{C}$. By means of the relations $\overset{\nabla}{\underline{I}} = -2\underline{D}(\mathbf{u})$ and $\lambda G = \eta$, the constitutive law (1) can be equivalently written as follows:

$$\lambda(\mathbf{u}\cdot\nabla\underline{C} - \underline{C}\nabla\mathbf{u}^T - \nabla\mathbf{u}\underline{C}) + \frac{1}{2}\underline{C}\cdot\underline{C} = \frac{1}{2}\underline{I}.$$

Theorem 1. *If $\underline{C}^{in}$ is symmetric positive definite (s.p.d.) on $\partial\Omega^-$ and $\underline{C}(\cdot)$ is continuous on Ω, then $\underline{C}$ is s.p.d.*

The proof follows [8], where instationary constitutive laws of the form $\overset{\nabla}{\underline{C}} + \alpha\underline{C} = \beta\underline{I}$ with $\beta > 0$ are considered. The authors' argument is based on the closed-form solution of differential Riccati equations. Another proof was given by Hulsen in [5] for more general laws, but only in the case where $\overset{\nabla}{\underline{C}} = \partial_t\underline{C} - \underline{C}\nabla\mathbf{u}^T - \nabla\mathbf{u}\underline{C}$.

The numerical schemes which preserve the positive definiteness of $\underline{C}$ seem to be more stable and usually, energy estimates can also be derived. In order to obtain such schemes, several approaches have been developed in the last years. Fattal and Kupferman proposed in [3] to write the constitutive law in terms of $\underline{\psi} = ln\underline{C}$ by using a specific decomposition of $\nabla\mathbf{u}$, to approximate $\underline{\psi}$ and to put $\underline{C}_h = e^{\underline{\psi}_h}$.

Lee and Xu employed in [8] the framework of Riccati equations. We adopt here this approach, but we employ a discontinuous Galerkin method instead of the characteristics one, and also a different constitutive equation. However, we need to assume that $\Omega = \cup_{i=1}^{N} K_i$ such that

$$\forall i, \quad \partial K_i^- \subset \partial\Omega^- \quad \text{or} \quad \partial K_i^- \subset \cup_{j<i}\partial K_j^+. \tag{5}$$

Note that (5) holds true if $\mathbf{u}_h$ is constant, cf. [9]. According to (3), the discrete conformation tensor $\underline{C}_h$ satisfies on any $K \in \mathscr{T}_h$ an algebraic Riccati equation:

$$\underline{A}\underline{C}_h + \underline{C}_h\underline{A}^T - \underline{C}_h\underline{B}\underline{C}_h + \underline{F} = \underline{0} \tag{6}$$

with $\underline{B} = \frac{1}{2\lambda}\underline{I}$, $\underline{F} = \frac{1}{2\lambda}\underline{I} + \underline{C}_h^*$ and $\underline{A} = \nabla\mathbf{u}_h - \frac{d}{2}\underline{I}$. The upwinding scheme implies

$$\underline{C}_h^* = \frac{1}{|K|}\sum_{e\in\partial K^-}\int_e |\mathbf{u}_h\cdot\mathbf{n}|\underline{C}_h^{ext}ds\ , \qquad d = \frac{1}{|K|}\int_{\partial K^+}\mathbf{u}_h\cdot\mathbf{n}ds > 0.$$

Note that $\underline{B}$ is always s.p.d. For the sake of simplicity, we consider in what follows that $\mathbf{u}_h$ is known. Obviously, the quadratic equation (6) does not have a unique solution. Nevertheless, we can prove that Newton's method converges towards the maximal s.p.d. solution of (6). For this purpose, let us notice that Newton's iterate

$\underline{C}_h^n$ satisfies on any K an algebraic Lyapunov equation:

$$(\underline{A} - \underline{C}_h^{n-1}\underline{B})\underline{C}_h^n + \underline{C}_h^n(\underline{A} - \underline{C}_h^{n-1}\underline{B})^T = -\underline{F} - \underline{C}_h^{n-1}\underline{B}\underline{C}_h^{n-1}$$

and recall (cf. [7]) the next result for the Lyapunov equation $\underline{\mathscr{A}}\,\underline{\mathscr{X}} + \underline{\mathscr{X}}\,\underline{\mathscr{A}}^T = -\underline{\mathscr{F}}$.

Lemma 1. *If $\underline{\mathscr{A}}$ is stable (i.e., $Re(\lambda) < 0$) and $\underline{\mathscr{F}}$ is s.p.d., then $\underline{\mathscr{X}}$ is s.p.d.*

Then one can prove the following statement on any K (see for instance [7], [11]):

Theorem 2. *If $(\underline{A} - \underline{C}_h^0\underline{B})$ is stable and if $\underline{F}$ and $\underline{C}_h^0$ are s.p.d., then $(\underline{A} - \underline{C}_h^n\underline{B})$ is stable and $\underline{C}_h^n$ is s.p.d. for all n. Moreover, $(\underline{C}_h^n)_n$ converges towards the maximal s.p.d. solution of (6).*

The previous result together with an induction argument on the triangles yield:

Theorem 3. *Assume (5) and $\underline{C}_h^{in}$ s.p.d. If $\underline{C}_h^0$ satisfies on any K the assumptions of Theorem 2, then Newton's iterates are s.p.d. and converge towards the maximal s.p.d. solution of (6) on Ω.*

Remark 1. One may note that $(d_0 + \partial_\tau(c_{1h} + c_2 + d_1))\,(\underline{\sigma}_h, \underline{\sigma}_h) = -2\lambda tr(\underline{\mathscr{A}}\,\underline{\sigma}_h^2)$ for any $\underline{\sigma}_h \in \underline{X}_h$, with $\underline{\mathscr{A}} = \underline{A} - \underline{C}_h^n\underline{B}$ stable by hypothesis. This yields the ellipticity of the bilinear form $(d_0 + \partial_\tau(c_{1h} + c_2 + d_1))\,(\cdot,\cdot)$.

Remark 2. Let us now look at the same DG scheme for the Oldroyd-B model:

$$\mathbf{u}\cdot\nabla\underline{C} - \nabla\mathbf{u}\underline{C} - \underline{C}\nabla\mathbf{u}^T + \frac{1}{We}\underline{C} = \frac{1}{We}\underline{I}\,.$$

One can see that $\underline{C}_h$ satisfies on any triangle a Lyapunov equation of matrix $\underline{\mathscr{A}} = \underline{A} - \frac{1}{2We}\underline{I}$, which according to Lemma 1 has to be stable in order to get a s.p.d. solution. Clearly, it may occur that this condition is violated for large Weissenberg numbers. The instationary case is generally easier to treat, since $\underline{\mathscr{A}}$ is now replaced by $\underline{\mathscr{A}} - \frac{1}{2\Delta t}\underline{I}$, which can be rendered stable for Δt small enough.

5 Numerical Results

The code is written in the in-house C++ library Concha. We first consider the 4:1:4 planar contraction/expansion, cf. Fig. 1 (left), with $a = 10^{-3}\,\mathrm{m}$, $\eta = 1000\,\mathrm{Pa\,s}$ and $\rho = 1000\,\mathrm{kg\,m^{-3}}$. On the inflow (left), we set $\mathbf{u}\cdot\mathbf{n} = 0.1\,\mathrm{m\,s^{-1}}$ whereas a Neumann condition is imposed on the outflow boundary (right); a symmetry condition is imposed on the top boundary and a null velocity elsewhere. We compute the largest Weissenberg number for a corresponding Newtonian fluid as follows: $We = \lambda\dot{\gamma} = \dfrac{6\lambda u}{a}$, with u the mean velocity in the thin channel. We first compare in

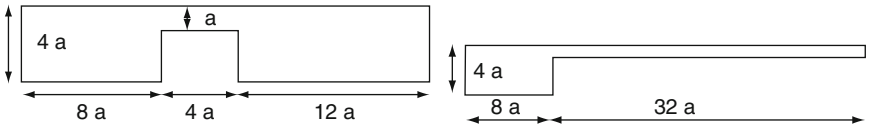


Fig. 1 Geometry of the 4:1:4 (*left*) and the 4:1 (*right*) test-cases

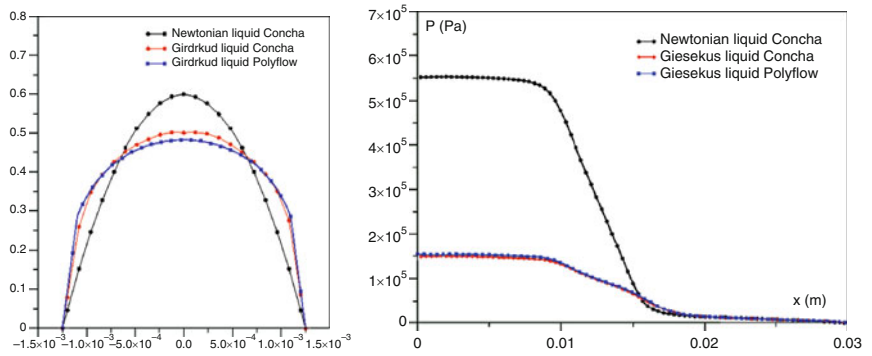


Fig. 2 Comparison Concha vs. Polyflow at $We = 7.68$: velocity (*left*) and pressure (*right*)

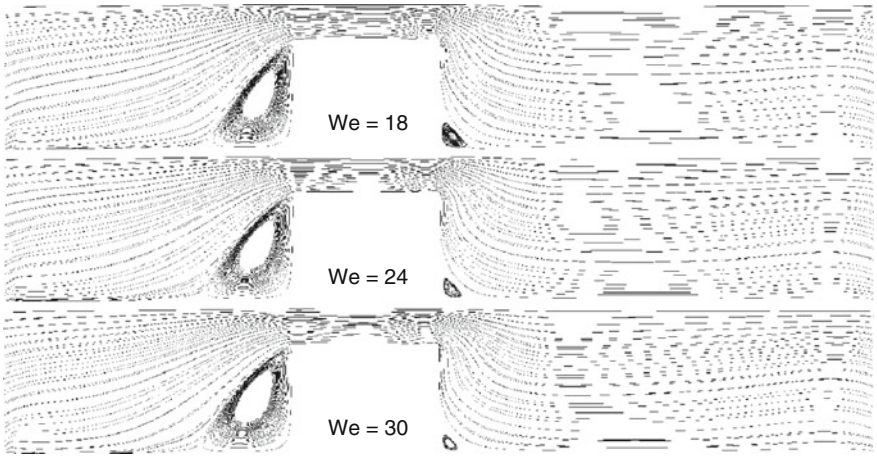


Fig. 3 Comparison of velocity magnitude for Giesekus ($We = 2.6$) and Newtonian flows

Fig. 2 our results with those given by Polyflow, which is the most popular code for the simulation of polymer liquids (http://www.ansys.com/products/polyflow).

We have used a mesh consisting of 25,794 triangles with Concha, respectively 14,866 with Polyflow. The highest Weissenberg number for which Polyflow still converges is $We = 7.68$. We observe a good agreement between the two approaches; the modification of the velocity profiles on the vertical axis in the thin channel and the shut down of the pressure on the symmetry axis are typical

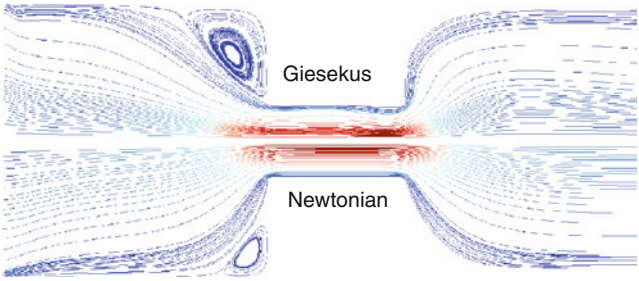


Fig. 4 Giesekus flow for large Weissenberg numbers ($We = 18$, 24 and 30)

Fig. 5 First eigenvalue of the conformation tensor for $We = 18$, 30 and 33.5

Fig. 6 Second eigenvalue of the conformation tensor for $We = 18$, 30 and 33.5

behaviors of non-Newtonian fluids. Next, we compare in Fig. 3 the velocity magnitude for a Newtonian and a Giesekus fluid while in Fig. 4 we show several simulations of the Giesekus flow for different Weissenberg numbers.

Finally, we consider the 4:1 planar contraction cf. Fig. 1 (right). We show in Fig. 5 and Fig. 6 the two eigenvalues of the conformation tensor, computed for different Weissenberg numbers. As expected, they are both strictly positive. For $We > 33.5$, the algorithm does not converge. This can be explained by the fact that the conformation tensor is no longer positive definite; we recall that the coupled system (3) is solved globally, so the velocity is computed at every Newton's iteration.

References

1. Becker, R., Capatina, D., Joie, J.: A dG method for the Stokes equations related to nonconforming approximations. INRIA Research Report (2009), http://hal.inria.fr/inria-00380772
2. Brenner, S.: Korn's inequalities for piecewise H_1 vector fields. Math. Comp. **73**, 1067–1087 (2004)
3. Fattal, R., Kupferman, R.: Constitutive laws of the matrix-logarithm of the conformation tensor. J. Non-Newtonian Fluid Mech. **123**, 281–285 (2004)
4. Giesekus, H.: A simple constitutive equation for polymer fluids based on the concept of deformation-dependent tensorial mobility. J. Non-Newtonian Fluid Mech. **11**, 69–109 (1982)

5. Hulsen, M.: A sufficient condition for a positive definite configuration tensor in differential models. J. Non-Newtonian Fluid Mech. **38**, 93–100 (1990)
6. Hulsen, M., Fattal, R., Kupferman, R.: Flow of viscoelastic fluids past a cylinder at high Weissenberg number: Stabilized simulations using matrix logarithms. J. Non-Newtonian Fluid Mech. **127**, 27–39 (2005)
7. Lancaster, P., Rodman, L.: Algebraic Riccati Equations. Clarendon Press, Oxford (1995)
8. Lee, Y., Xu, J.: New formulations, positivity preserving discretizations and stability analysis for non-Newtonian flow models. Comput. Meth. Appl. Mech. Eng. **195**, 1180– (2006)
9. Lesaint, P., Raviart, P.-A.: On a finite element method for solving the neutron transport equation. In: de Boor, C.A. (ed.) Mathematical Aspects of Finite Element Methods in Partial Differential Equations, pp. 89–123. Academic Press, NY (1974)
10. Mardal, K.-A., Winther, R.: An observation on Korn's inequality for nonconforming finite element methods. Math. Comp. **75**, 1–6 (2005)
11. Mehrmann, V.L.: The autonomous linear quadratic control problem. Lecture Notes in Control and Information Sciences, **163**, Springer, Berlin (1991)
12. Owens, R.G., Phillips, T. N.: Computational Rheology. Imperial College Press, London (2002)

A dG Method for the Strain-Rate Formulation of the Stokes Problem Related with Nonconforming Finite Element Methods

Roland Becker, Daniela Capatina, and Julie Joie

Abstract We study a discontinuous Galerkin method for the Stokes problem written in terms of the strain-rate tensor. We approach the velocity by polynomials of degree k and the pressure by polynomials of degree $k-1$ by element for $k = 1, 2$ or 3. The stabilization of the viscous term is new and involves the jump across the edges of the L^2-projection on P_{k-1} of the velocity. It allows us to recover, when the stabilization parameter γ tends towards infinity, some stable and well-known nonconforming approximations; moreover, the inf-sup constant is independent of γ. This allows us to conclude that our method is robust with respect to γ. For $k = 1$, a second stabilization term is added in order to retrieve a discrete Korn inequality. The choice of the strain-rate formulation presents two main advantages, steming from its equivalence with a three-fields formulation of the Stokes problem. First, it can be easily extended to non-Newtonian liquids. Second, it allows us to deal with more physical boundary conditions involving the normal stress. Optimal a priori error estimates are also derived and numerical tests illustrating the accuracy and the robustness of the scheme are presented.

1 Introduction

In the literature, there exist many finite element methods for the approximation of the Stokes problem (see for instance [4] for a presentation of continuous and nonconforming schemes). We are interested here in discontinuous Galerkin methods. One of the most used and well-known dG approximation is the symmetric interior penalty method, first introduced in [1] for the Laplace equation, and then generalized to Stokes and Navier–Stokes equations by Girault, Rivière and Wheeler [5]. The velocity is looked for in $\mathbf{P}_k$ and the pressure belongs to P_{k-1}, with $1 \leq k \leq 3$.

R. Becker (✉), D. Capatina and J. Joie
EPI CONCHA & LMA CNRS UMR 5142, INRIA Bordeaux-Sud-Ouest & Université de Pau, IPRA, BP1155, 64013 PAU, France
e-mail: roland.becker@univ-pau.fr, daniela.capatina@univ-pau.fr, julie.joie@univ-pau.fr

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_14,

The stabilization term is a penalization of the jumps of the velocities across the edges. Instead, we propose to penalize the L^2-projection on P_{k-1} of the jumps. This new stabilization allows us to prove that contrarily to [5], our method is robust for large stabilization parameters γ. Indeed, the solution of the dG formulation tends, as γ goes to infinity, towards the solution of the $\mathbf{P}_k \times P_{k-1}$ nonconforming approximation of the Stokes problem. Moreover, the inf-sup constant with respect to the energy norm of our method is independent of γ whereas that of [5] is $O(1/\sqrt{\gamma})$.

We study here the Stokes problem written in terms of the strain-rate tensor. The main advantage is the equivalence between its dG version and a three-fields formulation, allowing to recover the stress tensor in an obvious way. One may then be able to generalize it to non-Newtonian fluids or to impose other boundary conditions related to the normal stress. In order to retrieve a discrete Korn inequality for discontinuous velocities (cf. [3] or [6]), we consider an additional stabilization term. The proposed discretization is well-posed and yields optimal convergence rates, which are confirmed by numerical experiments.

We write the vectors in bold letters and the second-order tensors in underlined letters, $\underline{\tau} = (\tau_{ij})_{1\leq i,j\leq 2}$; the product of two tensors will be denoted by $\underline{\tau} : \underline{\sigma} = \sum_{i,j} \tau_{ij}\sigma_{ij}$.

2 The Stokes Problem

We consider the Stokes equations, which describe the steady flow of an incompressible, Newtonian fluid at low Reynolds numbers and which can be written as follows:

$$-2\mu\, div\underline{D}(\mathbf{u}) + \nabla p = \mathbf{f} \ \text{ in } \Omega \tag{1}$$

$$\nabla \cdot \mathbf{u} = 0 \ \text{ in } \Omega \tag{2}$$

where $\mathbf{u}$ is the velocity, p the pressure, $\underline{D}(\mathbf{u}) = \frac{1}{2}\left(\underline{\nabla}\mathbf{u} + (\underline{\nabla}\mathbf{u})^t\right)$ the strain-rate tensor and μ the viscosity. We take Ω a polygonal domain of $\mathbf{R}^2$ and the data $\mathbf{f} \in \mathbf{L}^2(\Omega)$.

For the sake of simplicity, we only consider here homogeneous Dirichlet boundary conditions. The corresponding variational formulation can be written as follows:

$$\begin{cases} (\mathbf{u}, p) \in \mathbf{H}_0^1(\Omega) \times L_0^2(\Omega) \\ 2\mu \displaystyle\int_\Omega \underline{D}(\mathbf{u}) : \underline{D}(\mathbf{v})dx - \int_\Omega p\nabla\cdot\mathbf{v}dx = \int_\Omega \mathbf{f}\cdot\mathbf{v}dx & \forall \mathbf{v} \in \mathbf{H}_0^1(\Omega) \\ -\displaystyle\int_\Omega q\nabla\cdot\mathbf{u}dx = 0 & \forall q \in L_0^2(\Omega) \end{cases} \tag{3}$$

Its well-posedness results from the Babŭska–Brezzi theorem and a Korn inequality.

3 Discontinuous Galerkin Discretization

We are now interested in the discretization of problem (3) by means of fully discontinuous finite elements. Let $(\mathscr{T}_h)_{h>0}$ be a family of triangulations of Ω consisting of triangles: $\overline{\Omega} = \bigcup_{T\in\mathscr{T}_h} T$. Let us first introduce some useful notations.

We denote by ε_h^{int} the set of internal edges of $\mathscr{T}_h$, by ε_h^{∂} the set of edges situated on the boundary $\partial\Omega$ and we put $\varepsilon_h = \varepsilon_h^{int} \cup \varepsilon_h^{\partial}$. As usually, let h_T be the diameter of the triangle T and let $h = \max_{T\in\mathscr{T}_h} h_T$. On every edge e belonging to ε_h^{int}, such that $\{e\} = \partial T^i \cap \partial T^j$, we define once for all the unit normal $\mathbf{n}_e$ oriented from T^i towards T^j. Then, for a given function φ, we define the jump across the edge e by $[\varphi] = \varphi_{/T^i} - \varphi_{/T^j}$ and the average by $\{\varphi\} = \frac{1}{2}\left(\varphi_{/T^i} + \varphi_{/T^j}\right)$. If $e \in \varepsilon_h^{\partial}$, we take for $\mathbf{n}_e$ the outward unit normal $\mathbf{n}$ and for $[\varphi]$ and $\{\varphi\}$ the trace of φ on e. We agree to denote the $L^2(e)$-orthogonal projection of a given function $\varphi \in L^2(e)$ on the polynomial space P_k by $\pi_k\varphi$. In what follows, we take $k = 1$, 2 or 3 and we introduce the finite dimensional spaces:

$$\mathbf{V}_h = \left\{\mathbf{v}_h \in \mathbf{L}^2(\Omega);\ (\mathbf{v}_h)_{/T} \in \mathbf{P}_k,\ \ \forall T \in \mathscr{T}_h\right\},$$
$$Q_h = \left\{q_h \in L_0^2(\Omega);\ (q_h)_{/T} \in P_{k-1},\ \ \forall T \in \mathscr{T}_h\right\}.$$

We consider the following discontinuous Galerkin approximation of (3):

$$\begin{cases} (\mathbf{u}_h, p_h) \in \mathbf{V}_h \times Q_h & \\ a_h(\mathbf{u}_h, \mathbf{v}_h) + b_h(p_h, \mathbf{v}_h) = \int_\Omega \mathbf{f}\cdot\mathbf{v}dx & \forall \mathbf{v}_h \in \mathbf{V}_h \\ b_h(q_h, \mathbf{u}_h) = 0 & \forall q_h \in Q_h \end{cases} \tag{4}$$

where : $a_h(\cdot,\cdot) = A_0(\cdot,\cdot) + A_1(\cdot,\cdot) + \gamma J(\cdot,\cdot) + \gamma_1 J_1(\cdot,\cdot)$

$$A_0(\mathbf{u}_h, \mathbf{v}_h) = 2\mu \sum_{T\in\mathscr{T}_h} \int_T \underline{D}(\mathbf{u}_h) : \underline{D}(\mathbf{v}_h)dx$$

$$A_1(\mathbf{u}_h, \mathbf{v}_h) = -2\mu \sum_{e\in\varepsilon_h} \left(\int_e \{\underline{D}(\mathbf{u}_h)\mathbf{n}_e\}\cdot[\mathbf{v}_h]\,ds + \int_e \{\underline{D}(\mathbf{v}_h)\mathbf{n}_e\}\cdot[\mathbf{u}_h]\,ds \right)$$

$$J(\mathbf{u}_h, \mathbf{v}_h) = \mu \sum_{e\in\varepsilon_h} \frac{1}{|e|} \int_e [\boldsymbol{\pi}_{k-1}\mathbf{u}_h]\cdot[\boldsymbol{\pi}_{k-1}\mathbf{v}_h]ds$$

$$J_1(\mathbf{u}_h, \mathbf{v}_h) = \mu \sum_{e\in\varepsilon_h^{int}} \frac{1}{|e|} \int_e [\boldsymbol{\pi}_1(\mathbf{u}_h\cdot\mathbf{n}_e)][\boldsymbol{\pi}_1(\mathbf{v}_h\cdot\mathbf{n}_e)]ds$$

$$b_h(q_h, \mathbf{v}_h) = -\sum_{T\in\mathscr{T}_h} \int_T q_h \nabla\cdot\mathbf{v}_h dx + \sum_{e\in\varepsilon_h} \int_e \{q_h\}[\mathbf{v}_h\cdot\mathbf{n}_e]ds$$

and where $\gamma > 0$ and $\gamma_1 > 0$ are two stabilization parameters independent of h. The stabilization terms $J(\cdot,\cdot)$ and $J_1(\cdot,\cdot)$ are added in order to retrieve the coercivity of $a_h(\cdot,\cdot)$. $J(\cdot,\cdot)$ is inherent to our dG method whereas $J_1(\cdot,\cdot)$ is needed to obtain a discrete Korn inequality on $\mathbf{V}_h$. Note that in the classical Interior Penalty (IP) method, $J(\cdot,\cdot)$ is replaced by: $J^{IP}(\mathbf{u},\mathbf{v}) = \mu \sum_{e\in\varepsilon_h} \frac{1}{|e|}\int_e [\mathbf{u}]\cdot[\mathbf{v}]ds$.

Remark 1. For $k = 2$ or 3, one has $KerJ \subset KerJ_1$ therefore it is not necessary to add $J_1(\cdot,\cdot)$. We only add it in order to get a unified presentation of the method for all k.

In view of the analysis of (4), we introduce the semi-norms on $\mathbf{H}^1(\Omega) + \mathbf{V}_h$:

$$|\mathbf{v}|_{1,h} = \left(\sum_{T\in\mathscr{T}_h} |\mathbf{v}|^2_{1,T}\right)^{1/2}, \qquad \|\underline{D}(\mathbf{v})\|_{0,h} = \left(\sum_{T\in\mathscr{T}_h} \|\underline{D}(\mathbf{v})\|^2_{0,T}\right)^{1/2},$$
$$[[\mathbf{v}]] = \left(2\mu\|\underline{D}(\mathbf{v})\|^2_{0,h} + \gamma J(\mathbf{v},\mathbf{v}) + \gamma_1 J_1(\mathbf{v},\mathbf{v})\right)^{1/2}.$$

Then we can show the next results (see [2] for the detailed proofs):

Lemma 1. *The application* $\mathbf{v} \to [[\mathbf{v}]]$ *is a norm on* $\mathbf{V}_h$*, for all* $k = 1,\ 2,\ 3$.

Lemma 2. *For any* $\mathbf{v} \in \mathbf{V}_h$*, there exists a constant* $c > 0$ *independent of* h *and* μ*, such that:*

$$|\mathbf{v}|_{1,h} \le c\left(\|\underline{D}(\mathbf{v})\|^2_{0,h} + \frac{1}{\mu}J_1(\mathbf{v},\mathbf{v}) + \frac{1}{\mu}J(\mathbf{v},\mathbf{v})\right)^{1/2}. \tag{5}$$

Proof. We use the following result for piecewise H^1 functions, established by Brenner [3] in a stronger form and then improved in [6]:

$$|\mathbf{v}|_{1,h} \le c\left(\|\underline{D}(\mathbf{v})\|_{0,h} + (\frac{1}{\mu}J_1(\mathbf{v},\mathbf{v}))^{1/2} + \phi(\mathbf{v})\right),$$

where $\phi : \mathbf{H}^1(\Omega) \to \mathbf{R}$ is a continuous semi-norm such that if $\phi(\mathbf{v}) = 0$ for a rigid motion $\mathbf{v}$, then $\mathbf{v}$ is a constant vector. We apply it with $\phi(\mathbf{v}) = (\sum_{e\in\varepsilon_h^\partial} \|\boldsymbol{\pi}_0\mathbf{v}\|^2_{0,e})^{1/2}$.

4 Theoretical Results

In this section, we briefly present the results that we have established for the discrete formulation (4); for the proofs see [2]. We first prove the well-posedness of (4), using the Babŭska–Brezzi theorem; more precisely, we establish the $\mathbf{V}_h$-coercivity of $a_h(\cdot,\cdot)$ for γ large enough and the inf-sup condition on $b_h(\cdot,\cdot)$. It is important to note that the inf-sup constant is shown to be independent of γ and γ_1.

Theorem 1. *For γ sufficiently large, the mixed problem (4) has a unique solution.*

Next, we investigate the robustness of our dG method (4) for large stabilization parameters γ and the relation between (4) and the $\mathbf{P}_k^{nc} \times P_{k-1}^{disc}$ nonconforming finite element approximation of the Stokes problem, which reads as follows:

$$\begin{cases} (\mathbf{u}_h^*, p_h^*) \in \mathbf{H}_h \times Q_h \\ A_0(\mathbf{u}_h^*, \mathbf{v}_h) + \gamma_1 J_1(\mathbf{u}_h^*, \mathbf{v}_h) + b_h(p_h^*, \mathbf{v}_h) = f_h(\mathbf{v}_h) \quad \forall \mathbf{v}_h \in \mathbf{H}_h \\ b_h(q_h, \mathbf{u}_h^*) = 0 \quad \forall q_h \in Q_h \end{cases} \tag{6}$$

where $\mathbf{H}_h = \left\{ \mathbf{v} \in \mathbf{L}^2(\Omega);\ (\mathbf{v})_{/T} \in \mathbf{P}_k,\ \forall T \in \mathscr{T}_h \text{ and } [\pi_{k-1}\mathbf{v}]_{/e} = 0, \forall e \in \varepsilon_h \right\}$.

Remark 2. Note that $\mathbf{H}_h = KerJ$ and for $k = 2$ or 3, $J_1(\mathbf{u}_h, \mathbf{v}_h) = 0$ for all $\mathbf{u}_h,\ \mathbf{v}_h \in \mathbf{H}_h$.

Theorem 2. *Let $(\mathbf{u}_h, p_h)$ be the solution of (4) and $(\mathbf{u}_h^*, p_h^*)$ the solution of (6). Then, for γ_1 fixed, one has that:*

$$\lim_{\gamma \to \infty} \left(\left[\left[\mathbf{u}_h - \mathbf{u}_h^* \right] \right] + \| p_h - p_h^* \|_{0,\Omega} \right) = 0.$$

Remark 3. When using the classical IP stabilization term $J^{IP}(\cdot,\cdot)$ instead of $J(\cdot,\cdot)$, the solution $(\mathbf{u}_h^{IP}, p_h^{IP})$ tends as $\gamma \to \infty$ towards the solution of the Stokes problem discretized by $\mathbf{P}_k^{cont} \times P_k^{disc}$ finite elements. Contrarily to $\mathbf{P}_k^{nc} \times P_{k-1}^{disc}$, they don't form a stable pair of spaces for the Stokes problem.

Remark 4. For $k = 1$, if we let both γ and γ_1 tend towards infinity, the obtained velocity will belong to $X_h = KerJ \cap KerJ_1$. One can easily see that $\mathbf{P}_{1/\partial\Omega}^{cont} \subset X_h \subset \mathbf{P}_{1/\partial\Omega}^{nc}$ with $\mathbf{P}_{1/\partial\Omega}^{cont}$, respectively $\mathbf{P}_{1/\partial\Omega}^{nc}$, the space of continuous finite elements, respectively nonconforming finite elements with homogeneous Dirichlet condition on $\partial\Omega$.

As regards the a priori error estimates for problem (4), one gets:

Theorem 3. *Let $(\mathbf{u}, p) \in \mathbf{H}^{k+1}(\Omega) \times H^k(\Omega)$ be the solution of the continuous Stokes problem and let γ be sufficiently large (as in Theorem 1). Then:*

$$[[\mathbf{u} - \mathbf{u}_h]] \leq c h^k (\sqrt{\mu} |\mathbf{u}|_{k+1,\Omega} + \frac{1}{\sqrt{\mu}} |p|_{k,\Omega})$$
$$\| p - p_h \|_{0,\Omega} \leq c h^k (\mu |\mathbf{u}|_{k+1,\Omega} + |p|_{k,\Omega})$$

with a constant c independent of h and μ. Moreover, if Ω is convex then

$$\| \mathbf{u} - \mathbf{u}_h \|_{0,\Omega} \leq c h^{k+1} (|\mathbf{u}|_{k+1,\Omega} + \frac{1}{\mu} |p|_{k,\Omega}).$$

5 Three Fields Formulation of the Stokes Problem

Let $\underline{X} = \left\{\underline{\theta} = (\theta_{ij})_{1\le i,j\le 2}; \theta_{ij} = \theta_{ji},\ \theta_{ij} \in L_2(\Omega),\ i,j = 1,2\right\}$ and let $\underline{\tau} = 2\mu\underline{D}(\mathbf{u})$ be the viscous stress tensor. We consider the following dG discretization of the three-fields formulation of the Stokes problem:

$$\begin{cases} (\mathbf{U}_h, P_h, \underline{\tau}_h) \in \mathbf{V}_h \times Q_h \times \underline{X}_h & & \\ k_h(\mathbf{U}_h, \mathbf{v}_h) + b_h(P_h, \mathbf{v}_h) + d_h(\underline{\tau}_h, \mathbf{v}_h) = \displaystyle\int_\Omega \mathbf{f}\cdot\mathbf{v}dx & \forall \mathbf{v}_h \in \mathbf{V}_h \\ b_h(q_h, \mathbf{U}_h) = 0 & \forall q_h \in Q_h \\ d_h(\underline{\theta}_h, \mathbf{U}_h) - e_h(\underline{\theta}_h, \underline{\tau}_h) = 0 & \forall \underline{\theta}_h \in \underline{X}_h \end{cases} \tag{7}$$

where $\underline{X}_h = \left\{\underline{\theta}_h \in \underline{X};\ (\underline{\theta}_h)_{/T} \in \underline{P}_{k-1},\ \ \forall T \in \mathcal{T}_h\right\}$ and where:

$$\begin{aligned} k_h(\cdot,\cdot) &= A_1(\cdot,\cdot) + \gamma J(\cdot,\cdot) + \gamma_1 J_1(\cdot,\cdot) \\ d_h(\underline{\theta}_h, \mathbf{v}_h) &= \sum_{T\in\mathcal{T}_h} \int_T \underline{\theta}_h : \underline{D}(\mathbf{v}_h)dx \\ e_h(\underline{\theta}_h, \underline{\tau}_h) &= \frac{1}{2\mu}\sum_{T\in\mathcal{T}_h} \int_T \underline{\theta}_h : \underline{\tau}_h dx. \end{aligned}$$

One can easily show the equivalence between (4) and (7) in the sense that the solution of (7) is given by $\mathbf{U}_h = \mathbf{u}_h$, $P_h = p_h$ and $\underline{\tau}_h = 2\mu\underline{D}(\mathbf{u}_h)$.

This formulation can be extended to non-Newtonian fluids in a natural manner. Indeed, when considering such fluids one cannot eliminate $\underline{\tau}$ from the constitutive law, and hence one deals with formulations of at least three unknowns. Moreover, this formulation allows us to dispose of a large panel of boundary conditions; one can thus impose the normal and/or tangential forces (see [2]).

6 Numerical Results

We present here some numerical tests illustrating the previous results. The developed C++ codes are written in the in-house library CONCHA. We recall that the additionnal stabilization term $J_1(\cdot,\cdot)$ is necessary only in the case $k = 1$. For this reason, we have chosen to illustrate the previous theoretical results only for $k = 1$ and in what follows, we consider $\gamma_1 = 10$. Moreover, the constants being independent of the viscosity, we have arbitrarily chosen to take $\mu = 1$. Similar results are obtained for other values of the viscosity. We first study the behavior of the numerical scheme (4) with respect to mesh refinement. We consider the exact solution of Stokes problem with non-homogeneous Dirichlet conditions:

$$\mathbf{u}(x,y) = \begin{pmatrix} \pi\cos(\pi x)\sin(\pi y) \\ -\pi\sin(\pi x)\cos(\pi y) \end{pmatrix}, \quad p(x,y) = \sin(\pi x)\sin(\pi y) \tag{8}$$

on $\Omega = [-1, 1] \times [-1, 1]$. At each refinement step, the discretization parameter h is divided by 2. We represent in Fig. 1 the error curves in terms of the total number of elements ne, in log scale. For the velocity (in the energy norm and in the L^2-norm) and for the pressure, they are in agreement with the theoretical results, that is:

$$\|\mathbf{u} - \mathbf{u}_h\|_{0,\Omega} = O(h^2), \quad [[\mathbf{u} - \mathbf{u}_h]] = O(h), \quad \|p - p_h\|_{0,\Omega} = O(h).$$

Now, we let γ vary and compare the results given by our numerical method with those obtained with the classical IP stabilization. We are interested in the computed errors and solutions for large γ on a fixed mesh, with γ_1 fixed. Let the exact solution be given by (8). We employ a mesh consisting of 4,096 elements. We compare in Fig. 2 the velocity and pressure errors computed by the two methods (ours in continuous lines, the IP in dotted lines), for different values of γ. One can notice that our stabilization yields a stable scheme independently of γ whereas the IP method yields bigger errors, which increase with γ.

We now consider a Poiseuille flow in the domain $\Omega = [0; 0.06] \times [-0.01; 0.01]$. On the inflow, we set $\mathbf{u} \cdot \mathbf{t} = 0$ and $\mathbf{u} \cdot \mathbf{n}$, whereas on the outflow we impose a homogeneous Neumann condition: $2\mu \underline{D}(\mathbf{u})\mathbf{n} - p\mathbf{n} = \mathbf{0}$. We have first considered

ne	$\|\mathbf{u} - \mathbf{u}_h\|_{0,\Omega}$	$[[\mathbf{u} - \mathbf{u}_h]]$	$\|p - p_h\|_{0,\Omega}$
64	0.732828	16.625998	3.06364
256	0.156187	7.740131	1.0945
1,024	0.037076	3.750639	0.470646
4,096	0.009035	1.847880	0.212294
16,384	0.002228	0.916672	0.100356
65,536	0.000553	0.456425	0.048785

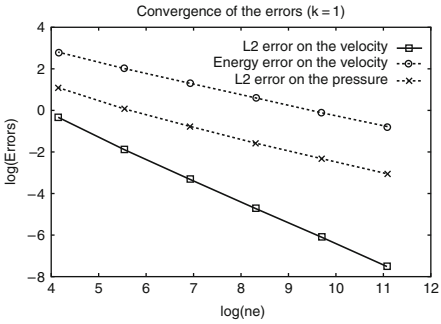


Fig. 1 Convergence rates for $k = 1$ ($\gamma = \gamma_1 = 10$, $\mu = 1$)

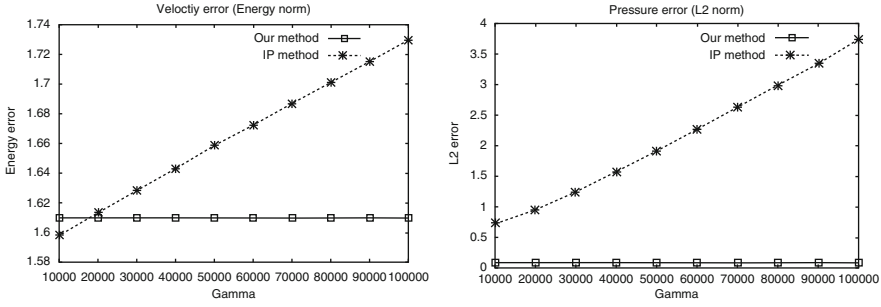
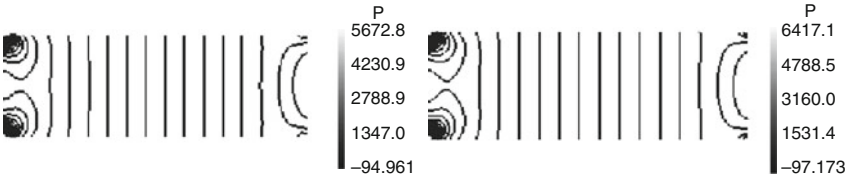


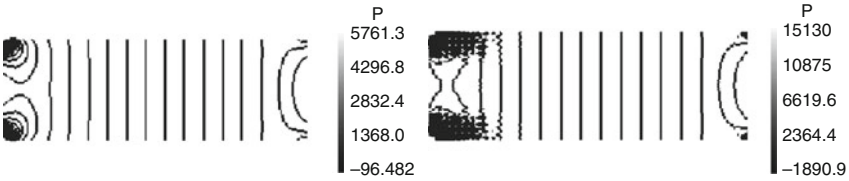
Fig. 2 Velocity error (*left*) and pressure error (*right*) with respect to γ for $k = 1$ ($\gamma_1 = 10$ and $\mu = 1$)

Fig. 3 Pressure obtained with Crouzeix-Raviart finite elements



(a) Pressure for $\gamma = 10(\mu = 1, \gamma_1 = 10)$. Ours: left; IP: right

(b) Pressure for $\gamma = 100(\mu = 1, \gamma_1 = 10)$. Ours: left; IP: right



(c) Pressure for $\gamma = 1000(\mu = 1, \gamma_1 = 10)$. Ours: left; IP: right

Fig. 4

a parabolic condition on the inflow, i.e., $\mathbf{u} \cdot \mathbf{n} = a(0.01^2 - y^2)$, and noticed a lack of stability of the IP method when γ become large. In the non-smooth case presented below, i.e., when imposing $\mathbf{u} \cdot \mathbf{n} = 1$ on the inflow boundary, the lack of stability is visible at rather small values of the stabilization parameter. Note that in this case, the solution does not belong to $\mathbf{H}^2(\Omega) \times \mathbf{H}^1(\Omega)$. In order to dispose of a reference solution, we have computed it by means of nonconforming finite elements of Crouzeix–Raviart (see Fig. 3 for the pressure). We have represented in Fig. 4a–c the pressure obtained with the two methods for different values of γ. As γ increases, the pressure computed with the IP method gets worse whereas our method is robust.

References

1. Arnold, D. N.: An interior penalty finite element method with discontinuous elements. SIAM J. Numer. Anal. **19**, 742–760 (1982)
2. Becker, R., Capatina, D., Joie, J.: A dG method for the Stokes equations related to nonconforming approximations. INRIA Research Report (2009). http://hal.inria.fr/inria-00380772
3. Brenner, S.: Korn's inequalities for piecewise H_1 vector fields. Math. Comp. **73**, 1067–1087 (2004)
4. Girault, V., Raviart, P.-A.: Finite element methods for Navier–Stokes equations. Springer, Berlin (1986)
5. Girault, V., Rivière, B., Wheeler, M.: A discontinuous Galerkin method with nonoverlapping domain decomposition for the Stokes and Navier–Stokes problems. Math. Comp. **74**, 53–84 (2005)
6. Mardal, K.-A., Winther, R.: An observation on Korn's inequality for nonconforming finite element methods. Math. Comp. **75**, 1–6 (2005)

Numerical Simulation of the Stratified Flow Past a Body

L. Beneš and J. Fürst

Abstract The article deals with the numerical simulation of 2D and 3D unsteady incompressible stratified flows. Initial system of equations is the Boussinesq approximation of the Navier–Stokes equations. The flow field in the towing tank with a moving sphere is modeled for a wide range of Richardson numbers. The obstacle is modeled via penalization technique. The resulting set of partial differential equations is then solved by the fifth-order finite difference WENO scheme, or by the second-order finite volume AUSM MUSCL scheme. For the time integration, the second-order BDF method was used. Both schemes are combined with the artificial compressibility method in dual time.

1 Introduction

Stratification plays important role in many industrial and environmental problems. Several years we are interested in modeling of the stratified and unstratified flow in various applications (pipe, atmospheric boundary layer etc.). The present work was motivated by a desire to obtain a better understanding of these effects.

2 Boussinesq Approximation

This type of flow can be often assumed to be incompressible, but yet the density is not constant owing to temperature changes, gravity, etc. For description of this flow, the Navier–Stokes equations for viscous incompressible flow with variable density is used. These equations are simplified by the Boussineq approximation.

L. Beneš (✉) and J. Fürst
Department of Technical Mathematics, Faculty of Mechanical Engineering, Czech Technical University in Prague, Karlovo náměstí 13, CZ-12135 Praha 2, Czech Republic
e-mail: benes@marian.fsik.cvut.cz, Jiri.Furst@fs.cvut.cz

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_15,

Density and pressure are divided into two parts: a background field (with subscript $_0$) plus a perturbation. The background field fulfill the hydrostatic balance equation $\partial p_0(z)/\partial z = -\rho_0(z)g$. The system of equations obtained is partly linearized around the average state ρ_*. The resulting set of equations can be written in the form

$$P W_t + F^i(W)_x + G^i(W)_y + H^i(W)_z = \nu(F^v(W)_x + G^v(W)_y + H^v(W)_z) + S(W). \tag{1}$$

$$F^i = [\rho u, u^2 + \frac{p}{\varrho_*}, uv, uw, u]^T,\ G^i = [\rho v, uv, v^2 + \frac{p}{\varrho_*}, vw, v]^T,\ H^i = [\rho w, uw, vw, w^2 + \frac{p}{\varrho_*}, w]^T,$$
$$F^v = [0, u_x, v_x, w_x, 0]^T, \quad G^v = [0, u_y, v_y, w_y, 0]^T, \quad H^v = [0, u_z, v_z, w_z, 0]^T.$$

where $W = [\rho, u, v, w, p]^T$, is vector of unknown variables respectively, the density perturbation, three velocity components and the pressure perturbation. $S = [-vd\rho_0/dz, 0, 0, -g, 0]^T$ is the gravity and source term and $P = diag(1, 1, 1, 1, 0)$. To describe the stratification, the following bulk Richardson number is used:

$$Ri = \frac{g\frac{d\rho_0}{dz}}{\varrho_* U^{ob^2}}$$

where U^{ob} is velocity of the moving obstacle.

3 Numerical Schemes

Two different numerical schemes were used for solving mentioned problem. The time discretization is the same in both cases. For the spatial were used either the flux-splitting method with WENO interpolation or the finite volume AUSM MUSCL scheme with the Hemker–Koren limiter.

Flux Splitting for Incompressible Flows
The discretization in space is achieved by standard fourth-order differences for viscous terms. For discretization of the inviscid fluxes the following special high-order flux-splitting method was used. Divide the inviscid flux $F^{in}(W)$ into two parts, the convective flux $F^c(W) = [\rho u, u^2, uv, uw, 0]^T$ and the pressure flux $F^p(W) = [0, p, 0, 0, \beta^2 u]^T$, then approximate the flux derivative by

$$F^i(W)_x\big|_i \approx \frac{1}{\Delta x}\left[F^c_{i+1/2} - F^c_{i-1/2}\right] + \frac{1}{\Delta x}\left[F^p_{i+1/2} - F^p_{i-1/2}\right]. \tag{2}$$

The high-order weighted ENO scheme [9] is chosen as the interpolation method (only the spatial index i in the x-direction is preserved, the remaining two indexes are omitted). The original WENO interpolation uses an upwind bias and it can be formally written in the following form (function `weno5` is described in [9]):

$$\phi_{i+1/2} = \begin{cases} \phi^+_{i+1/2} = \texttt{weno5}(\phi_{i-2}, \phi_{i-1}, \phi_i, \phi_{i+1}, \phi_{i+2}) & \text{if } u_{i+1/2} > 0, \\ \phi^-_{i+1/2} = \texttt{weno5}(\phi_{i+3}, \phi_{i+2}, \phi_{i+1}, \phi_i, \phi_{i-1}) & \text{if } u_{i+1/2} \le 0. \end{cases} \tag{3}$$

By mathematical analysis the convective part is discretized by simple upwinding, the third component of the pressure is approximated by backward differencing and the fourth component by a forward difference. The final scheme takes the form

$$u_{i+1/2} := (u^+_{i+1/2} + u^-_{i+1/2})/2, \quad p_{i+1/2} := (p^+_{i+1/2} + p^-_{i+1/2})/2, \tag{4}$$

$$F^c(W)_{i+1/2} := ((\rho u)^\pm_{i+1/2}, (u^2)^\pm_{i+1/2}, (uv)^\pm_{i+1/2}, (uw)^\pm_{i+1/2}, 0)^T$$

$$F^p(W) := (0, p_{i+1/2} + \beta \frac{u^+_{i+1/2} - u^-_{i+1/2}}{2}, 0, 0, u_{i+1/2} + \frac{p^+_{i+1/2} - p^-_{i+1/2}}{2\beta})^T,$$

where $+$ or $-$ is taken in the convective flux according to the sign of $u_{i+1/2}$.

A similar algorithm is applied in other directions for the fluxes G, H. The resulting scheme has high-order accuracy in space. It was validated for the case of compressible inviscid flows by a computation of shock-vortex interaction; see [8].

AUSM Scheme

The finite volume AUSM scheme was used for spatial discretization of the inviscid fluxes in our second scheme.

$$\int_\Omega (F^i_x + G^i_y + H^i_z) dS = \oint_{\partial\Omega} (F^i n_x + G^i n_y + H^i n_z) dl$$
$$\approx \sum_{k=1}^{6} \left[u_n \begin{pmatrix} \varrho \\ u \\ v \\ w \\ \beta^2 \end{pmatrix}_{L/R} + p \begin{pmatrix} 0 \\ n_x \\ n_y \\ n_z \\ 0 \end{pmatrix} \right] \Delta l_k, \tag{5}$$

where n is the normal vector, u_n the normal velocity vector, and $(q)_{L/R}$ are quantities on the left/right hand side of the face. These quantities are computed using MUSCL reconstruction with the Hemker–Koren limiter [7] in the form [3]:

$$q_R = q_{i+1} - \frac{1}{2}\delta_R, \quad q_L = q_i + \frac{1}{2}\delta_L,$$

$$\delta_{L/R} = \frac{a_{L/R}(b^2_{L/R} + 2) + b_{L/R}(2a^2_{L/R} + 1)}{2a^2_{L/R} + 2b^2_{L/R} - a_{L/R}b_{L/R} + 3},$$

$$a_R = q_{i+2} - q_{i+1}, \quad a_L = q_{i+1} - q_i, \quad b_R = q_{i+1} - q_i, \quad b_L = q_i - q_{i-1}.$$

Since the pressure is discretized using central differences, the scheme is stabilized following [4] by a pressure diffusion of the form

$$F_{\mathbf{d}i+1/2,j} = \left(0,\ 0,\ 0,\ 0,\ \eta \frac{p_{i+1,j} - p_{i,j}}{\beta_x}\right)^T, \qquad \beta_x = w_r + \frac{2\nu}{\Delta x}$$

where T denotes transpose and w_r is a reference velocity (in our case the maximum velocity in the flow field). Viscous fluxes are discretized using central differences on the dual mesh. This scheme is second-order accurate in space.

Time Integration
The spatial discretization yields a system of ODE in the physical time t variable, which is solved by the second-order BDF formula

$$P\frac{3W^{n+1}-4W^n+W^{n-1}}{2\Delta t}+(\widetilde{F}^i_x+\widetilde{G}^i_y+\widetilde{H}^i_y)(W^{n+1})-\nu(\widetilde{F}^v_x+\widetilde{G}^v_y+\widetilde{H}^v_y)(W^{n+1})=\widetilde{S}^{n+1}. \tag{6}$$

By $\widetilde{}$ is denoted above described numerical approximation of the fluxes. Arising set of equations is solved by an artificial compressibility method in the dual time τ by an explicit 3-stage second-order Runge–Kutta method of the second order.

4 Obstacle Modeling

We are interested in the modeling of the flows past a moving body. There are various possibilities how to model body (e.g. moving mesh, immersed boundary see [6]). In our computations, the obstacle is modeled very simply as a source term emulating a porous media with high resistance. The source term $S(W)$ in this case takes the form

$$\left[-w\frac{d\rho_0}{dz},0,0,-g,0\right]^T+\frac{\chi(x,y,z,t)}{K}\left[0,U^{ob}-u,V^{ob}-v,W^{ob}-w,0\right]^T, \tag{7}$$

where K corresponds to small permeability and $\chi(x,y,z,t)$ is the characteristic function of the obstacle, which moves with velocity (U^{ob},V^{ob},W^{ob}).

To estimate the influence of the permeability K, and also numerical tests were published in [1].

5 Numerical Results

Solved problem is well known as the towing tank problem. The towing tank is a brimfull channel with the moving obstacle inside. Technical parameters: dimensions $8\times 4\,m$ in 2D or $8\times 4\times 1\,m$ in 3D, $\rho_*=1\ kg\,m^{-3}$, the kinematic viscosity $\nu=10^{-4}\,m^2\,s^{-1}$ and stable density gradient $d\rho_0/dz=-0.1\,kg\,m^{-4}$. The obstacle is a sphere of radius $0.1\,m$, located $1m$ from the left wall and at the middle of height and width. At time $t=0$ the obstacle starts moving with constant velocity $U^{ob}=1\,m\,s^{-1}$.

Homogeneous Dirichlet boundary conditions for the velocity and Neumann conditions for the density and pressure disturbances were used in 2D. In 3D, these boundary conditions were extended by periodic boundary conditions in the

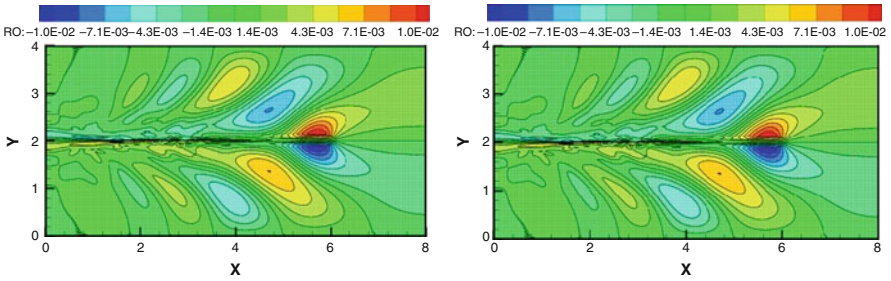


Fig. 1 Comparison of isolines of the density disturbances at the time $t = 5s$, $g = 100$, $Ri = 10$. AUSM MUSCL scheme *left* and WENO5 *right*

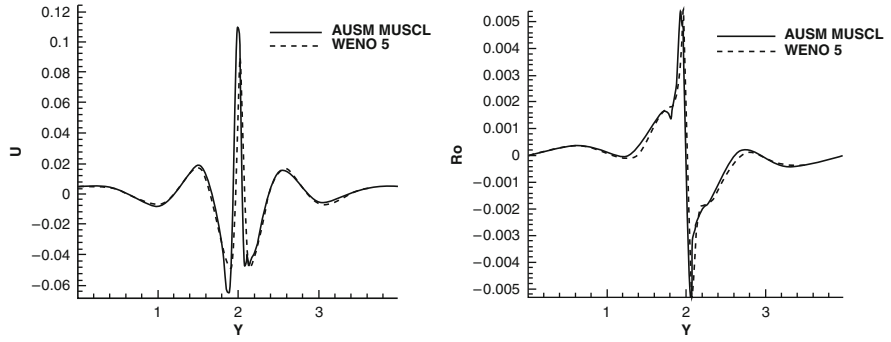


Fig. 2 Comparison of both schemes. Transversal distribution of the u-velocity component (*left*) and ρ (*right*), $x = 2.25$, $Ri = 10$, time $t = 5s$

y-direction. Cartesian grids with $320 \times 40 \times 160$ cells and 320×160 cells in 2D were used. The simulations were performed for wide range of stratifications $Ri \in < 0, 100 >$. The influence of mesh and permeability parameter were tested in our previous studies, [1, 5].

In the Figs. 1 and 2 we can see comparison of the schemes in 2D (with similar spherical obstacle) in the form of density isolines at the time $t = 5s$ and transversal distribution of the computed quantities. These figures exhibit good agreement between both methods, especially further from the obstacle, while small differences occur behind the sphere. The maximal values predicted by WENO 5 scheme at the height midpoint are somewhat lower. Figure 3 displays the comparison of the isosurfaces of the vorticity in 3D for the Richardson numbers $Ri = 10$ computed by WENO scheme and AUSM MUSCL. Small differences occur on the higher distances from body.

Stratified flows is characterized by a variation of fluid density in the vertical direction that can result in qualitative and quantitative changes of the flow by buoyancy. Figure 4 displays the dependence of the flow on the level of stratification. A comparison of the isosurfaces of density perturbation for three different Richardson numbers ($Ri = 1, 10, 100$) is presented at the time $t = 5s$. In the

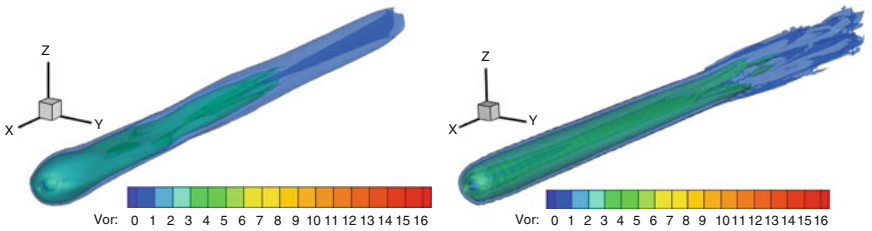


Fig. 3 Magnitude of the vorticity distribution for the Richardson numbers $Ri = 10$. WENO5 scheme (*left*) and AUSM MUSCL scheme (*right*), time $t = 5s$

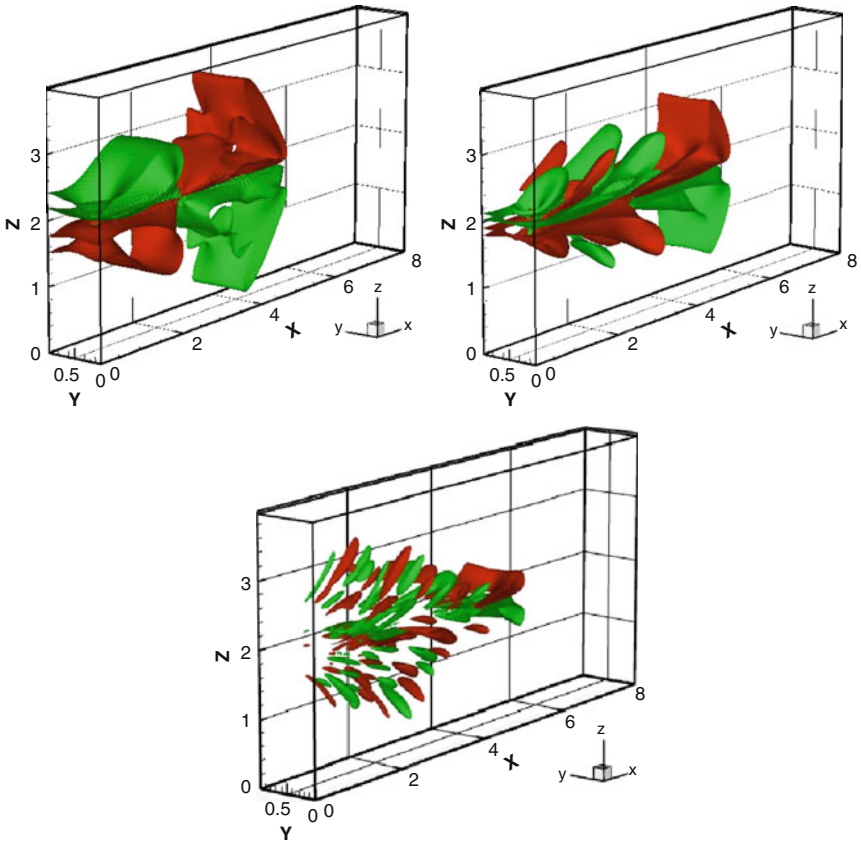


Fig. 4 Isosurfaces of the density perturbations at the time $t = 5s$ for $Ri = 1, 10, 100$, AUSM MUSCL scheme

case of low stratification, the solution takes the form of Karman vortex street. When the level of stratification increases, vortices are damped and internal gravity waves are generated. Its frequency is given by the Brunt–Väisälä frequency.

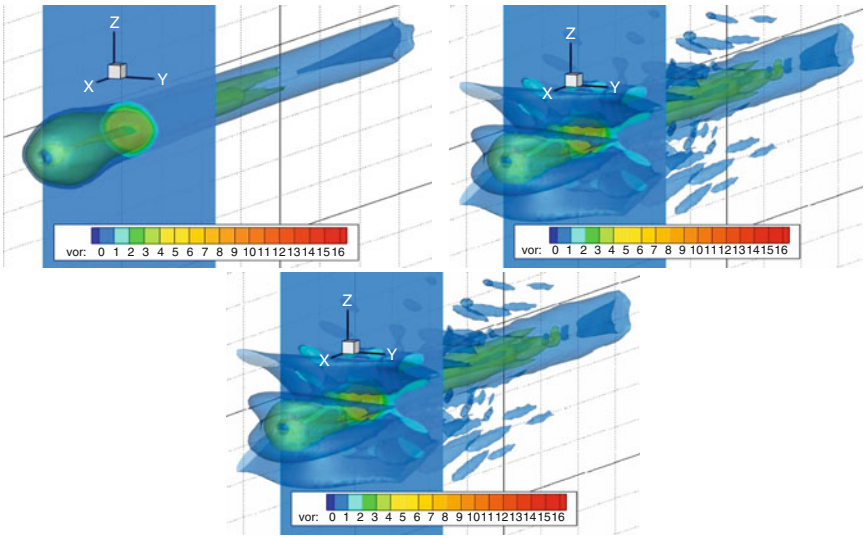


Fig. 5 Isosurfaces of the magnitude of the vorticity at the time $t = 5s$ for $Ri = 1, 10, 100$, WENO5 scheme

The isosurfaces of the vorticity in 3D for the Richardson numbers $Ri = 1, 10, 100$ are shown in Fig. 5. Stable stratification generally suppresses any vertical mixing of mass and momentum. This tendency can be seen at the y–z cross-section. In the case $Ri = 1$, the influence of stratification is small and the shape of vorticity in the cross-section is close to a circle. For the higher stratification $Ri = 10$ the vortices are damped especially in z direction, which leads to an asymmetry in the vorticity isosurface. Even greater asymmetry is in the case $Ri = 100$, the shape of isosurface is significantly different.

6 Conclusion

Experiments in the atmosphere are very expensive. So, the numerical simulations are often the only single source of information. Since the solution can depend on the numerical scheme, mesh etc, a comparison of solutions obtained using different methods eliminates this dependence.

Two numerical methods for simulation of 2D and 3D stratified flows have been developed and compared. Numerical results were obtained for Richardson numbers $Ri \in < 0, 100 >$. These results are in good mutual agreement and also match physical expectations. Small differences that emerged between schemes especially in 3D case require deeper investigation.

For simulation of the real atmospheric boundary layer flow, the significant question is the choice of appropriate boundary conditions. Our simple boundary conditions are suitable for laboratory type flows.

Acknowledgements This work was supported by Research Plans MSM 6840770003, GACR Project No.101/09/1539 and COST OC 167.

References

1. Beneš L., Fürst J., Fraunié Ph.: Numerical simulation of the towing tank problem using high order schemes. BAIL 2008 – Boundary and Interior Layers. Lecture Notes in Computational Science and Engineering 69, Springer, Berlin, 2009, ISSN 1439-7358
2. Berrabaa S., Fraunié P., Crochet M.: 2D Large Eddy Simulation of highly stratified flows : the stepwise structure effect. Advances in Computation . Scientific Computing and Applications, **7**, 179–185 (2001)
3. Blazek J.: Computational Fluid Dynamics: Principles and Applications. Elsevier Science, Amsterdam, 2001, ISBN 0080430090
4. 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)
5. Fraunie Ph., Beneš L., Fürst J.: Numerical simulation of the stratified flow. In: Proceeding of conference Topical Problems of Fluid Mechanics 2008, 5–8 (2008)
6. Fuka V., Brechler J.: LES of contaminant dispersion in an idealized geometry. Proceeding of CMFF09, 138–142, Budapest 2009, Hungary ISBN 978-963-420-984-3
7. Hemker P.W., Koren B.: Multigrid, defect correction and upwind schemes for the steady Navier–Stokes equations. Numerical methods for fluid dynamics III; pp. 153–170, Oxford University Press, Oxford, 1988
8. Kozel K., Angot Ph., Fürst J.: TVD and ENO schemes for multidimensional steady and unsteady flows. In: Finite Volumes for Complex Applications, 283–290, Hermes (1996)
9. Shu Chi-Wang, Jiang Guang-Shan: Efficient implementation of weighted eno schemes. Journal of Computational Physics, **126**, 202–228 (1996)

A Flexible Updating Framework for Preconditioners in PDE-Based Image Restoration Algorithms

Daniele Bertaccini and Fiorella Sgallari

Abstract We propose the solution of some discretized partial differential equation models for image denoising and deblurring by iterative linear system solvers accelerated by a simple but flexible framework for updating incomplete factorization preconditioners that presents a computational cost linear in the number of the image pixels. Here we perform some tests where the efficiency of the strategy is confirmed.

1 Motivations

Image restoration models based on partial differential equations, or PDE for short, can be discretized by finite differences, finite elements or finite volumes by using explicit, semi implicit (see e.g., [7]) and fully implicit (see e.g., [11]) time stepping in order to obtain a discrete model. Unfortunately, sometimes general preconditioning techniques for iterative methods for the solution of algebraic linear systems generated by discretized models with implicit or semi-implicit schemes can show an overall computational complexity in time and/or space higher than for the same iterative solver without preconditioning. In view of these facts, we propose here a technique that updates incomplete factorization-based preconditioners with a linear computational cost in the number of image pixels, as the techniques based on additive operator splitting (AOS) such as that proposed in [9] and in [12], which are very well suited for denoising only.

Recently, preconditioners for solving sequences of shifted linear systems arising in partial differential equations of evolutionary type were proposed in [2] and [3].

D. Bertaccini (✉)
Università di Roma "Tor Vergata", Dipartimento di Matematica, viale della Ricerca Scientifica, 00133 Roma (I)
e-mail: bertaccini@mat.uniroma2.it

F. Sgallari
Università di Bologna, Dipartimento di Matematica and CIRAM, Via Saragozza, 8, 40123 Bologna (I)
e-mail: sgallari@dm.unibo.it

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_16,

Here we generalize these preconditioners in order to update, downdate and regularize the incomplete factorizations of a sequence of symmetric positive definite matrices generated by the discretization of a nonlinear PDE model with selective diffusion (see [1] and [10]). In particular, we focus on using banded updates and banded approximations of the inverse of the component matrices of incomplete factorizations in LDL^T-form, where L is a lower triangular and D is a diagonal matrix, respectively. We compute the incomplete factorization to be updated not necessarily on the given observed image, perturbed by a certain level of blur and noise, and use the regularized updates for restoring images with, e.g., different blur. An analysis of these techniques without regularizing and by calculating the first preconditioner always on the given perturbed image can be found in [4].

In Sect. 2 a generalized Alvarez–Lions–Morel PDE model for image selective smoothing discretized with a semi implicit complementary volume scheme is briefly sketched. Section 3 introduces the proposed incomplete factorization preconditioners for the discrete operators in Sect. 2 and some remarks on their application. Section 4 includes some tests on images with noise and blur and comparisons with the Cholesky threshold preconditioner.

2 A Generalized Alvarez–Lions–Morel Model

In order to simplify the treatise, we will focus on the integration of a generalized Alvarez–Lions–Morel-like nonlinear equation for selective smoothing and deblurring (see [1] and [10])

$$u_t = g(|\nabla G_\sigma * u|)\,|\nabla u|_\epsilon \nabla \cdot \left(\frac{\nabla u}{|\nabla u|_\epsilon}\right) - \alpha |\nabla u|_\epsilon K^T \left(K\,u - u^0\right), \quad (t,x) \in I \times \Omega, \tag{1}$$

K is the blur operator, α is a parameter controlling the blur term, Ω can be assumed as a bounded rectangular domain, I is a scale (time) interval, g is a nonincreasing real function which tends to zero as its argument tends to infinity, G is a smoothing kernel and $|\nabla \cdot|_\epsilon$ means that we regularize in the sense of Evans and Spruck [6] to avoid zero in the denominators:

$$|\nabla u|_\epsilon = \sqrt{|\nabla u|^2 + \epsilon^2}. \tag{2}$$

Equation (1) is coupled with initial and boundary conditions

$$\begin{aligned} \frac{\partial u}{\partial \mathbf{n}} &= 0, \quad (t,x) \in I \times \partial\Omega \\ u(0,x) &= u^0(x), \quad x \in \Omega, \end{aligned} \tag{3}$$

$u(t,x)$ is the unknown image function, u^0 is the grey level of the image to be processed, $\mathbf{n}$ is the unit normal to the boundary of Ω.

We consider the linear semi-implicit fully discrete complementary volume scheme for (1)–(3) proposed in [7], but the same approach can be applied if the function $g(\cdot)$ is inside the divergence operator in (1).

Linear semi-implicit fully discrete complementary volume scheme for equation (1)–(3) proposed in [7] is generated by discretization of (1)–(3) choosing a uniform discrete time-scale increment $\tau = T/N$, and replacing time-scale derivative by backward differences. To simplify notation, let $\alpha = 0$ in (1). In order to generate a semi implicit scheme, nonlinear terms in (1) are computed on the basis of the previous scale step while linear terms on the current one:

$$u^n = u^{n-1} + \tau g^{n-1} |\nabla u^{n-1}|_\epsilon \nabla \cdot \left(\frac{\nabla u^n}{|\nabla u^{n-1}|_\epsilon} \right) \tag{4}$$

where

$$g^{n-1} = g\left(|\nabla G_\sigma * u^{n-1}| \right),$$

$g^{n-1} > 0$ according to its definition [5, 7]. By integrating over a co-volume V_i and using linear basis functions [7], we get the discrete system to be solved for u_i^n, $i = 1, \ldots, M$ (M are the nodes of the triangulation), $n = 1, 2, \ldots,$:

$$u_i^n = u_i^{n-1} + \frac{\tau}{b_i^{n-1}} \sum_{j \in C_i} a_{i,j}^{n-1} \left(u_j^n - u_i^n \right), \quad i = 1, \ldots, M, \quad n = 1, 2, \ldots, \tag{5}$$

with

$$b_i^{n-1} = \frac{|V_i|}{g(|\nabla G_\sigma * u^{n-1}|_i) |\nabla u_i^{n-1}|_\epsilon}, \quad a_{i,j}^{n-1} = \frac{1}{h_{i,j}} \sum_{T \in \varepsilon_{i,j}} \frac{|c_{i,j}|}{|\nabla u_T^{n-1}|_\epsilon},$$

where ∇u_T^{n-1} is related to the value assumed by the gradient of the piecewise linear function which is constant on every simplex T of the triangulation while $|c_{i,j}|$ is the length of the portion of the co-edge that is the perpendicular bisector of the edge connecting the related nodes; see [7, Sect. 2]. By rewriting the discrete system (5) as a linear system $C_n u_n = f_n$ whose solution vector is

$$u_n = (u_1^n, \ldots, u_M^n)^T,$$

the matrix C_n has as entries $c_{i,j}^n$, $i, j = 1, \ldots, M$, $n = 1, 2, \ldots,$ where

$$c_{i,j}^n = b_i^{n-1} + \tau \sum_{j \in C_i} a_{i,j}^{n-1}, \; i = j; \quad c_{i,j}^n = -\tau a_{i,j}^{n-1}, \; i \neq j, \tag{6}$$

we have the result.

Theorem 1. *The matrices of the linear systems generated at each time scale by discrete complementary volume scheme (5) are symmetric and diagonally dominant M-matrices.*

Proof. See [7, Proposition 1]. □

By [2, Theorem 4.2], we can expect a fast decay of the entries of the inverse of the matrices $\{C_n\}$ (6) by moving away from the main diagonal. However, it is also worth to recall that, as observed in [7, p. 682], with a decreasing ϵ, the diagonal dominance of the matrix of the linear system generated by (5) tends to disappear and this can drop down the decay properties of the matrix C_n, with a possible negative effect on the acceleration properties of the updates. More details and comments on this effect can be found in [4].

By including the blur operator in (5), the discretized model gives a sequence of linear system whose matrices have a larger and denser band with respect to those without blur contribution (i.e., with $\alpha = 0$ in (1)).

3 Updating Incomplete Factorizations

In order to solve the discretized counterpart of model (1), we need to solve a sequence of sparse linear systems

$$C_n u_n = f_n, \quad n = 1, \ldots, N, \tag{7}$$

where C_n is the nth matrix given by the complementary volume discretization of (1). Sequences of linear systems can be generated by either the quasi-Newton step of an implicit solver such as, e.g., in [11] or by a semi-implicit formulation of the discretization in time such as, e.g., in [1, 7] and here.

Let us assume C_1 (or another matrix C_0 generated by the discretization of the PDE model with slightly different parameters, e.g., different blur or noise) as the seed matrix C_{seed} and suppose we can compute the LDL^T factorization of C_{seed}, i.e.,

$$C_{seed} = LDL^T, \tag{8}$$

where L is unit lower triangular and D is diagonal. It is worth to note that the factorization (8) will *never* be computed in practice, we need to state this only to justify our approximations in the sequel. By denoting with C_j the matrix we want to precondition, we can look for a formal factorization for C_j updating (8), which, assuming C_j factorizable, can be written as

$$C_j = L(D + E_j)L^T, \tag{9}$$

where E_j serves as an update for the diagonal matrix D in (8). By a generalization of an argument from [3], i.e., by considering that C_{j+1} can be potentially completely different from C_j, we get

$$E_j = Z^T \Delta_j Z, Z = \left(L^T\right)^{-1} \tag{10}$$

where

$$\Delta_j = C_j - C_{seed}, \quad j = 1, \ldots, n. \tag{11}$$

Note that j starts from 1 because C_1 can differ from C_{seed} in this setting. Let us compute an incomplete factorization $\tilde{L}\,\tilde{D}\,\tilde{L}^T$ for the (sparse) seed matrix C_{seed} with a threshold chosen in order to get a $\tilde{L}$ whose nonzero entries are of the order of those of C_{seed}. Our candidate preconditioner for C_j is

$$P_j = \tilde{L}\left(\tilde{D} + \tilde{E}_j\right)\tilde{L}^T, \quad j = 1, \ldots, n, \tag{12}$$

and for C_j^{-1} is

$$Z_j = \tilde{Z}\left(\tilde{D} + \tilde{E}_j\right)^{-1}\tilde{Z}^T, \quad j = 1, \ldots, n. \tag{13}$$

A sparsified version $\tilde{Z}$ of $Z = \left(L^T\right)^{-1}$ can be generated by approximating the inverse of $\tilde{L}^T$ and discarding the entries whose absolute value is below a specified drop tolerance. The matrix $\tilde{Z}$ is sparse (banded) by the decaying properties of inverses of diagonally dominant matrices; see [2, Theorem 4.2]. The perturbation $\tilde{E}_j$ that we add to $\tilde{D}$ in (12) and (13) is an approximation of $E_j = Z^T\,\Delta_j\,Z$ that could in principle be given by $\tilde{E}_j = \tilde{Z}^T\,\tilde{\Delta}_j\,\tilde{Z}$, $\tilde{\Delta}_j$ an approximation of Δ_j. Here we take $\tilde{\Delta}_j = \Delta_j$, which is a narrow band matrix in our setting. Therefore, for each PCG iteration, we need to solve a linear system with matrix $\tilde{D} + \tilde{E}_j$ even if we use the preconditioner in the inverse factorized form (13). If the decay of the entries in Z is fast enough, e.g., if the matrices C_n are strongly diagonally dominant, then it could be enough to define $\tilde{E}_j$ as a *diagonal* approximation of $\tilde{Z}^T\,\tilde{\Delta}_j\,\tilde{Z}$. The underlying diagonal updates do not require to solve extra linear systems in order to apply the preconditioner (13). Whenever needed, we regularize the underlying diagonal updates by replacing each negative (or less than 10^{-2} times the drop tolerance) diagonal entry in $\tilde{D} + \tilde{E}_j$ by the corresponding one of $\tilde{D}$; see numerical experiments in Sect. 4.

The computational cost is expected to be much lower for PCG using (13) in a parallel or multicore setting because no triangular linear system is needed to be solved in order to apply the preconditioner. This is confirmed by numerical experiments. Comments and an analysis of the impact of the variation of the parameters of the model (1) on the spectral properties of the preconditioners (12), (13) can be found in [4]. We stress that the generalization proposed in this contribution with respect to the setting in [4] (here the seed incomplete factorization $\tilde{L}\,\tilde{D}\,\tilde{L}^T$ can be computed from a perturbed image with possibly different parameters with respect to the given one) does not permit the application of the convergence results in [4, Sect. 3.2] because $\lim_{\tau\to 0} C_j$ can differ from C_{seed}. In view of this, we plan to extend the convergence theory of [4] in a future research.

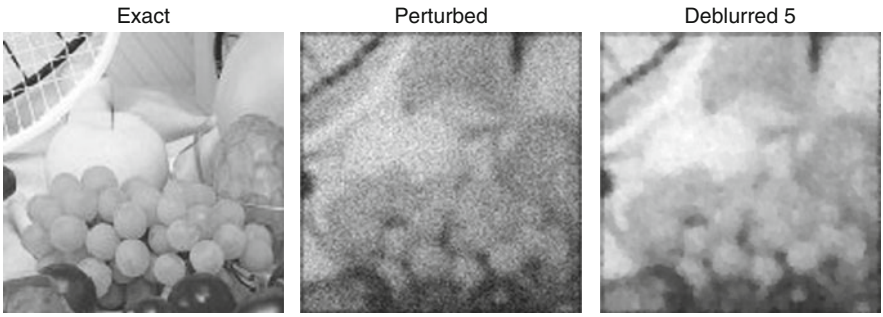


Fig. 1 (from *left* to *right*) original 128×128 fruit picture; with 10% noise and blur; after five time steps (parameters: $\tau = 10^{-5}$; $band = 3$ and $\sigma = 3$ for the blur operator)

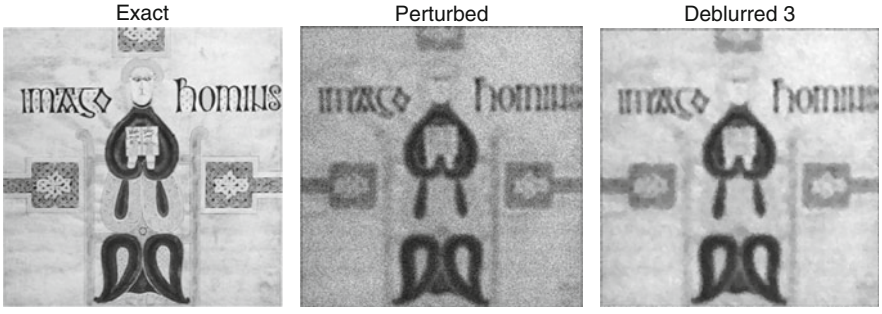


Fig. 2 (from *left* to *right*) original 256×256 picture; with 10% noise and blur; after three steps (parameters: $\tau = 10^{-5}$; $band = 3$ and $\sigma = 3$ for the blur operator)

4 Numerical Results

We report here some experiments using regularized updated preconditioners (*reg inv updated prec* in the tables) for the semi-implicit complementary volume scheme described in [7] generalized for denoising and deblurring images. The blur contribution is given by the Matlab function `blur.m`; see [8]. This function is used with parameters `band=` 3 and `sigma=` 3. The former specifies the bandwidth of the Toeplitz blocks and the latter the variance of the Gaussian point spread function; see [8] for more details. We used a preliminary Matlab R2009a 64bit implementation of our algorithms on a Intel T9550 2.66Ghz laptop with 4Gb RAM and second core disabled.

In the tables below, for each scale step, we report the global number of conjugate gradient iterations, the time (in seconds) needed for solving the underlying linear systems (in bold the best performance) and the ratio of the number of nonzero elements of the preconditioner over the nonzeros entries of C_{seed} ($nnz(P)/nnz(C)$).

In all experiments we consider a time scale of 10^{-5} (pixel intensities are scaled in the interval $[0, 1]$) and generate the steps needed to restore a given image. With a

Table 1 Timings and iterations for denoising and deblurring the 128×128 fruit picture (see Fig. 1) with 5 time steps. Noise level 10%, blur parameters $band = 3$ and $\sigma = 3$

Algorithm	Iterations	Time (s)	$nnz(P)/nnz(C)$
reg inv updated prec	147	0.87	0.985
reg inv updated prec[a]	85	**0.5**	0.811
Incomplete Cholesky (10^{-1})	37	1.5	0.35–0.38
Incomplete Cholesky (10^{-2})	15	2.9	0.67–0.69
Unpreconditioned	433	2.6	–

[a]The matrix C_{seed} for the updated inverse preconditioner is computed for a model with noise level 10% and **band** $= 5$, $\sigma = 3$

Table 2 Timings and iterations for denoising and deblurring the 256 × 256 book picture (see Fig. 2) with 3 time steps. Noise level 10%, blur parameters $band = 3$ and $\sigma = 3$.

Algorithm	Iterations	Time (s)	$nnz(P)/nnz(C)$
reg inv updated prec	72	**1.7**	1.7
Incomplete Cholesky (10^{-1})	37	2.5	0.46 – 0.47
Incomplete Cholesky (10^{-2})	15	9.2	0.76 – 0.79
Unpreconditioned	525	11.2	–

Table 3 Timings and iterations for denoising and deblurring the 256 × 256 book picture with five time steps. Noise level 20%, blur parameters $band = 3$ and $\sigma = 3$

Algorithm	Iterations	Time (s)	$nnz(P)/nnz(C)$
reg inv updated prec	146	**3.1**	1.27
Incomplete Cholesky (10^{-1})	78	4.3	0.4–0.46
Incomplete Cholesky (10^{-2})	29	11	0.7–0.76
Unpreconditioned	1050	22.2	–

larger time step τ, but still reasonable for image restoration, we continue to observe an interesting speedup with respect to recomputing the incomplete factorization. We used $\epsilon = 10^{-2}$ in (2) and the function g in (1) is $g(s) = 1/(1 + K s^2)$, $K = 0.1$. The convolution is realized as in [7] by using σ less than τ. The matrix $\tilde{Z}$ is generated by sparsifying the inverse of incomplete Cholesky factorizations of C_{seed} with a threshold of 0.01. Results are compared with threshold incomplete Cholesky factorizations recomputed at each step (*IC* in the tables, in brackets the drop tolerances used: 0.1 and 0.01) The conjugate gradient iterations are stopped when the relative residual is less than 10^{-5}. The timings in the tables do not include the cost of the incomplete factorization for the updated preconditioner for two reasons: (1) this cost is small with respect to the cost for the iterations for our preconditioners, in particular if an efficient implementation is considered; (2) one could use a *seed* incomplete factorization that is already available from another experiment starting with the same original image, without recomputing the factors $\tilde{L}$ and/or $\tilde{Z}$ in (12) or in (13).

The proposed images are 128 × 128 and 256 × 256, but preliminary tests confirm that when the size of image increases, the speedups do not deteriorate and both the computational cost per linear iteration and the memory space required remain

linear in the number of image pixels. Experiments, comparisons and analysis of convergence behavior in a slightly less general setting are provided in [4]. We note that all the proposed numerical tests with regularized diagonal updated preconditioners in inverse form (13) report good savings with respect to recomputing the preconditioner at each step, reusing the same preconditioner computed for the first linear system or with respect to incomplete Cholesky preconditioners with various thresholds. Surprisingly, sometimes we observe a better quality of the updates by providing the incomplete factorization of the matrix of the linear system arising in the first scale step of a model with different blur parameters; see Table 1.

References

1. Alvarez, L., Lions, P.L., Morel, J. M.: Image selective smoothing and edge detection by nonlinear diffusion II. SIAM J. Numer. Anal. **29**, 845–866 (1992)
2. Benzi, M., Bertaccini, D.: Approximate inverse preconditioning for shifted linear systems. BIT **43**, 231–244 (2003)
3. Bertaccini, D.: Efficient preconditioning for sequences of parametric complex symmetric linear systems. ETNA **18**, 49–64 (2004)
4. Bertaccini, D., Sgallari, F., Updating preconditioners for image restoration. Manuscript, 2009
5. Catté, F., Lions, P.L., Morel, J.M., Coll, T.: Image selective smoothing and edge detection by nonlinear diffusion. SIAM J. Numer. Anal. **29**, 182–193 (1992)
6. Evans, L.C., Spruck, J.: Motion of level sets by mean curvature. I. J. Differential Geom. **33**, 635–681 (1991)
7. Handlovičová, A., Mikula, K., Sgallari, F.: Semi-implicit complementary volume scheme for solving level set like equations in image processing and curve evolution. Num. Math. **93**, 675–695 (2003)
8. Hansen, P.C.: Regularization Tools, version 4.1. Netlib repository, URL: http://www.netlib.org (2008)
9. Lu, T., Neittaanmäki, P., Tai, X.C.: A parallel splitting up method and its application to Navier-Stoke equations. Appl. Math. Lett. **4(2)**, 25–29 (1991)
10. Marquina, A., Osher, S.: Explicit algorithms for a new time dependent model based on level set motion for nonlinear deblurring and noise removal. SIAM J. Sci. Comput. **22**, 387–405 (2000)
11. Walkington, N.: Algorithms for computing motion by mean curvature. SIAM J. Numer. Anal. **33**, 2215–2238 (1996)
12. Weickert, J., ter Haar Romeny, B.M., Viergever, M.A.: Efficient and reliable schemes for nonlinear diffusion filtering. IEEE Trans. Image Process. **7**, 398–410 (1998)

Stabilized Finite Element Method for Compressible–Incompressible Diphasic Flows

M. Billaud, G. Gallice, and B. Nkonga

Abstract This paper concerns the simulation of two immiscible fluids separated by a moving interface. In this goal, a global and simple numerical approach in which the gas is considered compressible and the liquid incompressible is elaborated. The numerical simulation of bubble dynamics phenomena is presented to illustrate the proposed method.

1 Introduction

Diphasic interface flows, such as bubble flows, are often present in many environmental and engineering problems. Numerical simulation is a good way of understanding and controlling such flows in many industrial process.

The present work deals with a numerical method able to accuratly predict diphasic flows. There are many intrinsic difficulties associated to such flows. In order to simplify the problem we suppose there is no shock (in particular in the gas), the two fluids are immiscible, and the surface tension effets are neglected.

The issue of modelling for diphasic interface flows has been addressed in numerous research papers, see [1, 2, 5, 7] for example. In these articles three different approaches are considered. In [1], both liquid and gas are described as compressible fluids contrary to [5] in which the two fluids are incompressible. The two approaches have some drawbacks related to the inherent physical properties of each

M. Billaud (✉)
IMB, Université Bordeaux 1, 351 Cours de la Libération 33405 Talence, France
e-mail: marie.billaud@math.u-bordeaux1.fr

G. Gallice
CEA-CESTA, BP2, Le Barp, France
e-mail: gerard.gallice@cea.fr

B. Nkonga
Laboratoire J. A. Dieudonné, Universit de Nice Sophia-Antipolis, Parc Valrose, Nice, France
e-mail: boniface.nkonga@unice.fr

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_17,

phase. For example, in the first one the volume conservation of the liquid is not respected despite in the second one the compressibility of the gas is not taken into account. Here, we investigate a third strategy [2, 7], in which the gas is considered as compressible and the liquid as incompressible.

In [2], a numerical method was developped to treat the coupling of compressible and incompressible flows, solving each phase with different schemes, a total variation diminushing (TVD) finite volume scheme for Euler equation in the gas phase and a Marker and Cell (MAC) scheme for the incompressible Navier–Stokes equations in the liquid one, leading to a complex and non-global numerical method. The aim of this note is to present a global and simple approach that relies on the three basic components: the stabilized finite element [3, 4] for spatial approximation of Navier–Stokes equations, the Level Set method for tracking precisely the interface with a discontinuous Galerkin scheme [6] to solve the associated transport equation and and averaging approach [8] to treat the discontinuities at the interface.

This short paper is organized as follows. In a second part, the model used to describe the considered flows is presented. In the third part the numerical method is detailed and its good behaviour is illustrated in a fourth part.

2 Global Model

Using the unified Navier–Stokes equations, recalled in the first section, the global model considered to describe the diphasic flow motion is presented.

2.1 Navier–Stokes Equations Unified Form

Most flows obey the Navier–Stokes equations. Written in conservative form, these equations are:

$$\partial_t \mathbf{U} + \sum_{i=1}^{d} \partial_{x_i} \mathbf{F}_i^a = \sum_{i=1}^{d} \partial_{x_i} \mathbf{F}_i^v + \mathbf{S}, \tag{1}$$

where d is the spatial dimension, $\mathbf{U}$ is the vector of the conservative unknowns, $\mathbf{F}_i^a$ and $\mathbf{F}_i^d$ are the advective and diffusive flux, respectively, in the ith direction, $\mathbf{S}$ is a source vector. The partial derivative in time and in the ith direction are respectively ∂_t et ∂_{x_i}.

Using the vector of primitive unknowns, (see [3]), ${}^t\mathbf{Y} = (p, {}^t\mathbf{u}, T)$, with p the pressure, $\mathbf{u}$ the velocity vector, T the temperature, it is possible to rewrite (1) in the equivalent quasi-linear form:

$$L(\mathbf{Y}, \boldsymbol{\chi}, \rho, \mu, \kappa)\mathbf{Y} = \mathbf{S}(\mathbf{Y}, \rho), \tag{2}$$

with

$$L(\mathbf{Y}, \boldsymbol{\chi}, \rho, \mu, \kappa) = A_0(\mathbf{Y}, \boldsymbol{\chi}, \rho)\partial_t + \sum_{i=1}^{d} A_i(\mathbf{Y}, \boldsymbol{\chi}, \rho)\partial_{x_i} - \sum_{i,j=1}^{d} \partial_{x_i}\left(K_{ij}(\mathbf{Y}, \mu, \kappa)\partial_{x_j}\right).$$

In addition, $A_0 = \partial_{\mathbf{Y}}\mathbf{U}, A_i = \partial_{\mathbf{Y}}\mathbf{F}_i^a$ are the i-th Euler Jacobian matrices, and $K = [K_{ij}]$ is the diffusivity matrix where $\sum_{j=1}^{d} K_{ij}\partial_{x_j}\mathbf{Y} = \mathbf{F}_i^v$.

The previous matrices are expressed in term of physical quantities inherent to the considered fluid, such as the unknown $\mathbf{Y}$, the density ρ, the viscosity μ, the thermal conductivity κ and two thermodynamic coefficients $\boldsymbol{\chi} = (\alpha_p, \beta_T)$ which are:

$$\alpha_p = -\frac{1}{\rho}\left(\frac{\partial\rho}{\partial T}\right)_p \quad \text{the } \textit{volume expansivity} \text{ coefficient,} \tag{3}$$

$$\beta_T = \frac{1}{\rho}\left(\frac{\partial\rho}{\partial p}\right)_T \quad \text{the } \textit{isothermal compressibility} \text{ coefficient.} \tag{4}$$

Here the flow is only submitted to the gravity field then $\mathbf{S}$ only depends on ρ and $\mathbf{Y}$.

Quasi-linear equations for primitive unknowns (2) are well defined for both compressible and incompressible flows (see [4]), hence we call them the *unified Navier–Stokes equations*. This unified formulation is a good starting point for a global numerical scheme suitable for compressible-incompressible coupling.

2.2 Governing Equations for Diphasic Flow

In the present work the diphasic flow field is computed in a bounded multi-dimensional domain $\Omega \subset \mathbb{R}^d$. This domain is divided into two regions: Ω_1 (liquid) and Ω_2 (gas) seperated by the interface Γ.

Since there is no shock, in each phase Ω_i, the flow obeys the Navier–Stokes equation written in the quasi-linear unified form (2):

$$L(\mathbf{Y}_i, \boldsymbol{\chi}_i, \rho_i, \mu_i, \kappa_i)\mathbf{Y}_i = \mathbf{S}(\mathbf{Y}_i, \rho_i) \text{ in } \Omega_i \times \mathbb{R}^{+*}, \tag{5}$$

where the quantities $\mathbf{Y}_i, \boldsymbol{\chi}_i, \rho_i, \mu_i, \kappa_i$ are specific to each fluid i. The liquid (Ω_1) is supposed to be an incompressible and its density is supposed constant. In this case the two thermodynamic coefficients (3) and (4) are equal to zero:

$$\boldsymbol{\chi}_1 = (\alpha_p^1, \beta_T^1) = \mathbf{0}. \tag{6}$$

Contrary to the liquid, the gas (Ω_2) is described by the compressible Navier–Stokes equations. To close the system the perfect gas equation of state is adopted

leading to the relation:

$$\rho_2 = \frac{p_2}{RT_2}, \tag{7}$$

where R is the gas constant. Using this equality the two thermodynamic coefficients become:

$$\boldsymbol{\chi}_2 = (\alpha_p^2, \beta_T^2) = \left(\frac{1}{T_2}, \frac{1}{p_2}\right). \tag{8}$$

Finally, to capture precisely the interface Γ, a Level Set approach is used. In this context the interface is given by the zero level set of a continuous function ϕ. Here, the following convention is chosen: $\phi < 0$ in Ω_1 and $\phi > 0$ in Ω_2. Obviously we have $\Gamma(t) = \{\mathbf{x} \in \Gamma; \phi(\mathbf{x}, t) = 0\}, \forall t \geq 0$. As the interface moves at the flow velocity, the following transport equation for ϕ can be easly derived:

$$\partial_t \phi + \mathbf{u} \cdot \nabla \phi = 0. \tag{9}$$

Since $\mathbf{Y}$ is supposed regular at the interface and using the level set function we obtain formally the *global model* for the set of the primitive unknowns defined over the all domain by:

$$\mathscr{L}(\mathbf{Y}, \phi)\mathbf{Y} = \mathscr{S}(\mathbf{Y}, \phi), \text{ in } \Omega \times \mathbb{R}^{+*} \tag{10}$$

with

$$\mathscr{L}(\mathbf{Y}, \phi)\mathbf{Y} = \mathscr{A}_0(\mathbf{Y}, \phi)\partial_t \mathbf{Y} + \sum_{i=1}^{d} \mathscr{A}_i(\mathbf{Y}, \phi)\partial_{x_i}\mathbf{Y} - \sum_{i,j=1}^{d} \partial_{x_i}(\mathscr{K}_{ij}(\mathbf{Y}, \phi)\partial_{x_j}\mathbf{Y}).$$

In the previous global operator $\mathscr{L}$ the following matrices are introduced:

$$\begin{aligned}
\mathscr{A}_0(\mathbf{Y}, \phi) &= A_0(\mathbf{Y}, \boldsymbol{\chi}(\phi), \rho(\phi)), \\
\mathscr{A}_i(\mathbf{Y}, \phi) &= A_i(\mathbf{Y}, \boldsymbol{\chi}(\phi), \rho(\phi)), \\
\mathscr{K}_{ij}(\mathbf{Y}, \phi) &= K_{ij}(\mathbf{Y}, \mu(\phi), \kappa(\phi)), \\
\mathscr{S}(\mathbf{Y}, \phi) &= \mathbf{S}(\mathbf{Y}, \rho(\phi)),
\end{aligned}$$

which depend on the discontinuous physical coefficients:

$$\begin{aligned}
\boldsymbol{\chi}(\phi) &= (1 - H(\phi))\chi_1 + H(\phi)\chi_2, \\
\rho(\phi) &= (1 - H(\phi))\rho_1 + H(\phi)\rho_2, \\
\mu(\phi) &= (1 - H(\phi))\mu_1 + H(\phi)\mu_2, \\
\kappa(\phi) &= (1 - H(\phi))\kappa_1 + H(\phi)\kappa_2,
\end{aligned}$$

constructed using H is the Heaviside function:

$$H(\phi) = \begin{cases} 1 \text{ when } \phi > 0, \\ 0 \text{ when } \phi < 0. \end{cases} \text{ in } \Omega \tag{11}$$

The global model (10) has the advantage of describing globally the diphasic flow and allows us to develop a global numerical method detailed in the next part.

3 Numerical Method

In this section the numerical approach used in order to simulate compressible–incompressible diphasic flow is presented.

3.1 Algorithm

For the sake of simplicity, the numerical method is based on a decoupled algorithm. Hence, on each time intervalle $[t^n, t^{n+1}]$, (9) and (10) are solved separately:

1. *Flow step* (fixed interface)

$$\begin{aligned} \mathscr{L}(\mathbf{Y}, \phi^n) &= \mathscr{S}(\mathbf{Y}, \phi^n) \quad \text{in } \Omega \times]t^n; t^{n+1}[\\ \mathbf{Y}(\mathbf{x}, t^n) &= \mathbf{Y}^n(\mathbf{x}) \qquad \mathbf{x} \in \Omega \\ &\Rightarrow \mathbf{Y}^{n+1} \end{aligned} \tag{12}$$

2. *Interface step* (given flow)

$$\begin{aligned} \partial_t \phi + \mathbf{u}^{n+1} \cdot \nabla \phi &= 0 \qquad \text{in } \Omega \times]t^n; t^{n+1}[\\ \phi(\mathbf{x}, t^n) &= \phi^n(\mathbf{x}) \; \mathbf{x} \in \Omega \\ &\Rightarrow \phi^{n+1} \end{aligned} \tag{13}$$

The two subproblems (12) and (13) are discretized in space on the same mesh, with the stabilized finite element method for (12) and a Galerkin Discontinuous scheme for (13). The details of the numerical method for (13) are discussed in [6]. Here we focus on the resolution of the subproblem (12).

3.2 Global Finite Element Formulation Method

To solve Navier–Stokes compressible and incompressible equation (2), a finite element method is chosen. It is well known that in the context of same order finite

element for approximation the pressure, velocity and temperature, the standard Galerkin approach suffers from numerical instabilities. These pathologies are associated first to the violation of inf-sup condition for incompressible flows, and second to the dominating convection terms present in both compressible and incompressible Navier–Stokes equations. To overcome these difficulties, it is necessary to introduce stabilization term in the standard Galerkin method. Since it allows to treat these two problems, the Galerkin Least Square (GLS) [4] method is consider in this paper.

The idea of the proposed approach is to use the GLS method to solve the global equation (10). In order to introduce this method, let $\mathscr{T}_h$ be a triangulation of Ω into triangular or tetrahedral element e. Using these notations the following space-time stabilized formulation is proposed.

Find $\mathbf{Y}_h^{n+1} \in \mathbf{S}_h$ such that $\forall \mathbf{W}_h \in \mathbf{T}_h$:

$$\begin{aligned}
&\sum_{e\in\mathscr{T}_h} \int_e \mathbf{W}_h \cdot \left(\overline{\mathscr{A}_0}_e^n \frac{\delta \mathbf{Y}}{\delta t} + \sum_{i=1}^{d} \overline{\mathscr{A}_i}_e^n \partial_{x_i} \mathbf{Y}_h^{n+1} - \overline{\mathscr{S}}_e^n \right) \\
&+ \sum_{i,j=1}^{d} \partial_{x_i} \mathbf{W}_h \cdot \overline{\mathscr{K}_{ij}}_e^n \partial_{x_j} \mathbf{Y}_h^{n+1} d\Omega \\
&+ \sum_{e\in\mathscr{T}_h} \int_e {}^t\overline{\mathscr{L}}_e^n \cdot \overline{\tau}_e^n \left(\overline{\mathscr{L}}_e^n \mathbf{Y}_h^{n+1} - \overline{\mathscr{S}}_e^n \right) d\Omega = 0. \qquad (14)
\end{aligned}$$

The differences $\delta\mathbf{Y}$ and δt are respectively $\mathbf{Y}^{n+1} - \mathbf{Y}^n$ and $t^{n+1} - t^n$. We also introduce the approximated trial and weight functions spaces $\mathbf{S}_h$ and $\mathbf{T}_h$. The first integral is the Galerkin contribution in integrated by parts form. The last integral is the least-square term with $\overline{\tau}_e^n$ the stabilization coefficient matrice (see [3, 4]).

As the global operator $\mathscr{L}$ is non-constant and discontinous across the interface and as the interface does not conform with the computational mesh, it is necessary to treat specifically hybrid elements (those cut by the interface). In addition, we have to design the stabilization parameter in these elements.

To overcome these difficulties, we extand to compressible-incompressible the averaged strategy used for incompressible diphasic flows [8]. In the proposed approach some explicit averaged terms defined on each element are introduced such as $\overline{\mathscr{A}_0}_e^n, \overline{\mathscr{A}_i}_e^n, \overline{\mathscr{K}_{ij}}_e^n, \overline{\mathscr{S}}_e^n, \overline{\tau}_e^n$. These operators are expressed in term of the averages $\overline{\mathbf{Y}}_e^n, \overline{\rho}_e^n, \overline{\mathbf{Y}}_e^n, \overline{\mu}_e^n, \overline{\kappa}_e^n, \overline{\boldsymbol{\chi}}_e^n = (\overline{\alpha_p}_e^n, \overline{\beta_T}_e^n)$ by:

$$
\begin{aligned}
\overline{\mathscr{A}_0}_e^n &= A_0(\overline{\mathbf{Y}}_e^n, \overline{\boldsymbol{\chi}}_e^n, \overline{\rho}_e^n),\\
\overline{\mathscr{A}_i}_e^n &= A_i(\overline{\mathbf{Y}}_e^n, \overline{\boldsymbol{\chi}}_e^n, \overline{\rho}_e^n),\\
\overline{\mathscr{K}_{ij}}_e^n &= K_{ij}(\overline{\mathbf{Y}}_e^n, \overline{\mu}_e^n, \overline{\kappa}_e^n),\\
\overline{\tau}_e^n &= \tau(\overline{\mathbf{Y}}_e^n, \overline{\rho}_e^n, \overline{\mu}_e^n, \overline{\kappa}_e^n),\\
\overline{\mathscr{S}}_e^n &= \mathbf{S}(\overline{\mathbf{Y}}_e^n, \overline{\rho}_e^n),
\end{aligned}
$$

To elaborate such averages the interface is supposed to be a planar surface and we define a color function ε_e^n, that represent the gas proportion in each element e at time t^n, given by:

$$\varepsilon_e^n = \frac{V_2^n}{V_e}, \tag{15}$$

where V_2^n is the gas volume in the element e whose the volume is V_e. Using ε_e^n, we introduce:

- *Arithmetic averages* for the unknowns, the thermodynamic coefficients, and the stabilization matrice;

$$
\begin{aligned}
\overline{\mathbf{Y}}_e^n &= (1-\varepsilon_e^n)\mathbf{Y}_1 + \varepsilon_e^n \mathbf{Y}_2,\\
\overline{\boldsymbol{\chi}}_e^n &= (1-\varepsilon_e^n)\boldsymbol{\chi}_1 + \varepsilon_e^n \boldsymbol{\chi}_2,\\
\overline{\tau}_e^n &= (1-\varepsilon_e^n)\tau_1 + \varepsilon_e^n \tau_2,
\end{aligned}
$$

- *Harmonic averages* for the density, the viscosity, and the conductivity;

$$
\begin{aligned}
\overline{\rho}_e^n &= (1-\varepsilon_e^n)\rho_1 + \varepsilon_e^n \rho_2,\\
\overline{\mu}_e^n &= (1-\varepsilon_e^n)\mu_1 + \varepsilon_e^n \mu_2,\\
\overline{\kappa}_e^n &= (1-\varepsilon_e^n)\kappa_1 + \varepsilon_e^n \kappa_2.
\end{aligned}
$$

For the sake of simplicity the average have been presented here in one dimension of space, yet the idea remains the same in two or three dimensions.

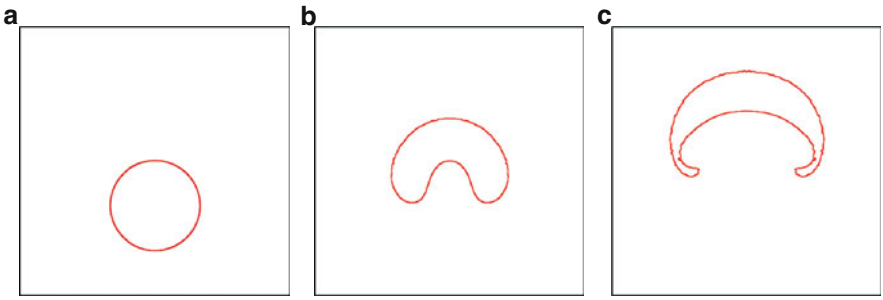


Fig. 1 Interface shape of a rising air bubble in water ($\rho_1/\rho_2 = 1000$, $\mu_1/\mu_2 = 10^5$, $Re = 100$)

4 Numerical Simulations

The proposed stategy is performed in both one and two spatial dimensions. We propose here a two-dimensional test-case. The dynamic of single air bubble in a box full of water.

We consider a square domain of lengh $L = 0.3\,\mathrm{m}$ in which a circular bubble of radius $R = 0.05\,\mathrm{m}$ is located in $(0.15\,\mathrm{m}, 0.15\,\mathrm{m})$. The flow is initially at rest at the pressure and is submitted to the gravity field $\mathbf{g} = 10\,\mathrm{m.s}^{-2}$.

The Fig. 1 shows the evolution of the bubble shape, for the ratios of density $\rho_1/\rho_2 = 1000$ and viscosity $\mu_1/\mu_2 = 10^5$. The subscripts 1 and 2 are associated to water and air respectively. The problem is characterized by the dimensionless Reynolds number which is $Re = (2R)^{3/2}\sqrt{g}\rho_1/\mu_1 = 100$. As we can see in the Fig. 1, the bubble rise in the water due to the effects of buoyancy and the interface is deforming. The simulation is performed for 14,000 times steps from $t = 0\mathrm{s}$ to $t = 0.44\,\mathrm{s}$ which shows the stability of the numerical method.

5 Conclusion

In this note, we have presented a global and simple numerical method dedicated to the prediction of compressible-incompressible diphasic flows. Using a global formulation of the Navier–Stokes equations for liquid-gas flows a numerical scheme have been developed based on a stabilized finite element method. To treat the discontinuities across the interface specific averages have been introduced. It appears that the proposed strategy gives good results for the numerical simulation of bubble dynamics.

References

1. Abgrall, R., Saurel, R.: A multiphase Godunov method for compressible multifluid and multiphase flows, Journal of Computational Physics, **140**, 425–467 (1999)
2. Caiden, R., Fedkiw, R.P., Anderson, C.: A numerical method for two phase flow consisting of separate compressible and incompressible regions, Journal of Computational Physics, **166**, 1–27 (2001)
3. Hauke, G.: Simple stabilizing matrices for the computation of compressible flows in primitives variables, Computer Methods in Applied Mechanics and Engineering, **190**, 6881–6893 (2001)
4. Hauke, G., Hughes, T.J.R.: A comparative study of different sets of variables for solving compressible and incompressible flows, Computer Methods in Applied Mechanics and Engineering, **153**, 1–44 (1998)
5. Marchandise, E., Remacle, J.-F.: A stabilized finite element method using a discontinuous level set approach for solving two phase incompressible flows, Journal of Computational Physics, **219**, 780–800 (2006)
6. Marchandise, E., Remacle, J.-F., Chevaugeon, N.: A quadrature-free discontinuous galerkin method for the level set equation. Journal of Computational Physics, **212**, 338–357 (2006)

7. Tanaka, N., Nakamura, N., and Nishimura, K.: Numerical analysis of boiling and condensing phenomena using gas-liquid unified algorithm, *oral talk*, ECCOMAS 2008
8. Turek, S.: FEM techniques for incompressible multiphase flows based on Level Set and phase field approach, *oral talk*, ENUMATH 2009

An Immersed Interface Technique for the Numerical Solution of the Heat Equation on a Moving Domain

François Bouchon and Gunther H. Peichl

Abstract A finite difference scheme for the heat equation with mixed boundary conditions on a moving domain is presented. We use an immersed interface technique to discretize the Neumann condition and the Shortley–Weller approximation for the Dirichlet condition. Monotonicity of the discretized parabolic operator is established. Numerical results illustrate the feasibility of the approach.

1 Introduction

Although moving boundary problems have a wide range of applications ([6, 8]) it is quite difficult to find numerical approaches which are supported by a theoretical analysis. In this paper we discuss a finite difference approximation to the heat equation on a doubly connected domain which moves and deforms with time according to a known dynamics. On one boundary component a Dirichlet condition, on the other a Neumann condition is specified. Immersed interface techniques appear to be well suited for this type of problems since they avoid the remeshing step that would be necessary at each time step if the mesh would follow the domain. They have recently been developed for the numerical treatment of partial differential equations with discontinuities in the coefficients ([11, 12]) or to cope with non rectangular domains ([9, 10]). A similar approach was pursued in [13] for a moving boundary problem. But in this paper the authors only allow a rigid translation of the domain. The moving boundary problem is converted to a sequence of fixed domain problems and information is substituted by an extrapolation technique at nodes which are

F. Bouchon (✉)
Laboratoire de Mathématiques, UMR CNRS 6620, Université Blaise-Pascal (Clermont-Ferrand 2), 63177 Aubière Cedex, France
e-mail: francois.bouchon@math.univ-bpclermont.fr

G.H. Peichl
Institute for Mathematics and Scientific Computing, Karl-Franzens-University Graz, A-8010 Graz, Austria
e-mail: gunther.peichl@uni-graz.at

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_18,

uncovered as the domain moves. A CFL- condition has to be imposed on all components of the boundary. The discretization is based on finite volume techniques. Second order convergence in space and time is reported.

Here we adapt a variant of the immersed interface technique originally constructed for an elliptic problem in [2] and then extended to the heat equation on a fixed domain in [3] for the moving boundary problem. The essential idea in [2] was to use a Shortley–Weller approximation of the Laplacian near the Dirichlet boundary and the standard five-point-stencil at any other interior node. This extends the solution one node across the Neumann boundary. The values at the exterior nodes are then determined by the Neumann condition. Therefore a "CFL"-like condition for the normal speed of the Neumann boundary is needed. Such a condition can be avoided for the Dirichlet boundary by a proper use of the Dirichlet condition. It is shown that the technique of [3] can be applied to show the monotonicity of the discrete parabolic operator. Numerical tests indicate a convergence rate of order $\tau + h^2$ which complies with the Euler-implicit scheme used to discretize the system.

2 Problem Formulation and Discretization

2.1 A Parabolic Problem on an Evolving Domain

We consider the following problem: Find the solution $u : D \equiv \cup_{t\in(0,T)} (\Omega(t) \times \{t\}) \to \mathbb{R}$ of the parabolic problem

$$\partial_t u(x,t) - \Delta u(x,t) = f(x,t) \qquad t \in (0,T), x \in \Omega(t), \tag{1}$$

$$u(x,0) = u_0(x) \qquad x \in \Omega(0), \tag{2}$$

$$u(x,t) = u_D(x,t) \qquad t \in (0,T), x \in \Gamma_D(t), \tag{3}$$

$$\partial_n u(x,t) = u_N(x,t) \qquad t \in (0,T), x \in \Gamma_N(t), \tag{4}$$

where the doubly connected bounded domain $\Omega(t) \subset \mathbb{R}^2$ depends continuously on t, as well as $\Gamma_D(t)$ and $\Gamma_N(t)$ which are the two disjoint parts of its boundary ($d(\Gamma_D(t), \Gamma_N(t)) > 0$, and $\Gamma_D(t) \cup \Gamma_N(t) = \partial\Omega(t)$). The data f, u_0, u_D, and u_N are such that the solution is twice continuously differentiable in time and four times in space with bounded derivatives.

Note that this problem is sometimes referred to in the literature (see [7]) as "parabolic problem on a non-cylindrical domain" since (1) must be satisfied for $(x,t) \in D \subset \mathbb{R}^2 \times \mathbb{R}$. The case where D would be cylindrical, $D = \Omega \times (0,T)$, corresponds to (1)–(4) with $\Omega(t) = \Omega$ for all $t \in [0,T]$. Results for existence, uniqueness and regularity of strong solutions can be found in [7] for parabolic problems defined on a non-cylindrical domain with Dirichlet boundary conditions, or parabolic problems on cylindrical domains with mixed boundary conditions.

We assume that $\Omega(t)$ satisfies the ball condition used in [2] and [3] for all $t \in [0, T]$:

There exists r_0 such that for all $t \in [0, T]$, for all $x \in \Gamma_N(t)$ one can find points $\xi_x \in \Omega(t)$ and $\eta_x \in \complement\bar{\Omega}(t)$ such that the balls $B(\xi_x, r_0)$ and $B(\eta_x, r_0)$ satisfy

$$B(\xi_x, r_0) \subset \Omega(t), \quad B(\eta_x, r_0) \subset \complement\bar{\Omega}(t), \quad \overline{B(\xi_x, r_0)} \cap \overline{B(\eta_x, r_0)} = \{x\}. \tag{5}$$

2.2 Derivation of the Scheme

In [2] a finite difference scheme was presented for the following elliptic problem:

$$-\Delta u(x) = f(x) \qquad x \in \Omega, \tag{6}$$

$$u(x) = u_D(x) \qquad x \in \Gamma_D, \tag{7}$$

$$\partial_n u(x) = u_N(x) \qquad x \in \Gamma_N. \tag{8}$$

The discretization of (6)–(8) resulted in the linear system

$$AU = F, \tag{9}$$

where U was the vector made of the unknowns $U_i \approx u(x_i)$ for $x_i \in \Omega_h \cup \Gamma_{N,h} \cup \Gamma_{D,h}$. Here Ω_h denotes the set of grid points interior to Ω away from the boundary Γ_D, $\Gamma_{D,h}$ denotes the set of grid points interior to Ω close to the boundary Γ_D and $\Gamma_{N,h}$ denotes the set of grid points which are exterior to Ω but close to the boundary Γ_N. A point is considered to be close to one boundary if at least one of its four neighbours is on the other side of the boundary. The vector F in (9) was a vector which depended on the data u_D, u_N and f.

The matrix A could be split in the following way:

$$A = \begin{pmatrix} A_1 & A_2 & \mathbb{O} \\ A_4 & A_5 & A_6 \\ \mathbb{O} & A_8 & A_9 \end{pmatrix}, \tag{10}$$

where the top block of lines corresponded to the discretization of (6) with the modifications due to the Shortley–Weller approximation (see [15]) for the points in $\Gamma_{D,h}$, the second block of lines corresponded to the discretization of (6) for the points in Ω_h, and the bottom block of lines corresponded to the discretization of (8) for the points in $\Gamma_{N,h}$, which was derived in such a way that the matrix A in (10) was an M-matrix. Thus, the second order convergence of the scheme could be proven. In [3] this method was adapted for the problem (1)–(4) for a fixed domain. The Euler-implicit scheme resulted in a linear system described by the matrix formulation:

$$(B + \tau^{-1} D)U^n = \tau^{-1} D U^{n-1} + F^n, \tag{11}$$

where U^n was the vector made of the unknowns $u_i^n \approx u(x_i, t^n)$, F^n was a vector which depended on the data u_D, u_N and f, D and B were matrices defined by

$$D = \begin{pmatrix} I & \mathbb{O} & \mathbb{O} \\ \mathbb{O} & I & \mathbb{O} \\ \mathbb{O} & \mathbb{O} & \mathbb{O} \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} B_1 & B_2 & \mathbb{O} \\ B_4 & B_5 & B_6 \\ \mathbb{O} & B_8 & B_9 \end{pmatrix}. \tag{12}$$

Here the first two blocks of lines corresponded to points in $\Omega_h \cup \Gamma_{D,h}$, and the bottom block of lines corresponded to the discretization of the Neumann Boundary condition for the points in $\Gamma_{N,h}$. The matrix B was shown to be monotone although it was not an L-matrix, and a perturbation technique similar to the work presented in [1] was applied to show that the matrix $B + \tau^{-1} D$ remained monotone. The convergence analysis showed that the error was bounded by

$$|e^n| \leq C \left(\tau + h^2/\tau\right), \tag{13}$$

where τ and h had to be small and of the same order of magnitude. The numerical tests showed the better convergence rates one in time and two in space.

The adaptation of the above scheme to a moving domain $\Omega(t)$ is complicated by the fact that nodes may enter or leave $\Omega(t)$ as it evolves. Thus the dimension of the vectors U^n as well as of the involved matrices depend on n. Since we are still going to use an Euler-implicit scheme the discretization of $-\Delta u(t^n, \cdot)$ and of the Neumann condition (4) will be described by a matrix B^n analogous to B in (12) of appropriate variable dimension.

In order to describe the discretization of the time derivative let Ω_h^n, $\Gamma_{D,h}^n$ and $\Gamma_{N,h}^n$ be the quantities for time t^n which correspond to Ω_h, $\Gamma_{D,h}$, respectively $\Gamma_{N,h}$ in the case of a stationary domain. We also denote $\Lambda_{D,h}^n = \{x_i \notin \Omega_h^n \text{ across } \Gamma_D(t^n)\}$. The usual forward difference approximation

$$\partial_t u(x_i, t^n) \approx \frac{1}{\tau}(u_i^n - u_i^{n-1})$$

can be used provided u_i^{n-1} has been calculated in the previous time step. This is the case for points $x_i \in \left(\Omega_h^n \cup \Gamma_{D,h}^n\right) \cap \left(\Omega_h^{n-1} \cup \Gamma_{D,h}^{n-1} \cup \Gamma_{N,h}^{n-1}\right)$.

For points which are close to the Neumann boundary $\Gamma_{N,h}^n$ it can be shown that the value of u_i^{n-1} will be available if the following "CFL"-like condition is satisfied:

$$v_{\max}\tau \leq \frac{\sqrt{2}h}{2}, \tag{14}$$

where $v_{\max}$ denotes the maximum modulus of the normal velocity v_n of the boundary $\Gamma_N(t)$.

Finally we turn to points $x_i \in \left(\Omega_h^n \cup \Gamma_{D,h}^n\right) \cap \Lambda_{D,h}^{n-1}$. Hence there are largest times $t_n - \alpha_i^n \tau$, $\alpha_i^n \in (0, 1)$, such that $x_i \in \Gamma_D(t_n - \alpha_i^n \tau)$. Using the Dirichlet

condition (3) we are lead to the following discretization of the time derivative

$$\partial_t u(x_i, t^n) \approx \frac{u_i^n - u_D(x_i, t^n - \alpha_i^n \tau)}{\alpha_i^n \tau}. \tag{15}$$

We emphasize that due to this use of the Dirichlet condition it is not necessary to impose a "CFL"-condition on $\Gamma_D(t)$ analogous to (14). Therefore the Dirichlet boundary is allowed to move faster than the Neumann boundary.

Let $\mathscr{D}^n = \{i \in \mathbb{N} : x_i \in \left(\Omega_h^n \cup \Gamma_{D,h}^n\right) \cap \Lambda_{D,h}^{n-1}\}$, $\mathscr{I}^n = \{i \in \mathbb{N} : x_i \in \left(\Omega_h^n \cup \Gamma_{D,h}^n\right) \cap \left(\Omega_h^{n-1} \cup \Gamma_{D,h}^{n-1} \cup \Gamma_{N,h}^{n-1}\right)\}$ and $\mathscr{E}^n = \{i \in \mathbb{N} : x_i \in \Gamma_{N,h}^n\}$. Note that $\mathscr{D}^n$ may be empty and that the sets $\mathscr{D}^n$, $\mathscr{I}^n$ and $\mathscr{E}^n$ are disjoint for a given value of n. With respect to this splitting of the nodes used for computation at time t^n we obtain the following discrete system

$$(B^n + \tau^{-1} D^n) U^n = \tau^{-1} \widetilde{D}^n \widetilde{U}^{n-1} + F^n, \tag{16}$$

with

$$B^n = \begin{pmatrix} B_1^n & B_2^n & \mathbb{O} \\ B_4^n & B_5^n & B_6^n \\ \mathbb{O} & B_8^n & B_9^n \end{pmatrix}, \quad D^n = \begin{pmatrix} D(\alpha)_{\mathscr{D}^n} & & \\ & I_{\mathscr{I}^n} & \\ & & \mathbb{O} \end{pmatrix}, \quad \widetilde{D}^n = \begin{pmatrix} \mathbb{O} & & \\ & I_{\mathscr{I}^n} & \\ & & \mathbb{O} \end{pmatrix}.$$

Above $D(\alpha)_{\mathscr{D}^n} = \operatorname{diag}(\frac{1}{\alpha_i^n} : i \in \mathscr{D}^n)$ and $I_{\mathscr{I}^n}$ denotes the identity matrix of dimension $|\mathscr{I}^n|$. The vector $\widetilde{U}^{n-1}$ on the right hand side of (16) takes into account that U^n and U^{n-1} in general have different dimensions. The vector F^n is determined by

$$F_i^n = \begin{cases} \frac{1}{\alpha_i^n \tau} u_D(x_i, t^n - \alpha_i^n \tau) + c_i^{SW} + f(x_i, t^n), & i \in \mathscr{D}^n, \\ c_i^{SW} + f(x_i, t^n), & i \in \mathscr{I}^n, \\ u_N(P_i, t^n) & i \in \mathscr{E}^n, \end{cases}$$

where c_i^{SW} is a contribution arising from the Shortley–Weller approximation and P_i denotes an appropriate point on $\Gamma_N(t^n)$. Note that the splitting of the system for the moving domain problem is done in a way slightly different from (12).

Let A^n be the matrix in (9) referring to the domain Ω_h^n. It was pointed out in [3] that A^n and B^n are related by

$$B^n = \mathscr{M}^n A^n \equiv \begin{pmatrix} I_{\mathscr{D}^n} & & \\ & I_{\mathscr{I}^n} & \\ & M^n & I_{\mathscr{E}^n} \end{pmatrix} A^n$$

with $M^n \leq 0$ and $M^n = \mathscr{O}(h)$. Hence $\mathscr{M}^n$ is monotone. Then

$$\begin{aligned} B^n + \tau^{-1} D^n &= \mathscr{M}^n (A^n + \tau^{-1} (\mathscr{M}^n)^{-1} D^n) \\ &= \mathscr{M}^n (A^n + \tau^{-1} \begin{pmatrix} D(\alpha)_{\mathscr{D}^n} & \mathbb{O} & \mathbb{O} \\ \mathbb{O} & I_{\mathscr{I}^n} & \mathbb{O} \\ \mathbb{O} & -M^n & \mathbb{O} \end{pmatrix}) \\ &= \mathscr{M}^n \left(\begin{pmatrix} A_1^n + \frac{1}{\tau} D(\alpha)_{\mathscr{D}^n} & A_2^n & \mathbb{O} \\ A_4^n & A_5^n & A_6^n \\ \mathbb{O} & A_8^n & A_9^n \end{pmatrix} + \tau^{-1} \begin{pmatrix} \mathbb{O} & \mathbb{O} & \mathbb{O} \\ \mathbb{O} & I_{\mathscr{I}^n} & \mathbb{O} \\ \mathbb{O} & -M^n & \mathbb{O} \end{pmatrix} \right) \\ &\equiv \mathscr{M}^n (\hat{A}^n + \tau^{-1} E^n). \end{aligned}$$

Theorem 1. *Let $\frac{h}{\tau}$ be bounded. Then $B^n + \tau^{-1} D^n$ is monotone for h and τ sufficiently small.*

Proof. The proof is identical to the one given in [3]: since A^n is an M-matrix and $\tau^{-1} D(\alpha)_{\mathscr{D}^n} \geq 0$ we conclude that $\hat{A}^n$ is monotone and $(\hat{A}^n)^{-1} \leq (A^n)^{-1}$. Therefore monotonicity of $B^n + \tau^{-1} D^n$ will follow once $\hat{A}^n + \tau^{-1} E^n$ is shown to be monotone. We now scale the matrix $\hat{A}^n + \tau^{-1} E^n$ according to

$$T^n \equiv \operatorname{diag}(hI_{\mathscr{D}^n}, hI_{\mathscr{I}^n}, I_{\mathscr{E}^n})(\hat{A}^n + \tau^{-1} E^n) \equiv Q^n + R^n. \tag{17}$$

Since the matrices $\hat{A}_i^n$ have entries of magnitude $\mathscr{O}(h^{-2})$ for $i = 1, \cdots 6$ and $\mathscr{O}(h^{-1})$ for $i = 8, 9$, and since the matrix M^n has entries of magnitude $\mathscr{O}(h)$, the matrices Q^n and R^n in (17) have entries of magnitude $\mathscr{O}(h^{-1})$ and $\mathscr{O}(1)$ respectively. Hence, since R^n is non negative, the perturbation T^n of the M-matrix Q^n in (17) can be shown to remain monotone using Theorem 2.5 in [1] for τ sufficiently small. □

Remark 1. We note that the possible unboundedness of the diagonal elements of $\hat{A}_1^n$ does not effect the proof of Theorem 1 since it does not use the nodes corresponding to indices in $\mathscr{D}^n$.

3 Numerical Results

Numerical tests have been run with the following data:

The Neumann and Dirichlet boundaries are given by

$$\begin{aligned} \Gamma_N(t) &= \left\{(x, y) \in [0, 1]^2, \psi_N(x, y, t) = 0\right\}, \\ \Gamma_D(t) &= \left\{(x, y) \in [0, 1]^2, \psi_D(x, y, t) = 0\right\}, \end{aligned}$$

where:

$$\begin{aligned} \psi_N(x, y, t) = 14.4 \times (x - (0.4 + 0.1\cos(3t)))^2 - 0.288 \times (3 - (1 - \cos(3t)))^2 \\ + (4 - \sin(10t)) \times (y - (0.4 + 0.2\sin(3t)))^2 \times (3 - (1 - \cos(3t)))^2, \end{aligned}$$

and

$$\psi_D(x,y,t) = 4 \times (x - (0.4 + 0.1\cos(3t)) - 0.1\sin(\min(16t, 3\pi)))^2 + (4 - \sin(10t)) \times (y - (0.4 + 0.2\sin(3t)))^2 - 0.04.$$

The boundary of the set $\Omega(t) = \{(x,y) \in [0,1]^2, \psi_N(x,y,t)\psi_D(x,y,t) < 0\}$ is given by $\Gamma_D(t) \cup \Gamma_N(t)$ (see Fig. 1), the Neumann boundary beeing the "exterior" boundary and the Dirichlet boundary is the boundary of the inner hole.

The data u_0, u_N, u_D and f have been chosen so that the exact solution of the continuous problem (1)–(4) is given by:

$$u(x,y,t) = (x^3 - 4xy^2 + 2y^4)\cos(t) + \sin(t) + \big(\cos(x) + \sin(y)\big)\exp(t^2). \tag{18}$$

The runs have been made with grid sizes $h = 1/I$, with $I \in \{200, 300, 400, 600, 800\}$, and with $\tau = T/N = 1/N$ with $N \in \{200, 400, 800, 1600, 3200, 6400\}$ and $N \geq I$ so that the "CFL"–condition (14) is satisfied.

Table 1 presents the local error between the computed and the exact solutions: $\max_{k,i,j} \left| u^k_{i,j} - u(x_{ij}, t_k) \right|$. This table shows an error of order $\mathcal{O}(\tau + h^2)$ since this error is (approximately) divided by 4 when h is divided by 2 and τ is divided by 4 (see for example the marked entries in Table 1). The observed convergence rates are the same as in [3].

The technique presented in [4,5] has been used to solve the linear system in (16). For the fixed domain case (see [3]), the preprocessing step requires $\mathcal{O}(n^3)$ floating points operations and needs only to be computed once. Here, since the domain evolves, this step must be computed at each time-step. The resulting asymptotic behavior of the CPU time required for the whole simulation is therefore $\mathcal{O}(h^{-3}\tau^{-1})$.

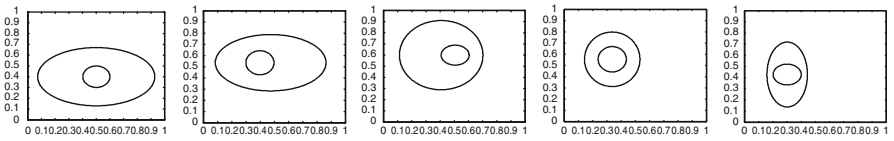


Fig. 1 Domain $\Omega(t)$ for $t = 0, 0.25, 0.5, 0.75, 1$

Table 1 Local error

h =	5×10^{-3}	3.33×10^{-3}	2.5×10^{-3}	1.67×10^{-3}	1.25×10^{-3}
$\tau = 5 \times 10^{-3}$	1.03×10^{-3}	*not cv*	*not cv*	*not cv*	*not cv*
$\tau = 2.5 \times 10^{-3}$	$\mathbf{6.27 \times 10^{-4}}$	5.06×10^{-4}	4.83×10^{-4}	*not cv*	*not cv*
$\tau = 1.25 \times 10^{-3}$	4.30×10^{-4}	3.01×10^{-4}	2.59×10^{-4}	2.43×10^{-4}	2.51×10^{-4}
$\tau = 6.25 \times 10^{-4}$	4.35×10^{-4}	2.00×10^{-4}	$\mathbf{1.56 \times 10^{-4}}$	1.27×10^{-4}	1.21×10^{-4}
$\tau = 3.125 \times 10^{-4}$	4.71×10^{-4}	1.95×10^{-4}	1.06×10^{-4}	7.49×10^{-5}	6.47×10^{-5}
$\tau = 1.5625 \times 10^{-4}$	4.89×10^{-4}	2.15×10^{-4}	1.08×10^{-4}	4.95×10^{-5}	$\mathbf{3.90 \times 10^{-5}}$

Table 2 CPU Time per time-step

h =	5×10^{-3}	3.33×10^{-3}	2.5×10^{-3}	1.67×10^{-3}	1.25×10^{-3}
CPU time *(sec.)*	5.40	41.18	75.38	195.29	550.63
CPU time/h^{-3}	6.75×10^{-7}	1.53×10^{-6}	1.18×10^{-6}	9.04×10^{-7}	1.08×10^{-6}

The average CPU time required for one time-step is reported in Table 2. The last line of Table 2 confirms the asymptotic operation count of order h^{-3} per time step.

4 Conclusion

The immersed interface technique of [2] and [3] has successfully been adapted for a parabolic problem on a moving domain. The second order convergence in space and first order in time has been observed on numerical tests, where a $\mathcal{O}(\tau^{-1}h^{-3})$ algorithm has been used. Further works will concentrate on similar problems on domains with less regularity (domains with corners) for which the ball condition (5) does not hold.

References

1. F. Bouchon, Monotonicity of some perturbations of irreducibly diagonally dominant M-matrices. *Numer. Math.*, Vol. 105 (2007), 591–601
2. F. Bouchon, G. H. Peichl, A second order immersed interface technique for an elliptic Neumann problem. *Numer. Meth. PDE*, Vol. 23 (2007), 400–420
3. F. Bouchon, G. H. Peichl, The immersed interface technique for parabolic problems with mixed boundary conditions. *SIAM J. Numer. Anal.*
4. B. L. Buzbee, F. W. Dorr, The direct solution of the Biharmonic equation on rectangular regions and the Poisson equation on irregular regions. *SIAM J. Numer. Anal.*, Vol. 11 (1974), 753–763
5. B. L. Buzbee, F. W. Dorr, J. A. George, G. H. Golub, The direct solution of the discrete Poisson equation on irregular regions. *SIAM J. Numer. Anal.*, Vol. 8 (1971), 722–736
6. J. Crank, Free and moving boundary problems. Oxford University Press, London (1984)
7. A. Friedman Partial differential equations of parabolic type, Prentice-Hall, NJ (1967)
8. R. Glowinski, J. P. Zolesio edts. Free and moving boundaries: Analysis, simulation and control, *Lecture Notes in Pure and Applied Mathematics, 252, Chapman & Hall CRC* (2007)
9. H. Johansen, P. Colella, A Cartesian grid embedded boundary method for Poisson's equation on irregular domains, *J. Comput. Phys.*, Vol. 147 (1998), 60–85
10. Z. Jomaa, C. Macaskill, The embedded finite difference method for the Poisson equation in a domain with an irregular boundary and Dirichlet boundary conditions, *J. Comp. Phys.*, Vol. 202 (2005), 488–506
11. R. J. LeVeque, Z. Li, The immersed interface method for elliptic equations with discontinuous coefficients and singular sources, *SIAM J. Numer. Anal.*, Vol. 31 (1994), 1019–1044
12. Z. Li, K. Ito, The immersed interface method: Numerical solutions of PDEs involving interfaces and irregular domains, SIAM Frontier Book Series, SIAM Publisher, Philadelphia, 2007
13. P. McCorquodale, P. Colella, H. Johansen, A Cartesian grid embedded boundary method for the heat equation on irregular domains. *J. Comp. Phys.*, Vol. 173, (2001), 620–635

14. J. A. Sethian Level set methods and fast marching methods: Evolving interfaces in computational geometry, fluid mechanics, computer vision and materials science, *Cambridge University Press*, 1999
15. G. H. Shortley, R. Weller, The numerical solution of Laplace's equation, *J. Appl. Phys.*, Vol. 9, (1938) 334–344

Lid-Driven-Cavity Simulations of Oldroyd-B Models Using Free-Energy-Dissipative Schemes

Sébastien Boyaval

Abstract In this work, we report on numerical tests in keeping with the study [Boyaval, Lelièvre, and Mangoubi, Free-energy-dissipative schemes for the Oldroyd-B model, *ESAIM: Mathematical Modelling and Numerical Analysis (M2AN)*, 43(3): 523–561, 2009], about Finite-Element discretizations of the Oldroyd-B system (for viscoelastic flows of some non-Newtonian fluids) which are stable in the sense of free-energy dissipation.

1 Introduction: Dissipative Oldroyd-B

During a period of time $[0, T)$ and in a physical domain $\Omega \subset \mathbb{R}^d$ ($d = 2$ or 3), the viscoelastic flow of some non-Newtonian fluids (like dilute polymeric fluids) can be described by a system coupling the incompressible Navier-Stokes equation for the velocity field $\mathbf{u} : (t, \boldsymbol{x}) \in [0, T) \times \Omega \to \mathbf{u}(t, \boldsymbol{x}) \in \mathbb{R}^d$ and the pressure field $p : (t, \boldsymbol{x}) \in (0, T) \times \Omega \to p(t, \boldsymbol{x}) \in \mathbb{R}$, with a *constitutive relation* for a tensor field $\boldsymbol{\sigma} : (t, \boldsymbol{x}) \in [0, T) \times \Omega \to \boldsymbol{\sigma}(t, \boldsymbol{x}) \in \mathbb{R}^{d\times d}_S$, where $\mathbb{R}^{d\times d}_S$ stands for the linear space of symmetric real square matrices.

Denoting by $\boldsymbol{I}$ the identity matrix, $\boldsymbol{f}$ a volume source term, and by Re, Wi and ε the usual positive dimensionless groups in rheology respectively known as the Reynolds number, the Weissenberg number and the elastic-to-viscous viscosity fraction (in fact $\varepsilon \in (0, 1)$), we now choose one widely used constitutive relation, the Oldroyd-B equation [2], hence the system:

$$\mathrm{Re}\left(\frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u}\cdot\nabla)\mathbf{u}\right) = -\nabla p + (1-\varepsilon)\Delta\mathbf{u} + \frac{\varepsilon}{\mathrm{Wi}}\mathbf{div}\boldsymbol{\sigma} + \boldsymbol{f} \tag{1}$$

$$\mathbf{div}\mathbf{u} = 0 \tag{2}$$

S. Boyaval
Laboratoire Saint-Venant, Université Paris-Est (Ecole des Ponts ParisTech)
6 & 8 Avenue Blaise Pascal, Cité Descartes, 77455 Marne-la-Vallée Cedex 2, France and MICMAC Project, INRIA, Domaine de Voluceau, BP. 105 – Rocquencourt, 78153 Le Chesnay Cedex, France
e-mail: sebastien.boyaval@inria.fr

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_19,

$$\mathrm{Wi}\left(\frac{\partial \boldsymbol{\sigma}}{\partial t} + (\mathbf{u}\cdot\nabla)\boldsymbol{\sigma} - (\nabla\mathbf{u})\boldsymbol{\sigma} - \boldsymbol{\sigma}(\nabla\mathbf{u})^T\right) + \boldsymbol{\sigma} = \boldsymbol{I}. \tag{3}$$

The numerical simulation of systems with constitutive relations like (1–2–3) is an active subject of research for decades [9]. Indeed, numerical instabilities are encountered with most usual discretizations, not only for complex geometries, but also in benchmark flows (used for comparisons with experimental results) like the flow past a condined cylinder or the 4:1 contraction. Here, we concentrate on the relaxation to equilibrium of flows in a square cavity $\Omega = (0,1)\times(0,1)$. We choose $\boldsymbol{f} = 0$, we supply (1–2–3) with homogeneous Dirichlet boundary conditions for the velocity field (no flow):

$$\mathbf{u} = 0 \text{ on } (0,T)\times\partial\Omega, \tag{4}$$

and we study the dissipative structure in time of the system (and its discretizations) for given initial conditions:

$$\mathbf{u}(0,\boldsymbol{x}) = \mathbf{u}^0(\boldsymbol{x}) \qquad \boldsymbol{\sigma}(0,\boldsymbol{x}) = \boldsymbol{\sigma}^0(\boldsymbol{x}) \in \mathbb{R}^{d\times d}_{SPD} \qquad \forall \boldsymbol{x}\in\Omega\,, \tag{5}$$

where $\mathbb{R}^{d\times d}_{SPD}$ is the set of symmetric positive definite matrices.

The choice of the Oldroyd-B model, a simple prototype for many differential constitutive relations, has inherent limitations for physicists and mathematicians. In particular, it cannot be used for long times in shear flows, and the existence results are not very developed yet. See [1,3] for a detailed discussion about known results. But it is enough for the purpose of our study: good discretizations of constitutive relations should mimick the dissipation of smooth solutions to the continuous system. This study is an important first step to understand the numerical instabilities observed in long-time numerical simulations, like blow-up or absence of convergence toward a stationary state beyond a limiting Weissenberg number, when the time and space discretization parameters are refined. Recall indeed that long-time simulations are often used to capture a stationary state: so, even when one is not sure that the latter exists or is unique, discretizations should at least not bring spurious energy to the system. In particular, we will try to use our discretizations (dissipative under no-flow boundary conditions) for the long-time simulations in a lid-driven cavity, a common benchmark flow that is believed to reach a stationary state [4].

2 Dissipative Discretizations of the Oldroyd-B System

The long-time dissipative structure of the Oldroyd-B system has been studied in [5, 6] where it is shown that the good positive quantity to consider for dissipation is a so-called free-energy (see also [10]):

$$F(\mathbf{u},\boldsymbol{\sigma}) := \frac{\mathrm{Re}}{2}\int_\Omega |\mathbf{u}|^2 + \frac{\varepsilon}{2\mathrm{Wi}}\int_\Omega \mathrm{tr}\left(\boldsymbol{\sigma} - \ln\boldsymbol{\sigma} - \boldsymbol{I}\right). \tag{6}$$

If $\boldsymbol{\sigma}^0 \in \mathbb{R}^{d\times d}_{SPD}$, then $\boldsymbol{\sigma} \in \mathbb{R}^{d\times d}_{SPD}$ holds at all times $t \geq 0$ and the matrix logarithm $\ln \boldsymbol{\sigma}$ is well-defined at all times $t \geq 0$. Furthermore, the free energy of smooth solutions $(\mathbf{u}, \boldsymbol{\sigma})$ to the homogenous Cauchy-Dirichlet problem introduced above in Sect. 1 decays to zero when $t \to \infty$. Hence, smooth solutions converge as $t \to \infty$ to the unique stationary state $(\mathbf{u}_\infty, \boldsymbol{\sigma}_\infty) = (0, \boldsymbol{I})$.

Unfortunately, it is difficult to propagate this observation for supposedly smooth solutions of the continuous system to the discrete level. Approximation spaces for $(\mathbf{u}, \boldsymbol{\sigma})$ are typically linear, whereas one needs to compute the logarithm $\ln \boldsymbol{\sigma}_h$ of the discrete tensor field $\boldsymbol{\sigma}_h$ approximating $\boldsymbol{\sigma}$ and to manipulate the inverse matrix $\boldsymbol{\sigma}_h^{-1}$ in order to obtain an estimation of the free-energy dissipation at the discrete level. (There is no reason why $\ln \boldsymbol{\sigma}_h$ and $\boldsymbol{\sigma}_h^{-1}$ should lie in the same linear space as the tensor field $\boldsymbol{\sigma}_h$, except if $\boldsymbol{\sigma}_h$ is piecewise constant, which we write $\boldsymbol{\sigma}_h \in \mathbb{P}_0$.) In [3], we have thus derived discretizations which satisfy a discrete free-energy dissipation using the Finite-Element (FE) method with $\boldsymbol{\sigma}_h \in \mathbb{P}_0$.

Given $\boldsymbol{\sigma}_h \in \mathbb{P}_0$, the main difficulties in deriving a discrete free-energy dissipation are linked to the discretization of the nonlinear terms $(\mathbf{u}\cdot\boldsymbol{\nabla})\boldsymbol{\sigma}$, $(\boldsymbol{\nabla}\mathbf{u})\boldsymbol{\sigma}$ and $\boldsymbol{\sigma}(\boldsymbol{\nabla}\mathbf{u})^T$ in (3). In [3], we suggest to discretize the time with a Backward-Euler scheme (hence $(\boldsymbol{\nabla}\mathbf{u})\boldsymbol{\sigma}$ and $\boldsymbol{\sigma}(\boldsymbol{\nabla}\mathbf{u})^T$ become implicit terms). And we suggest to treat the advective term $(\mathbf{u}\cdot\boldsymbol{\nabla})\boldsymbol{\sigma}$ either by a characteristic method or a Discontinuous Galerkin (DG) method. At this point, different choices are still possible for the FE spaces of the approximations $(\mathbf{u}_h, p_h)$ to $(\mathbf{u}, p)$. For each possible choice, taking care of the inf-sup compatibility condition due to the incompressibility constraint (2) (the famous Ladyshenskaya-Babuška-Brezzi condition), then we can show the existence and long-time stability of discrete solutions [1, 3].[1]

In the next Section, we numerically test some of the different choices encountered during the derivation of the free-energy-dissipative schemes.

Remark 1. In our schemes, it is essential that the field $\boldsymbol{\sigma}_h$ remains positive-definite to give a meaning to $\ln \boldsymbol{\sigma}_h$ and $\boldsymbol{\sigma}_h^{-1}$. (In turn, an upper-bound on the discrete free-energy implies that $\boldsymbol{\sigma}_h$ has remained positive-definite at all times.) But since our free-energy-dissipative schemes are implicit, in particular because of the term $(\boldsymbol{\nabla}\mathbf{u}_h)\boldsymbol{\sigma}_h + \boldsymbol{\sigma}_h(\boldsymbol{\nabla}\mathbf{u}_h)^T$, one can only compute *approximations* of the solution $\boldsymbol{\sigma}_h$ (for instance using a fixed-point strategy), which could turn pointwise negative for some $\boldsymbol{x} \in \Omega$ (see next Sect. 3). Now, in [3], we also studied discretizations of the system reformulated when using the variable $\ln \boldsymbol{\sigma}$ instead of $\boldsymbol{\sigma}$ like in [4]. Such formulations forcefully retain the positive-definiteness of the tensor field and the existence of discrete solutions is slightly easier to establish. Yet, there is little difference in the construction process of discetizations that are free-energy-dissipative under no-flow boundary conditions. As a matter of fact, the possibility of enhanced numerical stability in the numerical simulations of the Oldroyd-B system thanks to a preserved positive-definiteness was also studied by [7, 8], where other techniques

[1] Piecewise linear discretizations for $\boldsymbol{\sigma}_h$ – continuous $\mathbb{P}_1$ and discontinuous $\mathbb{P}_1 + \mathbb{P}_0$ – have also been studied in [1, 3], but then, coarser approximations of $\boldsymbol{\sigma}_h$ than piecewise linear are still needed in the nonlinear terms and decrease the order of the scheme.

preserving the positivity than the log are proposed. But these studies do not consider the correct dissipative structure of Oldroyd-B (that is, the dissipation of the free energy (6)), and in spite of the numerous previous efforts in the same direction, the link between the positivity-preservation and the numerical (in)stability does not seem to be fully understood yet (that is, explained by a precise analysis at the elementary discrete level, corroborated from numerical simulations). In the future, a precise analysis of the numerical instabilities should thus take into account the numerical treatment of the nonlinearity in conjunction with the positivity-constraint (as a pre-requisite) and the free-energy-dissipation criterium (for dissipative problems). The possibility of bifurcations should also still be studied (in [1], we had some existence results, but no uniqueness for the discrete system). Here, we limit ourselves to simple preliminary numerical observations, which could motivate such further studies.

3 Numerical Results

We show numerical results for two Cauchy-Dirichlet problems in a cavity $\Omega = (0,1)\times(0,1)$ ($d=2$), at very low Reynolds number $\mathrm{Re} = 10^{-7}$ (close to creeping flows), with $\varepsilon = 0.5$ (which avoids instabilities due to incompatibility of the discrete spaces for $\mathbf{u}_h$ and $\boldsymbol{\sigma}_h$ [9]). For six different Weissenberg numbers $\mathrm{Wi} = .8, 1, 1.2, \ldots, 2$, we consider either of the two following choices for the Dirichlet boundary condition on the lid $\Gamma = [0,1]\times\{1\}$ of the cavity:

$$\mathbf{u}_h(x,1,t) = \begin{cases} \left(8x^2(1-x)^2(1+\tanh\left(8(t-.5)\right)\right), & \forall (x,t)\in[0,1]\times[0,t_{max}) \\ 0, & \forall (x,t)\in[0,1]\times(t_{max},T) \end{cases}$$

with either $t_{max} = T/5$ (to observe the relaxation to equilibrium for $t > t_{max}$) or $t_{max} = T$ (to capture a hypothetic stationary state). Initial conditions at $t=0$ are the stationary state $\left(\mathbf{u}_h^0 \equiv 0, \boldsymbol{\sigma}_h^0 \equiv \boldsymbol{I}\right)$ and we choose $T = 10$.

The two problems are solved using FE schemes like those discussed in the previous section (see [3] for details), using a constant time step $\Delta t = 10^{-2}$ for the backward Euler time-discretization, and three different regular meshes, that are built with triangular (isosceles) elements of size bounded above by $h = 0.1$, 0.05 and 0.03. We test two different choices of FE spaces for $(\mathbf{u}_h, p_h, \boldsymbol{\sigma}_h)$ ($\mathbb{P}_2\times\mathbb{P}_1\times\mathbb{P}_0$ and $\mathbb{P}_2\times\mathbb{P}_0\times\mathbb{P}_0$, where $\mathbb{P}_2$ means continuous quadratics) and two different discretizations for the convective term of the constitutive equation (characteristic and DG method). The nonlinear (implicit) terms will be treated using *one* Picard iteration (hence linearized [9]): at n-th iteration, we compute the solution $(\mathbf{u}_h^{n+1}, p_h^{n+1}, \boldsymbol{\sigma}_h^{n+1})$ to the discrete scheme knowing the result $(\mathbf{u}_h^n, p_h^n, \boldsymbol{\sigma}_h^n)$ of the previous iteration,

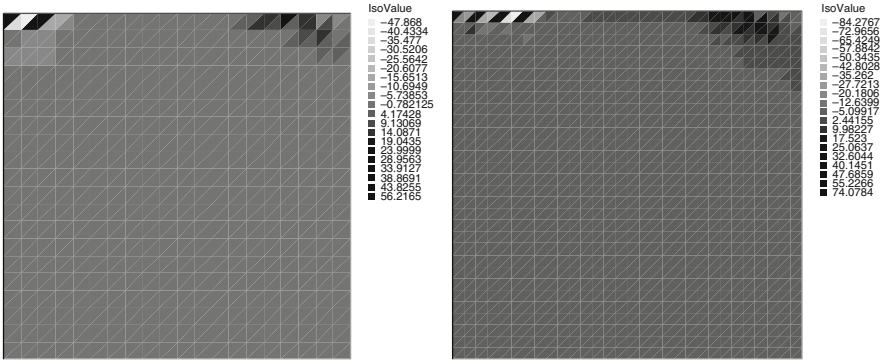


Fig. 1 Diagonal component of the tensor field at $t = 3$ for $t_{\max} = T$ and two different meshes (*left*: $h = 0.05$; and *right*: $h = 0.03$)

e.g. after replacing $(\nabla \mathbf{u}_h^{n+1})\boldsymbol{\sigma}_h^{n+1}$ by $(\nabla \mathbf{u}_h^{n})\boldsymbol{\sigma}_h^{n+1}$ in our scheme[2]). The resulting large linear system is solved using the GMRES method.

All computations are done with FreeFem++ (v2.240002). Two typical results for the diagonal component of the tensor field at $t = 3$ when $t_{max} = T$ are shown in Fig. 1 (DG method, $\mathbb{P}_2 \times \mathbb{P}_0 \times \mathbb{P}_0$). The difference between the two refinements in the area close to the right-top corner is an indication about the origin of possible instabilities (possibly a local lack of regularity of the solutions). And in fact, when the free-energy-dissipation and the positivity of $\boldsymbol{\sigma}_h$ are lost during the simulation, it is exactly in this singular area that a pointwise value of the tensor field first becomes negative, and then propagates through Ω in the cases where there is blow-up.

3.1 First Discretization: $(\mathbf{u}_h, p_h, \boldsymbol{\sigma}_h) \in \mathbb{P}_2 \times \mathbb{P}_1 \times \mathbb{P}_0$

In [3], we were not able to show a free-energy dissipation with this choice of FE spaces.[3] But first, we *numerically* check the relaxation to equilibrium for $t_{\max} = 2$ and nevertheless observe a free-energy-dissipation after $t_{\max} = 2$ until $T = 10$, see Fig. 2. (Yet, the characteristic method fails when $h = 0.03$, for $\boldsymbol{\sigma}_h$ becomes negative before $t_{max} = 2$ and the solutions blows up a few iterations later, whatever the Weissenberg number Wi used here.[4])

[2] This brute-force linearization seemed to yield correct approximations at small times after the start-up of the flow, but it may also have long-time ununderstood consequences explaining some of the instabilities in the results that we show here, see Remark 1.

[3] More precisely, this was possible only when the velocity field $\mathbf{u}_h$ used to advect $\boldsymbol{\sigma}_h$ was projected or interpolated in a consistent and weakly-incompressible manner.

[4] With the DG approach and a finer mesh $h = 0.025$, the positivity of $\boldsymbol{\sigma}_h$ is also lost before $t_{max} = 2$, whatever Wi, but no blow-up follows: the computations run until final time T.

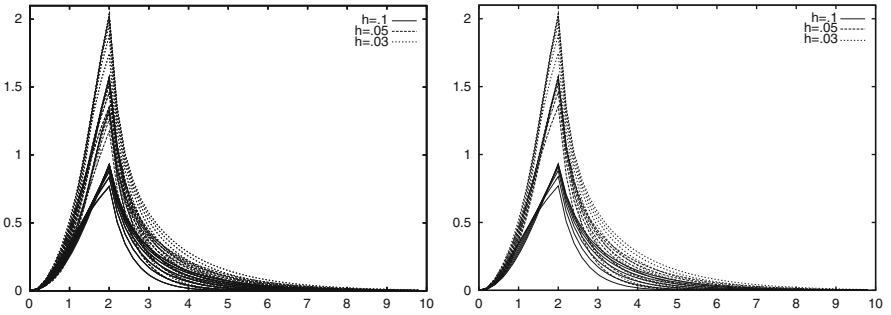


Fig. 2 Free energy with respect to time using the characteristic (*left*)/DG method (*right*) with $\mathbb{P}_1$ approximations of the pressure. We use $t_{\max} = 2$, three meshes and $\mathrm{Wi} \in [.8, 2]$. Note: for $h = 0.03$, the characteristic method fails

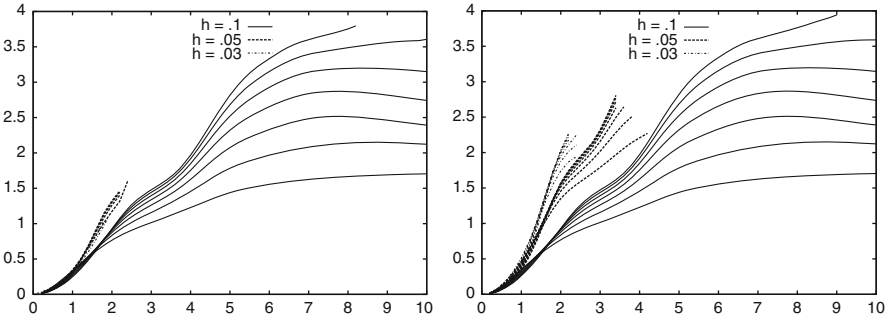


Fig. 3 Free energy with respect to time when $t_{\max} = T$, using the characteristic (*left*)/DG method (*right*) with $\mathbb{P}_1$ approximations of the pressure. When the curve stops, positivity has been lost (*left*: the computations then blow up). We use three different meshes $h = 0.1, 0.05, 0.03$ and $\mathrm{Wi} \in [.8, 2]$. Note: for $h = 0.03$, characteristics fail

Then, we try to capture a stationary state using $t_{\max} = T$, see Fig. 3. The *characteristic method* converges to a stationary state after $t = 2$ only for the coarser mesh $h = 0.1$. And its convergence rate is all the slower as Wi increases (the difference between two successive iterations decreases more slowly when Wi increases – a so-called "high-Weissenberg number problem" [9] ?). For finer meshes, the positivity is lost soon after $t = 2$, all the earlier as Wi increases. On the contrary, the *DG method* always converges to a stationary state. Yet, the positivity of $\boldsymbol{\sigma}_h$ is also lost for meshes finer than $h = 0.1$, soon after $t = 2$ (all the sooner as Wi increases and h decreases, but later than the characteristic method, and never followed by blow-up).

3.2 *Second Discretization:* $(\mathbf{u}_h, p_h, \boldsymbol{\sigma}_h) \in \mathbb{P}_2 \times \mathbb{P}_0 \times \mathbb{P}_0$

Compared to the previous section, we use here a lower-order FE space for the pressure field, and we can show that the DG approach is free-energy-dissipative (under no flow boundary conditions). For $t_{\max} = 2$, we similarly observe that all our

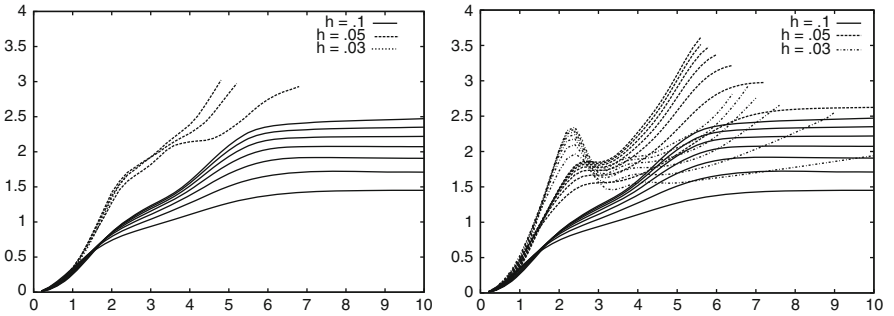


Fig. 4 Free energy with respect to time when $t_{\max} = T$, using the characteristic (*left*) / DG method (*right*) with $\mathbb{P}_0$ approximations of the pressure. When the curve stops, positivity has been lost (*left*: the computations then blow up). We use three different meshes $h = 0.1, 0.05, 0.03$ and $\mathrm{Wi} \in [.8, 2]$

schemes numerically dissipate the free-energy after $t_{\max}$ (under no flow boundary conditions). Notice yet that the positivity is never lost before $t_{\max} = 2$ here. (We do not show the results, which are very similar to those in Fig. 2 of Sect. 3.1.)

Second, we try to capture a stationary state using $t_{\max} = T$. The *characteristic method* reaches a stationary state only for the coarser mesh $h = 0.1$, all the slower as Wi increases (the convergence rate of the difference between two succesive iterations goes slower to zero). For finer meshes, no convergence to a stationary state could be observed because the positivity of $\boldsymbol{\sigma}_h$ is lost soon after $t = 2$, (all the sooner as Wi increases, and then the computations stop: there is blow up). On the contrary, the *DG method* seems to converge to a stationary state for all meshes (though again all the slower as Wi increases, and h decreases). The conformation tensor also loses its positivity after $t = 2$ (all the sooner as $\mathrm{Wi} \in [1, 2]$ increases and h decreases) for the two finer meshes, again a bit later than the characteristic method (see Fig. 3). Remarkably, for $\mathrm{Wi} = 0.8$ and $h = 0.05$ or $h = 0.03$, the positivity was never lost! (Thus, for $\mathrm{Wi} = 0.8$, we could observe that convergences of the free energy and the kinetic energy were not monotone with h.)

Finally, the only scheme here that is free-energy-dissipative under no-flow boundary conditions (that is, DG in Sect. 3.2) seems to behave better than the other ones for the lid-driven-cavity problem, although little difference is observed during pure relaxation to equilibrium (under no-flow boundary conditions). But there are still numerical instabilities, which look like a *high-Weissenberg number problem* [3, 9] with unclear origin (many different numerical problems could be mixed together). The next step in a rigorous study could be to ensure positivity while approximating correctly the nonlinear terms in a free-energy-dissipative scheme.

References

1. J. W. Barrett and S. Boyaval. Existence and approximation of a (regularized) Oldroyd-B model. *(preprint submitted for publication http://fr.arxiv.org/abs/0907.4066)*, 2009
2. R. B. Bird, C. F. Curtiss, R. C. Armstrong, and O. Hassager. *Dynamics of polymeric liquids*, volume 1: Fluid Mechanics. Wiley, New York, 1987

3. S. Boyaval, T. Lelièvre, and C. Mangoubi. Free-energy-dissipative schemes for the Oldroyd-B model. *ESAIM: Mathematical Modelling and Numerical Analysis*, 43(3):523–561, may 2009
4. R. Fattal and R. Kupferman. Time-dependent simulation of visco-elastic flows at high weissenberg number using the log-conformation representation. *J. Non-Newtonian Fluid Mech.*, 126:23–27, 2005
5. D. Hu and T. Lelièvre. New entropy estimates for the Oldroyd-B model, and related models. *Commun. Math. Sci.*, 5(4):906–916, 2007
6. B. Jourdain, C. Le Bris, T. Lelièvre, and F. Otto. Long-time asymptotics of a multiscale model for polymeric fluid flows. *Archive for Rational Mechanics and Analysis*, 181(1):97–148, 2006
7. Y. J. Lee and J. Xu. New formulations, positivity preserving discretizations and stability analysis for non-Newtonian flow models. *Comput. Methods Appl. Mech. Engrg.*, 195:1180–1206, 2006
8. A. Lozinski and R. G. Owens. An energy estimate for the Oldroyd-B model: theory and applications. *J. Non-Newtonian Fluid Mech.*, 112:161–176, 2003
9. R. G. Owens and T. N. Philips. *Computational rheology*. Imperial College Press, 2002
10. P. Wapperom and M. A. Hulsen. Thermodynamics of viscoelastic fluids: the temperature equation. *J. Rheol.*, 42(5):999–1019, 1998

Adaptive Multiresolution Simulation of Waves in Electrocardiology

Raimund Bürger and Ricardo Ruiz-Baier

Abstract A new fully adaptive multiresolution method is applied for the simulation of the complex dynamics of waves in excitable media in electrocardiology, where the membrane kinetics are given by the Aliev–Panfilov or Luo–Rudy II models. Numerical experiments show that the automatical adaptation strategy tracks the spatio-temporal pattern accurately at a substantially reduced computational cost if compared with fine-grid simulations. The nonlinear dynamics of complex multiscale patterns can thus be computed efficiently, also in the chaotic and turbulent regime which are currently beyond the frontiers of methods using regular discretizations.

1 Introduction

Nonlinear reaction-diffusion systems are widely used models of excitable chemical and biological media that usually exhibit rich spatio-temporal multiscale dynamics. Even in homogeneous media, nontrivial spatial structures (pulses, fronts, spiral waves and others) can emerge, and an impulse over a certain threshold initiates a wave of activity moving across the excitable medium. One of the most studied applications of such waves is the propagation of electrical activity in cardiac tissue. This phenomenon involves the interaction of different ion species across a combination of active and passive ion channels and diffusion of charge through a heterogeneous substrate with dynamically changing conductances.

It is the purpose of this contribution to provide further support for the adaptive multiresolution method for excitable media described in [4] by two new, original

R. Bürger (✉)
CI^{2}MA and Departamento de Ingeniería Matemática, Universidad de Concepción, Casilla 160-C, Concepción, Chile
e-mail: rburger@ing-mat.udec.cl

R. Ruiz-Baier
Modeling and Scientific Computing IACS-CMCS, École Polytechnique Fédérale de Lausanne EPFL, Station 8, CH-1015 Lausanne, Switzerland
e-mail: ricardo.ruiz@epfl.ch

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_20,

numerical examples for the Aliev–Panfilov (AP) and Luo–Rudy II (LRII) models in electrocardiology. The new feature of the example for the AP model is an implanted obstacle, while the LRII model is remarkably more involved than the models used in [4] since it includes a vector of seven gating variables, not just a scalar one.

We first consider a spatially two-dimensional model of waves in excitable media given by a reaction-diffusion system of the generic form

$$\partial_t u = \Delta A(u) + f(u,v), \quad \partial_t v = g(u,v), \quad (x,t) \in Q_T := \Omega \times [0,T], \tag{1}$$

where $\Omega \subset \mathbb{R}^2$ is an open, bounded, connected polygonal domain with boundary $\partial\Omega$, along with zero-flux boundary and initial conditions. The unknowns are the excitation and recovery variables u and v, which vary on fast and slow time scales, respectively. The functions f and g express the local reaction kinetics and A is a diffusion coefficient to be defined later.

The model by Aliev and Panfilov for propagation in cardiac tissue [1] is employed as the first of two specific examples. It consists in (1) along with

$$\begin{aligned} f(u,v) &= \{-C_1 u \text{ if } u < \rho_1,\ C_2 u + a \text{ if } u \in [\rho_1,\rho_2],\ C_3(1-u) \text{ if } u > \rho_2\} - v, \\ g(u,v) &= (ku - v) \cdot \{\eta_1 \text{ if } u < \rho_2,\ \eta_2 \text{ if } u > \rho_2,\ \eta_3 \text{ if } u < \rho_1 \text{ and } v < v_1\}, \end{aligned} \tag{2}$$

where $C_1, C_2, C_3, \eta_1, \eta_2, \eta_3, v_1$ and k are certain dimensionless parameters. The AP model (2) models the electrical activity in ventricular tissue more accurately than the well-known FitzHugh–Nagumo model (see [7]).

The second example is the LRII model [8] coupled with a monodomain description of the electrical wave propagation. It has the general form

$$\partial_t u = D\Delta u + f(u,\mathbf{v}) + I_{\text{ext}}(x,t), \quad \partial_t \mathbf{v} = g(u,\mathbf{v}), \quad (x,t) \in Q_T := \Omega \times [0,T], \tag{3}$$

where u is the membrane potential, $\mathbf{v} = (K_1, X, h, j, m, f, d)^{\mathrm{T}}$ is the vector of dimensionless ion-channel gating variables, and the total ionic current density

$$f(u,\mathbf{v}) = I_{\text{Na}}(u,\mathbf{v}) + I_{\text{si}}(u,\mathbf{v}) + I_{\text{K}}(u,\mathbf{v}) + I_{\text{K}_1}(u,\mathbf{v}) + I_{\text{K}_\text{p}}(u) + I_{\text{b}}(u)$$

is the sum of a fast inward sodium current I_{Na}, a slow inward current I_{si}, a time-dependent potassium slow outward current I_{K}, an outward potassium current I_{K_1}, a plateau potassium current I_{K_p}, and a total background current I_{b}:

$$\begin{aligned} I_{\text{Na}} &= G_{\text{Na}} m^3 h j (u - E_{\text{Na}}), & I_{\text{si}} &= G_{\text{si}} d f (u - E_{\text{si}}), & I_{\text{K}} &= G_{\text{K}} X X_i (u - E_{\text{K}}), \\ I_{\text{K}_1} &= G_{\text{K}_1} K_{1\infty} (u - E_{\text{K}_1}), & I_{\text{K}_\text{p}} &= G_{\text{K}_\text{p}} K_{\text{p}} (u - E_{\text{K}_\text{p}}), & I_{\text{b}} &= 0.03921(u + 59.87) \end{aligned}$$

with $G_{\text{Na}} = 23, G_{\text{si}} = 0.07, G_{\text{K}} = 0.705, G_{\text{K}_1} = 0.604, G_{\text{K}_\text{p}} = 0.0183 (\text{in mS cm}^{-2})$, $E_{\text{Na}} = 54.4$, $E_{\text{K}} = -77$, $E_{\text{K}_1} = -87.26$, $E_{\text{K}_\text{p}} = -87.26$, $E_b = -59.87$ (in mV). In addition, $E_{\text{si}} = 7.7 - 13.0287 \ln[\text{Ca}]_+$. The calcium ionic concentration satisfies

the Nernst equilibrium $d_t[\mathrm{Ca}]_+ = -10^{-4} I_{\mathrm{si}} + G_{\mathrm{si}}(10^{-4} - [\mathrm{Ca}]_+)$, and all gate variables $\rho = h, j, m, d, f, X, K_1$ evolve according to $d_t \rho = \alpha_\rho(u)(1-\rho) - \beta_\rho(u)\rho$, which precisely corresponds to the second equation in (3). Here, $\alpha_\rho(u)$ and $\beta_\rho(u)$ define the opening and closure rate of the gates, which are given by $\alpha_h = \alpha_j = 0$ for $u \geq -40\,\mathrm{mV}$, $\alpha_h = 0.135\mathrm{e}^{-0.147(u+80)}$ for $u < -40\,\mathrm{mV}$, and

$$\beta_h = \begin{cases} 3.56\mathrm{e}^{0.079u} + 3.1 \times 10^5 \mathrm{e}^{0.35u} & \text{for } u < -40\,\mathrm{mV}, \\ (0.13 + 0.13\mathrm{e}^{-0.09(u+10.66)})^{-1} & \text{otherwise}, \end{cases}$$

$$\alpha_j = (u + 37.8)\frac{\mathrm{e}^{0.2} + 2.7 \times 10^{-10}\mathrm{e}^{-0.04u}}{-7.87 \times 10^{-6}(1 + \mathrm{e}^{0.3(u+79.2)})} \quad \text{for } u < -40\,\mathrm{mV},$$

$$\beta_j = \begin{cases} 0.1212\mathrm{e}^{-0.01052u}(1 + \mathrm{e}^{-0.1378(u+40.14)})^{-1} & \text{for } u < -40\,\mathrm{mV}, \\ 0.3\mathrm{e}^{-2.535\times 10^{-7}u}(1 + \mathrm{e}^{-0.1(u+32)})^{-1} & \text{otherwise}, \end{cases}$$

$$\alpha_{K_1} = \frac{1.2}{1 + \mathrm{e}^{0.2385(u - E_{K_1} - 59.215)}}, \quad \alpha_m = \frac{0.32(u + 47.13)}{1 - \mathrm{e}^{-0.1(u+47.13)}}, \quad \beta_m = 0.08\mathrm{e}^{-0.0909u},$$

$$\alpha_d = \frac{0.095\mathrm{e}^{-0.01(u-5)}}{1 + \mathrm{e}^{-0.072(u-5)}}, \quad \beta_d = \frac{0.07\mathrm{e}^{-0.02(u+44)}}{1 + \mathrm{e}^{0.05(u+44)}}, \quad \alpha_f = \frac{0.012\mathrm{e}^{-0.008(u+28)}}{1 + \mathrm{e}^{0.15(u+28)}},$$

$$\beta_f = \frac{0.0065\mathrm{e}^{-0.02(u+30)}}{1 + \mathrm{e}^{-0.2(u+30)}}, \quad \alpha_X = \frac{0.0005\mathrm{e}^{0.083(u+50)}}{1 + \mathrm{e}^{0.057(u+50)}}, \quad \beta_X = \frac{0.0013\mathrm{e}^{-0.06(u+20)}}{1 + \mathrm{e}^{-0.04(u+20)}},$$

$$\beta_{K_1} = \frac{0.4912\mathrm{e}^{0.08(u - E_{K_1} + 5.476)}}{1 + \mathrm{e}^{-0.5143(u - E_{K_1} + 4.75)}} + \mathrm{e}^{0.0618(u - E_{K_1} - 594.31)}.$$

The gating variables X_i, K_p are assumed to rapidly reach a steady state, and therefore to depend only on the potential u. We set $X_i(u) = 1$ for $u \leq -100\,\mathrm{mV}$ and

$$X_i = (2.837\mathrm{e}^{0.04(u+77)} - 1)((u + 77)\mathrm{e}^{0.04(u+35)})^{-1} \quad \text{for } u > -100\,\mathrm{mV},$$
$$K_p = \left(1 + \mathrm{e}^{0.1672(7.488-u)}\right)^{-1}.$$

For overviews on multiresolution techniques for related problems, see [5, 9, 10]; references to other techniques to solve the system (1) are given in [4].

The remainder of the paper is organized as follows. In Sect. 2 we recall the numerical method for solving (1) on uniform fine meshes. This method is a classical finite volume (FV) scheme with a first-order Euler time discretization, and plays the role of a *reference numerical scheme*, i.e., it approximates the solution of (1) on a uniform mesh. In Sect. 3 we outline the MR procedure, which allows to construct space adaptive schemes based on the reference method (for details, see [4, 10]). The numerical results are presented in Sect. 4.

2 Reference Numerical Scheme

We consider a standard admissible mesh for $\Omega \subset \mathbb{R}^2$ formed by a family $\mathscr{T}$ of control volumes K of maximum diameter h and a family of points $(x_K)_{K\in\mathscr{T}}$, where x_K is the center of K. We let $N(K)$ denote the set of neighbors of K which share a common edge with K. Here $\mathscr{E}_{\text{int}}(K)$ is the set of edges of K in the interior of Ω and $\mathscr{E}_{\text{ext}}(K)$ the set of edges of K lying on the boundary $\partial\Omega$. For all $L \in N(K)$, $d(K,L)$ denotes the distance between x_K and x_L, and we denote by $\sigma = K|L$ ($\sigma = K|\partial\Omega$, respectively) the interface between K and L (between K and $\partial\Omega$, respectively). Moreover, $|K|$ stands for the two-dimensional measure of K and $|\sigma|$ for the one-dimensional measure of σ. Numerical fluxes on all edges σ are defined as by $F_{K,\sigma} = \tau_\sigma(u_L - u_K)$ for $\sigma = K|L \in \mathscr{E}_{\text{int}}(K)$ and $F_{K,\sigma} = 0$ for $\sigma \in \mathscr{E}_{\text{ext}}(K)$, which includes the zero-flux boundary conditions and where the transmissibility coefficients τ_σ are defined by $\tau_\sigma := |\sigma|/|d(K,L)|$ for $\sigma = K|L \in \mathscr{E}_{\text{int}}(K)$. We set $t^n := n\Delta t$ for $n = 0,\dots,N = \lceil T/\Delta t\rceil$. We define $f_K^n := f(u_K^n, v_K^n)$, $g_K^n := g(u_K^n, v_K^n)$, $u_K^0 := |K|^{-1}\int_K u_0(x)\mathrm{d}x$ and $v_K^0 := |K|^{-1}\int_K v_0(x)\mathrm{d}x$. To advance the numerical solution from t^n to $t^{n+1} = t^n + \Delta t$, we use the following finite volume scheme: Given u_K^n, v_K^n for all $K \in \mathscr{T}$, determine u_K^{n+1} and v_K^{n+1} from

$$|K|\Delta t^{-1}\big(u_K^{n+1} - u_K^n\big) + \textstyle\sum_{\sigma\in\mathscr{E}_{\text{int}}(K)\cup\mathscr{E}_{\text{ext}}(K)}\tau_\sigma\left(A\big(u_L^{n+1}\big) - A\big(u_K^{n+1}\big)\right) = |K| f_K^n, \tag{4}$$

$$\Delta t^{-1}\big(v_K^{n+1} - v_K^n\big) = g_K^n \ \text{ for all } K \in \mathscr{T}. \tag{5}$$

A CFL stability condition for the scheme (4) and (5) is given by

$$\Delta t h^{-1} \max_{K\in\mathscr{T},\, t^n<T}\left(\big|f_{u,K}^n\big| + \big|f_{v,K}^n\big| + \big|g_{u,K}^n\big| + \big|g_{v,K}^n\big|\right) + 4D\Delta t h^{-3} \le 1. \tag{6}$$

The resulting FV scheme produces a unique numerical solution. Solutions converge to a weak solution of (1) as the discretization parameters tend to zero [6].

3 Adaptivity: Multiresolution Framework

To be concise, we only consider Cartesian meshes on $\overline{\Omega} = [0,1]^2$, but the MR analysis could be carried out for more general meshes (see, e.g., [9]). We start by determining a nested mesh hierarchy $\mathscr{T}_0 \subset \cdots \subset \mathscr{T}_H$, using a partition of Ω. Each grid $\mathscr{T}_l$ is formed by the control volumes K^l on each level $l = 0,\dots,H$, where $l = 0$ corresponds to the coarsest and $l = H$ to the finest level. The *refinement sets* are defined by $\mathscr{M}_{K^l} := \{L_i^{l+1}\}_i$, where $\overline{K^l} := L_1^{l+1} \cup \cdots \cup L_{m_l}^{l+1}$, $m_l := \#\mathscr{M}_{K^l}$, where L_i^{l+1} is a control volume at level $l+1$, $L_i^{l+1} \subset K^l$. For $x \in K^l$ the *scale box function* is defined as $\tilde{\varphi}_{K^l}(x) := |K^l|^{-1}\chi_{K^l}(x)$, and therefore

the average of any function $u(\cdot,t) \in L^1(\Omega)$ on K^l can be expressed as the inner product $u_{K^l} := \langle u, \tilde{\varphi}_{K^l} \rangle_{L^1(\Omega)}$.

Cell averages and box functions satisfy the two-level relation

$$u_{K^l} = \sum_{L_i^{l+1} \in \mathscr{M}_{K^l}} |L_i^{l+1}||K^l|^{-1} u_{L_i^{l+1}}, \quad \tilde{\varphi}_{K^l} = \sum_{L_i^{l+1} \in \mathscr{M}_{K^l}} |L_i^{l+1}||K^l|^{-1} \tilde{\varphi}_{L_i^{l+1}}, \tag{7}$$

which defines a projection operator needed to move from finer to coarser levels. There is a transformation between the cell averages on level $l = H$ and the cell averages on level $l = 0$ plus a series of detail coefficients. This relation defines a prediction operator needed to move from coarser to finer resolution levels. A polynomial prediction is chosen, which in the particular case of Cartesian meshes is defined by $\tilde{u}_{L_i,l+1} = u_{L,l} - Q_x - Q_y + Q_{xy}$ for $i = 1, \ldots, \#\mathscr{M}_{K^l}$, where Q_x, Q_y and Q_{xy} are standard polynomial interpolators applied to the neighbors and diagonal neighbors of the control volume L^l, see [2–4, 10].

The error induced by the prediction operator at the cell K^l is defined as the difference between the cell average and the predicted value, i.e., $d_{K^l} := u_{K^l} - \tilde{u}_{K^l}$, and we may also write $d_{K^l,j} := \langle u, \tilde{\psi}_{K^l,j} \rangle$ for $j = 1, \ldots, \#\mathscr{M}_{K^l}$. For a multicomponent solution (u, v), on each cell K^l we compute for the refinement stages

$$d_{K^l} = \min\{|u_{K^l} - \tilde{u}_{K^l}|, |v_{K^l} - \tilde{v}_{K^l}|\}, \tag{8}$$

and use the maximum for the coarsening stages of the algorithm.

Roughly speaking, the more regular a function u is over K^l, the smaller is the corresponding detail coefficient. This motivates the so-called *thresholding* procedure, which consists in discarding all control volumes corresponding to details that are smaller in absolute value than a level-dependent tolerance ε_l. Choosing ε_l too small or too large will make the MR device inefficient (the compression rate is poor) or deteriorate the quality of the solution due to large thresholding errors, respectively.

For problems considered in [5] and [10], the reference scheme has a known order of convergence in space ($\alpha = 1/2$ and $\alpha = 2$ respectively). The latter constant is at present unknown for FV discretizations of degenerate parabolic equations. However, in [2, 3] we found that a methodology based on the ideas of [5, 10] can also be successfully applied to degenerate reaction-diffusion systems when α is a convergence rate obtained from numerical experiments. This approach is also applied here. Let us denote by α the experimental convergence rate of (4) and (5), which by means of standard preliminary computations (see e.g., [2]) we have found to be $\alpha = 1.2$.

Let the level-dependent tolerances ε_l be given by $\varepsilon_l = 2^{2(l-H)}\varepsilon_R$ for $l = 0, \ldots, H$. If the general time evolution operator is L^1-contractive and the reference numerical scheme is stable in the sense of (6) and the reference tolerance ε_R is set to

$$\varepsilon_{\mathrm{R}} = C\,\frac{2^{-(\alpha+2)H}}{|\Omega|\max\limits_{K\in\mathscr{T},\, t^n<T}\left(\left|f^n_{u,K}\right|+\left|f^n_{v,K}\right|+\left|g^n_{u,K}\right|+\left|g^n_{v,K}\right|\right)+D|\Omega|^{3/2}2^{2+H}}, \tag{9}$$

then the error due to thresholding is of the same order as the discretization error, and therefore the order of the underlying scheme is preserved. The constant C in (9) has to be determined by test calculations on a uniform grid (and possibly in one space dimension only), prior to the proper MR simulation, see e.g., [4, Example 4].

We organize the cell averages and corresponding details at different levels in a *dynamic graded tree*. The root is the basis of the tree. A parent node has four sons, and the sons of the same parent are called brothers. A node without sons is a leaf. A given node has $s' = 2$ nearest neighbors in each spatial direction, needed for the computation of the fluxes of leaves; if these neighbors do not exist, we create them as virtual leaves. Brothers are also considered nearest neighbors. We denote by Λ the set of all nodes of the tree and by $\mathscr{L}(\Lambda)$ the restriction of Λ to the leaves. We apply this MR representation to the spatial part of the function $\mathbf{u} = (u, v)$, which corresponds to the numerical solution of the underlying problem for each time step, so we need to update the tree structure for the proper representation of the solution during the evolution. To this end, we apply the above thresholding strategy, but always ensure the graded tree structure of the data.

We define the *data compression rate* $\eta := N/(2^{-(2H)}N + \#\mathscr{L}(\Lambda))$, where N is the number of control volumes in the full finest grid at level $l = H$, and $\#\mathscr{L}(\Lambda)$ is the number of leaves. The *speed-up* between the CPU times of the numerical solutions obtained by the FV and MR methods is defined by $\mathscr{V} := \mathrm{CPU\,time_{FV}}/\mathrm{CPUtime_{MR}}$.

4 Numerical Results

Example 1 corresponds to the AP model (1), (2) on $\Omega = (0, 256)^2$ (in millimeters). The physiological parameters are $\rho_1 = 0.0026$, $\rho_2 = 0.837$, $C_1 = 20$, $C_2 = 3$, $C_3 = 15$, $a = 0.06$, $k = 3$, $\nu_1 = 1.8$, $\eta_1 = 0.01$, $\eta_2 = 1.0$, $\eta_3 = 0.3$, $D = 2\,\mathrm{cm}^2/\mathrm{s}$ [1, 11]. We consider an inhomogeneity in the conductivity of the medium by setting $A(u) = 0$ if $\|(x - 230, y - 160)\| \le 20$ and $A(u) = Du$ otherwise. This circular obstacle could represent a scar on the cardiac tissue. For simplicity, we impose no-flux boundary conditions on the border of the obstacle. We let $u_0 = 0.9$ if $x \le 128$ and $y = 129$, $u_0 = 0$ otherwise, $v_0 = 2$ if $y \le 128$ and $v_0 = 0$ otherwise.

Figure 1 shows the evolution of u for this example. Clearly, the spiral turbulence, which otherwise dominates the evolution of the system, remains away from the obstacle, and the MR-based adaptive mesh adequately captures the excitation fronts. From (9) we obtain $\varepsilon_{\mathrm{R}} = 4.50 \times 10^{-4}$, and this value indeed produces experimental rates of convergence of about $h^{1.2}$ (see the upper part of Table 1).

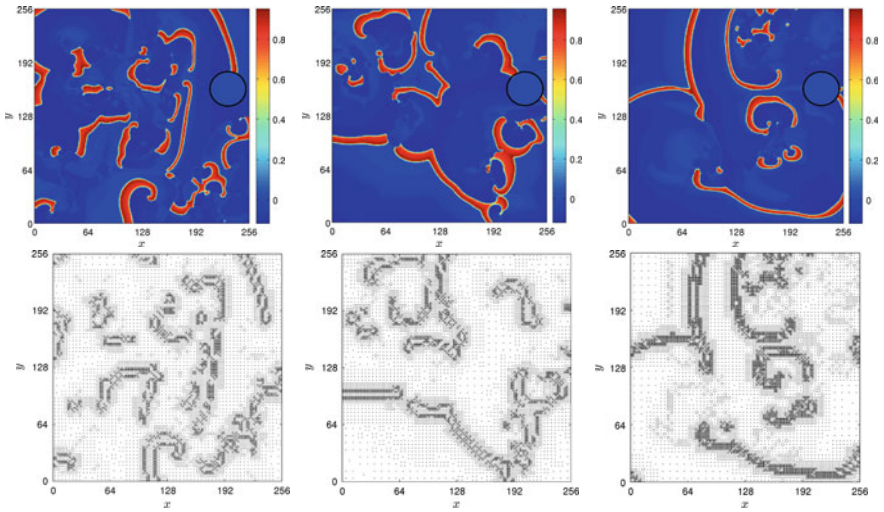


Fig. 1 Example 1. Aliev–Panfilov model: transmembrane potential u and corresponding graded tree structure for times (from *left* to *right*) $t = 0.1$ s, $t = 1$ s and $t = 1.4$ s

Table 1 Examples 1 and 2. Convergence history in different norms, and compression rates

Ex.	h_H	L^1-error	L^1-rate	L^2-error	L^2-rate	$L^\infty-$error	$L^\infty-$rate	η	$\mathscr{V}$
	2^2	4.31×10^{-2}	–	3.44×10^{-2}	–	8.11×10^{-2}	–	9.4112	9.4293
	2^1	1.83×10^{-2}	1.2371	1.47×10^{-2}	1.2309	3.43×10^{-2}	1.2407	11.0309	13.9917
1	2^0	7.92×10^{-3}	1.2133	6.12×10^{-3}	1.2632	1.45×10^{-2}	1.2386	13.1710	17.4209
	2^{-1}	3.31×10^{-3}	1.2482	2.64×10^{-3}	1.2238	6.11×10^{-3}	1.2490	17.2136	20.8701
	2^{-2}	1.42×10^{-3}	1.2710	1.13×10^{-3}	1.2461	2.60×10^{-3}	1.2589	21.8554	28.0526
	2^{-5}	6.72×10^{-2}	–	5.29×10^{-2}	–	8.15×10^{-2}	–	7.3650	11.7923
	2^{-6}	2.99×10^{-2}	1.1675	2.35×10^{-2}	1.1704	3.62×10^{-2}	1.1715	9.8097	16.6464
2	2^{-7}	1.31×10^{-2}	1.1967	1.03×10^{-2}	1.1928	1.59×10^{-2}	1.1899	12.3146	21.9165
	2^{-8}	5.72×10^{-3}	1.2033	4.51×10^{-3}	1.2049	6.91×10^{-3}	1.2031	15.1622	28.1796
	2^{-9}	2.53×10^{-3}	1.2089	1.90×10^{-3}	1.2160	3.03×10^{-3}	1.2097	20.7391	34.1880

In Example 2 we employ the LRII model (3) on $\Omega = (0, 8)^2$ (in centimeters) with $D = 1.25 \times 10^{-3}\,\mathrm{cm}^2/\mathrm{ms}$. Initially the tissue has an constant rest state $u = -84\,\mathrm{mV}$. To produce computational fibrillation, a reentrant wave is generated using a wavefront which after 0.25 ms is broken at the center of the domain. The external stimuli $I_{\mathrm{ext}} = -100\,\mu\mathrm{A/cm}^2$ for $t < 1\,\mathrm{ms}$ and $x < 0.2\,\mathrm{cm}$, and $I_{\mathrm{ext}} = -50\,\mu\mathrm{A/cm}^2$ for $315\,\mathrm{ms} < t < 316\,\mathrm{ms}$, $x \geq 4.5\,\mathrm{cm}$, $y \geq 4.5\,\mathrm{cm}$ are applied. The domain is initially discretized in $N = 256^2$ control volumes, the time step is set according to (6), and we use $\varepsilon_{\mathrm{R}} = 5.15 \times 10^{-3}$. The initial value for [Ca] is 2×10^{-4} mmol/L.

Figure 2 shows the numerical solution for u along with the corresponding representation of the leaves of the dynamic graded tree, which form the adaptive mesh generated by the MR algorithm. We observe that the wave created by the first

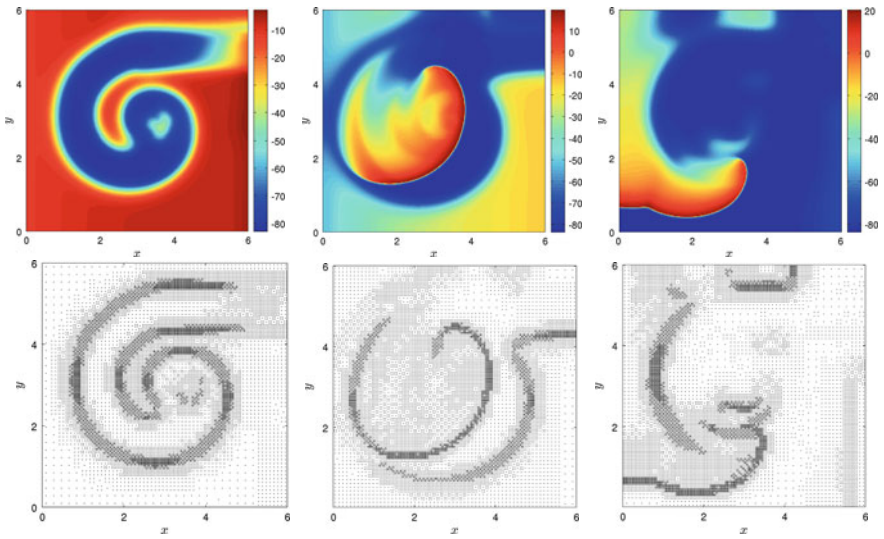


Fig. 2 Example 2. LRII model: transmembrane potential u and corresponding graded tree structure for times (from *left* to *right*) $t = 100$ ms, $t = 200$ ms and $t = 300$ ms

stimulus applied at the left border of the domain propagates to the right, and after applying the second stimulus, a rotating spiral wave forms. The MR device performs an automatic refinement/coarsening stage to accurately capture the high gradient fronts.

The lower half of Table 1 reports the convergence history of the MR method together with η and $\mathscr{V}$ (corresponding to $t = 200$). As in Example 1, the errors maintain a convergence rate of around $h^{1.2}$ and $\mathscr{V}$ grows linearly with the number of control volumes in the finest mesh N_H, and η reaches similar levels as in Example 1. In the reference scheme, for Example 1, most of the computational time is spent in resolving the diffusive part, while for Example 2, the stiffness of the ODE system for the gating variables requires the major part of the overall computational cost. In contrast to Example 1 and other multicomponent problems (see e.g., [2]), we here do not use (8) for the refinement and coarsening procedures, but only use the information on u, i.e., the whole system is evolved over a mesh whose construction is based on the local regularity of u. This simplification avoids the computation of details for all seven gating variables while maintaining a reasonable accuracy level.

Acknowledgements RB is supported by Fondecyt proj. 1090456, Fondap proj. 15000001, and BASAL project CMM, U. de Chile and Centro de Investigación en Ingeniería Matemática (CI^{2}MA), U. de Concepción. RR is supported by the Europ. Res. Council Advanced Grant "Mathcard, Mathematical Modelling and Simulation of the Cardiovascular System", proj. ERC-2008-AdG 227058.

References

1. Aliev, R.R., Panfilov, A.V.: A simple two-variable model of cardiac excitation. Chaos Solit. Fract. **7**, 293–301 (1996)
2. Bendahmane, M., Bürger, R., Ruiz-Baier, R.: A multiresolution space-time adaptive scheme for the bidomain model in electrocardiology. Numer. Meth. Partial Diff. Eqns. (to appear)
3. Bendahmane, M., Bürger, R., Ruiz-Baier, R., Schneider, K.: Adaptive multiresolution schemes with local time stepping for two-dimensional degenerate reaction-diffusion systems. Appl. Numer. Math. **59**, 1668–1692 (2009)
4. Bürger, R., Ruiz-Baier, R., Schneider, K.: Adaptive multiresolution methods for the simulation of waves in excitable media. J. Sci. Comput. **43**, 261–290 (2010)
5. Cohen, A., Kaber, S., Müller, S., Postel, M.: Fully adaptive multiresolution finite volume schemes for conservation laws. Math. Comp. **72**, 183–225 (2003)
6. Eymard, R., Gallouët, T., Herbin, R.: Finite Volume Methods. In Ciarlet, P.G., and Lions, J.L. (eds.), Handbook of Numerical Analysis, vol. VII. North-Holland, Amsterdam, pp. 713–1020 (2000)
7. Keener, J., Sneyd, J.: Mathematical Physiology I: Cellular Physiology II: Systems Physiology, Second Edition. Springer, New York (2009)
8. Luo, C., Rudy Y.: A dynamic model of the cardiac ventricular action potential – simulations of ionic currents and concentration changes. Circ. Res. **74**, 1071–1097 (1994)
9. Müller, S.: Adaptive Multiscale Schemes for Conservation Laws. Springer, Berlin (2003)
10. Roussel, O., Schneider, K., Tsigulin, A., Bockhorn, H.: A conservative fully adaptive multiresolution algorithm for parabolic PDEs. J. Comput. Phys. **188**, 493–523 (2003)
11. Shajahan, T.K., Sinha, S., Pandit, R.: Spiral-wave dynamics depends sensitively on inhomogeneities in mathematical models of ventricular tissue. Phys. Rev. E **75**, 011929 (2007)

On the Numerical Approximation of the Laplace Transform Function from Real Samples and Its Inversion

R. Campagna, L. D'Amore, A. Galletti, A. Murli, and M. Rizzardi

Abstract Many applications are tackled using the Laplace Transform (LT) known on a countable number of real values [J. Electroanal. Chem. 608, 37–46 (2007), Int. J. solid Struct. 41, 3653–3674 (2004), Imaging 26, 1183–1196 (2008), J. Magn. Reson. 156, 213–221 (2002)]. The usual approach to solve the LT inverse problem relies on a regularization technique combined with information a priori both on the LT function and on its inverse (see for instance [http://s-provencher.com/pages/contin.shtml]).

We propose a fitting model enjoying LT properties: we define a *generalized spline* that interpolates the LT function values and mimics the asymptotic behavior of LT functions. Then, we prove existence and uniqueness of this model and, through a suitable error analysis, we give a priori approximation error bounds to confirm the reliability of this approach. Numerical results are presented.

1 Introduction

We focus on the recovery of a real function $f(t)$, $t \geq 0$, given its Laplace transform F:

$$F(z) = \int_0^\infty e^{-zt} f(t) dt \ , \tag{1}$$

for real discrete values of z. This means that the Laplace Transform is known *only on a finite* set of real samples, $(x_i, F(x_i))_{i=1,\ldots,n}$, and the inverse function $f(t) = \mathscr{L}^{-1}[F(z)]$ is searched for.

R. Campagna (✉), L. D'Amore, and A. Murli
University of Naples Federico II, Complesso Universitario M.S. Angelo, Via Cintia, 80126 Naples, Italy
e-mail: rosanna.campagna@unina.it, luisa.damore@dma.unina.it, almerico.murli@dma.unina.it

A. Galletti and M. Rizzardi
University of Naples Parthenope, Centro Direzionale, Isola C4, 80143 Naples, Italy
e-mail: ardelio.galletti@uniparthenope.it, mariarosaria.rizzardi@uniparthenope.it

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_21,

We propose a fitting model enjoying Laplace Transform properties, that describes the restriction of F on the real axis in its convergence region. By this way it is possible to use any numerical method to compute $f(t)$. The organization of the paper is as follows. In Sect. 2 we recall the main asymptotic properties of Laplace transform functions, and we characterize it by introducing the set of *rational decay functions*. Section 3 is devoted to the definition of the fitting model, that is the *rational approximation model*, and to the proof of its existence and uniqueness. In Sect. 4, through a suitable error analysis, we give a priori approximation error bounds. In order to show the usefulness of this approach and the reliability of the error approximation bounds, numerical experiments are given in Sect. 5. Conclusions and future works are discussed in Sect. 6.

2 Preliminary

We suppose that the data $(x_i, F(x_i))_{i=1,\dots,n}$ belong to an unknown Laplace transform with *asymptotic rational decay*: almost all Laplace transform functions F are rational functions or have an *asymptotic rational decay*. Moreover they are analytic in the real half line (α_f, ∞), with α_f related abscissae of convergence. Let be $m \geq 0$ and $C^{\omega}(\Omega_m)$ the set of analytic functions in $\Omega_m = (m, +\infty)$.

Definition 1. If $\alpha > 0$ we refer to

$$\mathscr{R}^{\alpha}_{decay}(\Omega_m) = \left\{ F \in C^{\omega}(\Omega_m) : \begin{array}{r} \exists\, G(x) = a_1 x^{-\alpha} + a_2 x^{-(\alpha+h)} \quad \text{with } a_1 \neq 0, \\ a_1, a_2 \in \Re, \;\; h > 0 \; s.t. \text{ for } k = 0, 1, \dots \\ F^{(k)}(x) = G^{(k)}(x) + o\left(x^{-(\alpha+h+k)}\right) \end{array} \right\}$$

as the *asymptotic rational decay functions set* of order α.

Briefly, functions $F \in \mathscr{R}^{\alpha}_{decay}(\Omega_m)$ are analytic and can be written as

$$F(x) = \frac{a_1}{x^{\alpha}} + \frac{a_2}{x^{\alpha+h}} + o\left(x^{-(\alpha+h)}\right). \tag{2}$$

In other words, F assumes the form $a_1 x^{-\alpha}$ or goes to zero as $a_1 x^{-\alpha}$ and admits an asymptotic expansion with the second term of the form $a_2 x^{-(\alpha+h)}$. So $a_1 x^{-\alpha}$ and $a_2 x^{-(\alpha+h)}$ are respectively the *end behavior model function* of F and of $F - a_1 x^{-\alpha}$. Examples of functions with asymptotic rational decay are given by the following.

Proposition 1. *[2] Let $F(z)$ be a complex-valued function such that $H(z) = F(1/z)$ is holomorphic in the closed complex disk*

$$A_r = \{z \in \mathscr{C} : |z| \leq r\}.$$

Then $F \in \mathscr{R}^{\alpha}_{decay}(\Omega_{1/r})$ with $\alpha = \min\{n \in \mathscr{N} : H^{(n)}(0) \neq 0\}$.

Let us denote by $\mathscr{P}_n$ the linear space of polynomials of degree at most n, and let us introduce the set of *proper rational functions*:

$$\mathscr{R}_{s,m} = \left\{ \frac{p_s(x)}{q_m(x)} : m > s \geq 0,\ p_s \in \mathscr{P}_s \text{ and } q_m \in \mathscr{P}_m \text{ are coprime} \right\}. \quad (3)$$

Most of the Laplace transforms are *proper rational functions*; for these functions we have the following result:

Proposition 2. *[2] Let be F a Laplace transform with convergence abscissa $\alpha_f < 0$. If $F \in \mathscr{R}_{s,m}$, in the form*

$$F(x) = \frac{p_s(x)}{q_m(x)} = \frac{d_s x^s + d_{s-1}x^{s-1} + \cdots + d_0}{c_m x^m + c_{m-1}x^{m-1} + \cdots + c_0} \quad (4)$$

with $c_m = 1$, $d_s \neq 0$, then $F \in \mathscr{R}^{m-s}_{decay}(\Re^+)$ and it can be written as in (2) with

$$\alpha = m - s,\ a_1 = d_s,\ \ a_2 = d_{s-l-1} - d_s c_{m-l-1},\ \ h = l + 1$$

where we set $d_{-1} = d_{-2} = \cdots = d_{s-m} = 0$ and

$$l = \min\{k \in \{0, \ldots, m-1\} : d_{s-k-1} - d_s c_{m-k-1} \neq 0\}.$$

3 The Approximation Model

In this section we introduce the fitting model of $(x_i, F(x_i))_{i=1,\ldots,n}$. To tackle with the scarce information about the Laplace Transform from which the samples rise, we assume the model to be *interpolating*. Taking into account the main asymptotic behavior presented in Sect. 2, we set the model to be a suitable generalized spline; in literature is actually known the great usefulness of spline functions in applications due to their structural properties as well as excellent approximation power [4, 8]. Since the model inherits the high accuracy level of polynomial splines in approximating *smooth* functions between the nodes $[x_1, x_n]$, and the asymptotic behavior of rational decay functions in $[x_n, +\infty[$, according to the Laplace transform decrease towards zero, we refer to it as the *rational approximation model*. The pieces representing the restrictions of the whole model in each subinterval $[x_j, x_{j+1}] \subset [x_1, \infty)$, $j = 1, \ldots, n-1$, will be tied together by continuity of successive derivatives. Boundary conditions will affect the behavior of the fitting model outside the nodes.

We assume that F is a Laplace transform in the form (2), α_f its abscissa of convergence and

$$\Delta_n = \{x_1 < x_2 < \ldots < x_n\}, \quad (5)$$

with $x_1 > \max\{0, \alpha_f\}$, is a set of real sample data in which F is known:

$$y_j = F(x_j) \qquad j = 1, \ldots, n.$$

Let us denote by $\mathscr{S}_3(\Delta_n)$ the real linear space of twice continuously differentiable cubic splines defined on Δ_n and by

$$\mathscr{U}^\alpha = span\,\{x^{-\alpha}\}\ , \quad \alpha \in \Re^+ \tag{6}$$

the one dimensional linear space of functions defined in $\Re^+$.

We introduce the set of real-valued functions whose restriction in $[x_1, x_n]$ belongs to $\mathscr{S}_3$, and whose restriction in $[x_n, \infty)$ belongs to $\mathscr{U}^\alpha$. These functions will be interpolating F in Δ_n and approximating F in the whole interval $[x_1, +\infty[$. This set can be viewed as a set of *generalized splines* [8].

Definition 2. Let Δ_n be as in (5) and $\mathscr{U}^\alpha$ as in (6). We denote by $\mathscr{S}_{\mathscr{U}^\alpha}$ the linear space

$$\mathscr{S}_{\mathscr{U}^\alpha} = \mathscr{S}_{\mathscr{U}^\alpha}(\mathscr{S}_3(\Delta_n), \mathscr{U}^\alpha; \{x_n\}) = \left\{ s : \begin{array}{c} \text{there exist } s_1 \in \mathscr{S}_3(\Delta_n), s_2 \in \mathscr{U}^\alpha, \\ \text{with } s_1(x_n) = s_2(x_n) \quad \text{and} \\ s = s_1 \text{ on } [x_1, x_n), s = s_2 \text{ on } [x_n, \infty) \end{array} \right\}$$

We refer to $\mathscr{S}_{\mathscr{U}^\alpha}$ as the *rational approximation model*.

A function s belonging to $\mathscr{S}_{\mathscr{U}^\alpha}$ has smoothness properties between the knots and decays to zero as a Laplace transform with rational decay. Moreover

Definition 3. We denote by $s_{\Delta_n,F}$ a function in $\mathscr{S}_{\mathscr{U}^\alpha}$ interpolating a function F at Δ_n, that is

$$s_{\Delta_n,F}(x_j) = F(x_j) \quad j = 1, \ldots, n\ . \tag{7}$$

We refer to $s_{\Delta_n,F}$ as a *generalized rational decay approximation spline*.

3.1 Existence and Uniqueness: Boundary Conditions

Let $j \in \{2, \ldots, n\}$ be such that $y_j / y_{j-1} > 0$, we firstly introduce

$$\alpha_j = \frac{\log(y_{j-1}/y_j)}{\log(x_j/x_{j-1})}, \qquad \beta_j = y_j x_j^{\alpha_j}\ . \tag{8}$$

α_j and β_j can be viewed as the parameters of

$$g_j(x) = \beta_j x^{-\alpha_j}, \qquad g_j \in \mathscr{U}^{\alpha_j}\ , \tag{9}$$

that interpolates F at x_{j-1}, x_j. Moreover, in the following, we set

$$h_{j+1} = x_{j+1} - x_j \quad (j = 1, \ldots, n-1)\ , \quad \|\Delta_n\| = \max_{j=1,\ldots,n-1} h_{j+1}\ . \tag{10}$$

In order to obtain a unique $s_{\Delta_n,F} \in \mathscr{S}_{\mathscr{U}^\alpha}$ interpolating F at Δ_n, we adjoin three suitable boundary conditions.

Asymptotic Behavior Condition

1. In $[x_n, \infty[$ we force the model to satisfy two backward interpolation conditions to (x_{n-1}, y_{n-1}) and (x_n, y_n), i.e., it holds

$$\beta x_{n-1}^{-\alpha} = y_{n-1} \qquad \text{and} \qquad \beta x_n^{-\alpha} = y_n \tag{11}$$

and so $\alpha \equiv \alpha_n$ and $\beta \equiv \beta_n$.

Clamped Spline Boundary Conditions

We set the restriction of $s_{\Delta_n,F}$ between the knots to be a clamped spline function where we assign two boundary conditions. We assume that, near the ends x_1 and x_n, F can be approximated by its interpolating functions

$$g_2(x) = \beta_2 x^{-\alpha_2}, \qquad g_n(x) = \beta_n x^{-\alpha_n}$$

respectively at nodes x_1, x_2 and x_{n-1}, x_n. We get two conditions for $s_{\Delta_n,F}$ by setting

2. the first slope of $s_{\Delta_n,F}$ at x_1 as

$$s^{(1)}_{\Delta_n,F}(x_1) = g_2^{(1)}(x_1) = -\alpha_2 y_1/x_1; \tag{12}$$

3. the last slope of $s_{\Delta_n,F}$ at x_n as

$$s^{(1)}_{\Delta_n,F}(x_n) = g_n^{(1)}(x_n) = -\alpha_n y_n/x_n. \tag{13}$$

Starting from the definition of rational approximation model and from the boundary conditions we obtain the following result.

Theorem 1. ***(Existence and uniqueness)*** *[2] Let Δ_n be as in (5), $y_2/y_1 > 0$, $y_{n-1}/y_n > 1$ and $\alpha_2, \beta_2, \alpha_n, \beta_n$ as in (8). Then there exists only one function $s_{\Delta_n,F} \in \mathscr{S}_{\mathscr{U}^{\alpha_n}}$ that verifies conditions (11)–(13).*

4 Approximation Error Analysis

This section is devoted to the error analysis. Let us give some definitions and notations that will be used in the rest of the paper.

Definition 4. Let F be a Laplace transform, Δ_n as in (5), $x_{n+1} = +\infty$ and $s_{\Delta_n,F} \in \mathscr{S}_{\mathscr{U}^{\alpha}}$ a function interpolating F at Δ_n. Then we refer to the following terms

$$E_j(F, \Delta_n) = \max_{x \in [x_j, x_{j+1})} |s_{\Delta_n,F}(x) - F(x)|, \qquad j = 1, \ldots, n \tag{14}$$

as the *local approximation errors*.

Errors E_j $(j = 1, \ldots, n)$ are estimated for Laplace transforms F belonging to $\mathscr{R}^{\alpha}_{decay}(\Re^+)$ and with non-positive abscissa of convergence ($\alpha_f \leq 0$).

Let us consider the errors E_j $(j = 1, \ldots, n-1)$ in $[x_1, x_n]$. The restriction of $s_{\Delta_n,F}$ to this interval is a clamped spline with boundary slopes as in (12) and (13). Starting from a result for complete splines [4], by assuming that

$$e_1 = F^{(1)}(x_1) - \alpha_2 y_1/x_1 \qquad e_n = \alpha_n y_n/x_n - F^{(1)}(x_n) \tag{15}$$

can be neglected, we get:

Theorem 2. *[2] Let be Δ_n as in (5), $F \in C^4[x_1, x_n]$, $s_{\Delta_n,F}(x)$ the unique complete cubic spline interpolating F at nodes Δ_n and verifying the boundary conditions*

$$s^{(1)}_{\Delta_n,F}(x_1) = F^{(1)}(x_1) \qquad s^{(1)}_{\Delta_n,F}(x_n) = F^{(1)}(x_n)\ .$$

For $j = 1, \ldots, n-1$, if there exist L_j for which $|F^{(4)}(x)| \leq L_j$ for $x \in [x_j, x_{j+1}]$, then

$$E_j \leq B_j = h^2_{j+1} \cdot R + \frac{h^4_{j+1}}{4} L_j = \mathscr{O}(\|\Delta_n\|^4) \quad \text{with} \quad R = \max_{j=1,\ldots,n} r_j \quad \text{and} \tag{16}$$

$$r_1 = \frac{3}{4} h_2^2 L_1, \quad r_j = \frac{3}{4} \max(h_j, h_{j+1})^2 \max(L_j, L_{j+1}), \quad r_n = \frac{3}{4} h_n^2 L_{n-1} \tag{17}$$

Now let us consider the error E_n in $[x_n, +\infty)$. In this interval $s_{\Delta_n,F}$ is completely described by the values α_n and β_n. We obtain the following result:

Theorem 3. *[2] Let $F \in \mathscr{R}^{\alpha}_{decay}(\Omega_m)$ be as in (2), $\Delta_n \subset \Omega_m$ be as in (5), α_n be as in (8) and $D_n = \frac{a_2}{a_1} \frac{1}{x_n^h}$. Then*

$$E_n \leq B_n = \frac{|a_1\ D_n|}{x_n^{\alpha}} + \max\left\{ \frac{|a_1\ D_n|}{x_n^{\alpha}}\ ,\ \frac{|\alpha_n - \alpha|}{\alpha} |y_n| \right\} + o\left(\frac{1}{x_n^{\alpha+h}} \right) \tag{18}$$

5 Numerical Experiments

In this section we show numerical results. We assume that the Laplace transform $F(s) = \arctan s^{-1}$ is known only on $n = 50$ real samples $\{x_1, \ldots, x_n\} \in [0.36, 22]$ with $h_{min} = \min_i h_i = 0.2$ and $h_{max} = \max_i h_i = 0.6$.

Local Error Bounds

We compare the true local errors $E_1, \ldots, E_n$ with their bounds (16) and (18).
A heuristic sharper result can be obtained by substituting the maximum value R with the values r_j, by introducing the quantities

$$\bar{B}_j = h_{j+1}^2 \cdot r_j + \frac{h_{j+1}^4}{4} L_j, \qquad j = 1, \ldots, n-1 .$$

The following Figs. 1 and 2 compare local errors E_j ("$-\diamond$" in both subfigures) with their theoretical bounds B_j ("$-$o" on the left side) and with the heuristic ones $\bar{B}_j$ ("$-$o" on the right side). Heuristic bounds seem to better estimate the errors E_j than the theoretical ones, because they take into account the information about the local upper bounds L_j of the fourth derivative of F.

Let us consider error E_n. Firstly observe that $F \in \mathscr{R}^1_{decay}(\Re^+)$ and it can be written in the form (2) with $\alpha = 1$, $h = 2$, $a_1 = 1$, $a_2 = 1/3$. Then, by using Theorem 3, we get the bound $B_n = 6.9959e-005$ as a reliable approximation of the true error $E_n = 1.6364e-005$. Observe that it is $F(x_n) = 4.7201e-002$, i.e., the rational approximation model $s_{\Delta_n,F}$ offers a sharp approximation of the asymptotic behavior of F.

Numerical Inversion

We apply to $s_{\Delta_n,F}$ a numerical method $\mathscr{A}$ for inverting the Laplace transform. In this test, $\mathscr{A}$ represents the Rjabov algorithm which uses only real values [3]. We compare the computed solution $\mathscr{A}[s_{\Delta_n,F}(s)](t)$ with the *true* inverse Laplace transform function $f(t) = \mathscr{L}^{-1}[F(s)] = t^{-1}\sin(t)$. Figure 2 compares the true solution $f(t)$ ("$-\diamond$" in both subfigures) to the computed solutions $\mathscr{A}[F(s)](t)$ (solid line on the left side) and $\mathscr{A}[s_{\Delta_n,F}(s)](t)$ (solid line on the right side). As expected, the accuracy of $\mathscr{A}[s_{\Delta_n,F}(s)](t)$ as approximation of $f(t)$ depends on the magnitude of the approximation error introduced by $s_{\Delta_n,F}$ as well as on the stability of the numerical method $\mathscr{A}$.

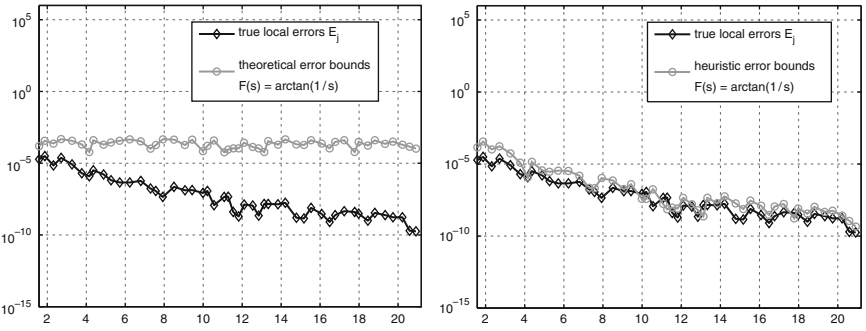


Fig. 1 *Left*: true local errors E_j ("$-\diamond$") **vs** theoretical bounds B_j ("$-$o"). *Right*: true local errors E_j ("$-\diamond$") **vs** heuristic bounds $\bar{B}_j$ ("$-$o"). x-values belong to the sample interval [0.36,22]

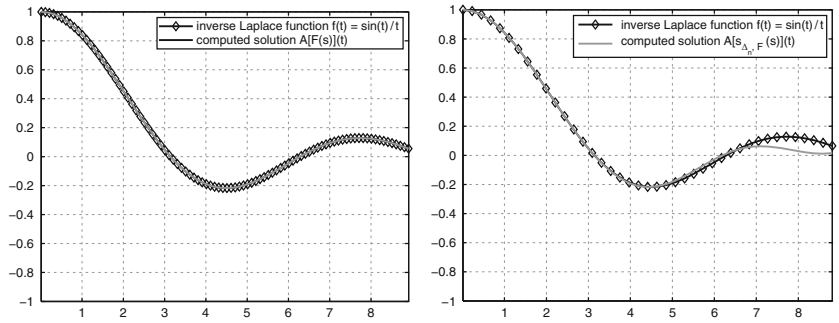


Fig. 2 Computed solutions. *Left*: $\mathscr{A}[F(s)](t)$ ("—") **vs** $f(t) = t^{-1}\sin(t)$ ("—◇"). *Right*: $\mathscr{A}[s_{\Delta_n,F}(s)](t)$ ("—") **vs** $f(t) = t^{-1}\sin(t)$ ("—◇")

6 Conclusions and Future Work

We deal with the problem of numerical inverting the Laplace transform in case of real samples. We introduce a fitting model for approximating Laplace transform functions. We show results concerning existence and uniqueness and we give approximation error bounds. Finally by means of a numerical example we show that this approach is reliable and justified. Future works are: to extend the fitting model for approximating (few) Laplace functions that have *exponential asymptotic decay*; to get computable error bounds from the theoretical results; to perform comparisons with other approaches employed in presence of discrete data [1, 6].

References

1. Borgia, G.C., Brown, R.J.S., Fantazzini, P.: Uniform penalty inversion of multiexponential decay data II. J. Magn. Reson., **147**, 273–285 (2000)
2. Campagna, R., D'Amore, L., Galletti, A., Murli, A., Rizzardi, M.: Existence and Uniqueness of a fitting model of the Laplace Transform from real samples. Preprint n. 26 (2009), Dipartimento di Matematica e Applicazioni "R. Caccioppoli", Università degli Studi di Napoli Federico II
3. Cuomo, S., D'Amore, L., Murli, A., Rizzardi, M.: Computation of the inverse Laplace transform based on a collocation method which uses only real values. JCAM, **198**, 98–115 (2007)
4. de Boor, C.: A Practical Guide to Splines. Springer, New York (1978)
5. Faure, PF, Rodts, S.: Proton NMR relaxation as a probe for setting cement pastes. Magn. Reson. Imaging, **26**, 1183–1196 (2008)
6. http://s-provencher.com/pages/contin.shtml
7. Montella, C., Michel, R., Diard, J.P.: Numerical inversion of Laplace transforms. A useful tool for evaluation of chemical diffusion coefficients in ion-insertion electrodes investigated by PITT. J. Electroanal. Chem., **608**, 37–46 (2007)
8. Schumaker, L.L.: Spline Functions: Basic Theory. Wiley, New York (1981)
9. van der Weerd, L., Melnikov, S.M., Vergeldt, F.J., Novikov, E.G., Van As, H.,: Modelling of self-diffusion and relaxation time NMR in multicompartment systems with cylindrical geometry. J. Magn. Reson., **156**, 213–221 (2002)
10. Zhao, X.: An efficient approach for the numerical inversion of Laplace transform and its application in dynamic fracture analysis of a piezoelectric laminate. Int. J. solid Struct., **41**, 3653–3674 (2004)

A Motion-Aided Ultrasound Image Sequence Segmentation

D. Casaburi, L. D'Amore, L. Marcellino, and A. Murli

Abstract We focus on *segmentation and tracking* of left ventricle and atrium (LVA) deformations in ultrasound images. We propose a fast, reliable and automatic approach to extract the LVA contour during the cardiac cycle. The approach combines a preliminary speckle reduction -based on non linear coherent diffusion model- with a motion-aided LVA border segmentation- based on geodesic level set active contours. A markers-controlled evolution of the segmentation level set surface is employed as a prior knowledge about the shape of the LVA chamber. The extent of this result is the deployment of an automatic stopping criterion. Computational kernels are sparse linear systems solved using GMRES iterative method equipped with AMG multigrid preconditioner. Experiments on real data are discussed.

1 Introduction

Computer Aided Diagnosis (CAD) is a growing application domain of medical analysis. In order to improve the diagnostic performance and to reduce the dependence of human expertise reliable and automatic software tools are strongly required. Here we are concerned with Echography (ultrasound imaging of the heart), one of the driving application areas of CAD.

- *Related works*: Literature on methods for segmenting and tracking the left ventricle and atrium (LVA) deformation is extensive (the reader may refer to [8] for a complete review). Many techniques uses an a priori anatomical knowledge

L. Marcellino (✉)
University of Naples Parthenope, Centro Direzionale, Isola C-4, 80143 Naples, Italy
e-mail: marcellino@uniparthenope.it

D. Casaburi, L. D'Amore, and A. Murli
University of Naples Federico II, Complesso Universitario M.S. Angelo, Via Cintia, 80126 Naples, Italy
e-mail: daniela.casaburi@dma.unina.it, luisa.damore@dma.unina.it, almerico.murli@dma.unina.it

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_22,

(general shape, location, and orientation of objects) incorporated in the form of initial conditions and data constraints by using level set models. The first application of level set models to 2-D medical data segmentation belongs to Malladi [7] and Caselles [9]. Level set methods have been applied both to filtering and segmentation of 2-D and 3-D ultrasound and, particularly, RT3DE (Real Time 3-D Echography) data [3]. Zhou et al. [16] consider LVA shape tracking by combining predictions with observations. An interesting approach is to use an adaptive threshold to create a marker-controller filling to closed objects [12].

- *The present work*: Most of existing efforts do not attempt to tackle segmentation and motion problems in a joint or simultaneous fashion. Since these two problems are not independent from each other more consistent results can be achieved by treating the spatial boundary and the motion tracking problem as a unified process. In the present work we propose a level-set formulation of a motion-aided segmentation. The segmentation model is based on a level set equation where the edge indicator of the segmented image is obtained using information provided by the optic flow computation. In addition we use an automatic markers-controlled evolution for the segmentation surface. A preliminary speckle reduction is performed on each frame of the sequence. In conclusion, the main contribution of this approach is the exploitation of information provided by the motion field

  1. To take into account the presence of subjective contours in the LVA border segmentation when defining the edge indicator function inside the segmentation model equation,
  2. To automatically determine the initial condition of segmentation model for each frame,
  3. To automatically stop the evolution of segmentation surface, using a set of markers points.

The paper is organized as follows. In Sect. 2, we introduce the mathematical models we are going to use. In Sect. 3, we describe the numerical approach and main computational kernels. Finally, in Sect. 4, results on an in-vivo sequence made of $26 \times 300 \times 300$ (1 cardiac cycle) ultrasound images are presented.

2 The Motion Aided Segmentation Model Equation

Let us give the following definition[1]:
[The image sequence brightness function]: Let $J \subset \Re$ be a bounded interval. Given $t \in J$, let $z(t) \equiv (x(t), y(t)) \in \Omega$, where $\Omega = \Omega_x \times \Omega_y \subset \Re^2$ is the image plane. The image plane Ω should depend on the acquisition time t. In practice, it is the same at each t because it refers to the rectangular plane of the image acquisition.

[1] For results concerning well posedness, stability and convergence the reader may refer, for instance, to [11].

Then, for simplicity of notations, we omit the dependence of Ω on t. We define the image sequence on J as the piecewise differentiable function:

$$I : t \in J \longrightarrow z(t) \in \Omega \longrightarrow I(z(t), t) \equiv I(t) \in [0, 255]$$

The starting point of this work is the level set equation based on *Riemannian mean curvature flow* [10]: *P1 [Left ventricle and atrium border segmentation].* Let $u(\tau, x(t), y(t), t)$ denote the segmentation function defined in $Q \equiv [0, Nscale] \times \Omega \times J$, $g(s) = \frac{1}{1+Ks^2}$ be the diffusion function and $*$ be the convolution with the gaussian function G_σ. At each t, the following PDE problem describes the LVA segmentation:

$$\frac{\partial u}{\partial \tau} = \sqrt{\epsilon + |\nabla u|^2}\, div \left(g(|\nabla G_\sigma * \widetilde{I}(t)|) \frac{\nabla u}{\sqrt{\epsilon + |\nabla u|^2}} \right) \tag{1}$$

with zero Dirichlet boundary conditions and u_0 as initial condition:

$$\begin{cases} u(\tau, x(t), y(t), t) = 0 & \tau \in [0, Nscale],\ (x(t), y(t)) \in \partial\Omega \quad t \in J \\ u(0, x(t), y(t), t) = u_0(x(t), y(t), t) & (x, y) \in \Omega \quad t \in J. \end{cases} \tag{2}$$

$g(s) = 1/\left(1 + Ks^2\right)$ $(K > 0)$ is the Perona–Malik edge-indicator function, G_σ is the Gaussian function. Function u_0 is the so called point-of-view surface, i.e., the initial state of the segmentation function u. A key challenge in many applications is the placement of the initial contour u_0. Since the contour moves either inward or outward, its initial placement will determine the segmentation that is obtained. Moreover, to reduce the dependence on a manual inspection it is desired that the initial contour for each frame of the sequence is automatically selected. We have performed many experiments to understand the influence of the position of the initial contour on the segmentation of the LVA chamber as this moves during a complete cardiac cycle. Finally, we use u_0 defined as follows:

Definition 1. (**Initial Condition of P1**): Let $t = t_0$ be the acquisition time of the first frame, then:

$$u_0(t_0) = \max_{i=1,\ldots,n} \omega_i\,(x(t_0), y(t_0)) \tag{3}$$

where:

$$\omega_i\,(x(t_0), y(t_0)) = \begin{cases} \frac{1}{|(x(t_0), y(t_0)) - (x_i(t_0), y_i(t_0))| + 1} & \text{if } (x(t_0), y(t_0)) \in D_i \\ \frac{1}{R+1} & \text{if } (x(t_0), y(t_0)) \in \Omega - D_i \end{cases} \tag{4}$$

where $D_i, i = 1, ..., n$ are circles of center $C_i = (x_i(t_0), y_i(t_0))$ (*focus-points*) and fixed radius R.

At subsequent frames ($t > t_0$), u_0 has been selected in automatic way using the functions ω_i previously introduced. More precisely, $\forall t > t_0$, and $\forall i = 1, \ldots, n$ are considered those functions ω_i such that: $\forall (x(t), y(t)) \in D_i, \widetilde{I}(x(t), y(t)) \leq H$ where $H = \alpha |max_{\Omega} I(t) - min_{\Omega} I(t)| + min_{\Omega} I(t) (\alpha \in [0, 1])$. The key feature of the segmentation model that we will employ for the LVA border detection is the definition of $\widetilde{I}$ in (1) inside the edge indicator function g. As is known, main difficulties arise in regions in the image corresponding to the so-called *subjective contours*. In our case, subjective contours are those corresponding to the horizontal position of the mitral valve, in those frame where it is open.

We propose to integrate the segmented image with the contour obtained at a previous time using its motion trajectory.[2] More precisely, computation of $\widetilde{I}(t)$ in (1) at each t, requires:

1. A preprocessing for speckle-noise reduction,
2. The integration of subjective contours inside the contour to segment.

1. *P2 [Speckle reduction]*: We use the nonlinear coherent diffusion model proposed in [1]. Let

$$I_S(x(t), y(t), t) = I(x(t), y(t), t) \cdot \eta_m(x(t), y(t)) + \eta_a(x(t), y(t))$$

be the noisy sequence, where η_m and η_a are multiplicative and additive noise. The following PDE problem describes the speckle reduction of image sequence I:

$$\frac{\partial I(\tau, x(t), y(t), t)}{\partial \tau} = \nabla (D |\nabla I|) \qquad (x, y) \in \Omega, \quad t \in J \quad \tau \geq 0$$

with zero Neummann boundary conditions and I_S as initial condition:

$$\begin{cases} \frac{\partial I}{\partial n} = 0 & \tau \in [0, Nscale], (x(t), y(t)) \in \partial \Omega, \quad t \in J \\ I(0, x(t), y(t), t) = I_S(x(t), y(t), t) & (x(t), y(t)) \in \Omega, \quad t \in J. \end{cases} \tag{5}$$

where D is the *diffusion matrix* defined as follows:

$$D = \begin{pmatrix} w_1 & w_2 \end{pmatrix} \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \begin{pmatrix} w_1^T \\ w_2^T \end{pmatrix} \tag{6}$$

w_1, w_2, *are the eigenvectors of the structure matrix* J:

$$J = G_\sigma * (\nabla I \nabla I^T) = \begin{pmatrix} G_\sigma * I_x^2 & G_\sigma * I_x I_y \\ G_\sigma * I_x I_y & G_\sigma * I_y^2 \end{pmatrix} \tag{7}$$

[2] During a single cardiac cycle, which lasts approximately 1 s, the heart contracts from end diastole (ED) to end systole (ES) and expands back to ED. Over this time, Echocg systems can acquire approximately 25 images of the heart. Because adjacent frames are imaged over a short time period (approximately 50 ms), the LVA boundaries exhibit strong temporal correlation. Thus, previous LVA boundaries may provide information regarding the location of the current LVA boundary.

where λ_1 *and* λ_2 *are defined as*:

$$\begin{aligned}\lambda_1 &= \begin{cases} \beta \cdot \left(1 - \frac{(\mu_1-\mu_2)^2}{s^2}\right) & (\lambda_1-\lambda_2)^2 \le s^2 \quad (s,\beta = const) \\ 0 & otherwise. \end{cases} \\ \lambda_2 &= \beta \end{aligned} \tag{8}$$

2. *P3 [Optic flow computation]*: We first recall the *motion trajectory* then, following [14], we compute the (apparent) motion field (or the so-called *optic flow*) by imposing that the spatial brightness gradient, does not change along the motion trajectory. *Let* $t_i, t_{i+1} \in J$, *where* $t_{i+1} > t_i$, $\Delta t = t_{i+1} - t_i$. *The motion trajectory of a point z(t)=(x(t),y(t))* $\in \Omega$ *is the line (or arc) L, defined by the successive positions of* $z(t)$, *as* t *moves from* t_i *towards* t_{i+1}. *The equation for* L *is*:

$$L: \begin{cases} \Delta x = x(t_{i+1}) - x(t_i) = \Delta t \cdot u(t_i) \\ \Delta y = y(t_{i+1}) - y(t_i) = \Delta t \cdot v(t_i) \end{cases}$$

where $(u(t), v(t)) = (\frac{d}{dt}x(t), \frac{d}{dt}y(t))$ *the components of the motion field at each* $z(t) \in \Omega$, *are obtained by solving the following PDE system*:

$$\begin{cases} \frac{\partial u}{\partial \tau} = \alpha \cdot div\left[\phi'(\nabla u \nabla u^T + \nabla v \nabla v^T)\nabla u\right] - 2[I_{xx}u + I_{yx}v + I_{tx}] \cdot I_{xx} + \\ \qquad\qquad -2[I_{xy}u + I_{yy}v + I_{ty}] \cdot I_{xy} \\ \frac{\partial v}{\partial \tau} = \alpha \cdot div\left[\phi'(\nabla u \nabla u^T + \nabla v \nabla v^T)\nabla v\right] - 2[I_{xx}u + I_{yx}v + I_{tx}] \cdot I_{yx} + \\ \qquad\qquad -2[I_{xy}u + I_{yy}v + I_{ty}] \cdot I_{yy} \end{cases} \tag{9}$$

with zero initial conditions and Dirichlet boundary conditions, and $I_t = \frac{\partial I}{\partial t}$, $\alpha > 0$ *is the regularization parameter and* $\phi'(s^2) = \epsilon + \frac{(1-\epsilon)}{2\sqrt{1+\frac{s^2}{\lambda^2}}}$.

Now, given $t_{i+1} > t_i$, and $\Gamma(t_i)$, the LVA border obtained at time t_i, then $\widetilde{I}(t_{i+1})$ in (1) is obtained as $\widetilde{I}(t_{i+1}) = I_{DS}(t_{i+1}) + \Gamma^{pre}(t_{i+1})$ where $\Gamma^{pre}(t_{i+1}) = \{(x(t_i) + u(t_i)\Delta t, y(t_i) + v(t_i)\Delta t), \quad s.t. \quad (x(t_i), y(t_i)) \in \Gamma(t_i)\}$ is the prediction of the position at time t_{i+1} of the curve $\Gamma(t_i)$ and I_{DS} is the despeckled image sequence.

In conclusion, the recovering of missing parts of the LVA border is obtained by combining the image to segment with the contour obtained at a previous time. Using motion trajectory, we first predict the position of the LVA border at a subsequent time, then we integrate this information inside the frame to segment.

The overall problem consists of three successive steps: P2, i.e., speckle reduction, P3, i.e., motion field computation, and P1, i.e., left ventricle segmentation. The related dataflow is the following: $I_S(t) \longrightarrow I_{DS}(t) \longrightarrow (u(t), v(t)) \longrightarrow \widetilde{I}(t) \longrightarrow \Gamma(t)$.

3 Numerical Approach

Discretization has been performed using a semi-implicit scheme with respect the scale parameter leading to linear problems at each scale step: nonlinear terms are treated from the previous scale step while the linear ones are considered on the current scale step. In particular, if $\Delta\tau$ is the scale parameter stepsize, i.e., $\Delta\tau = \frac{T}{N}$, then for $n = 1, 2, \ldots, N$ $\tau_n = n \cdot \Delta\tau$, $n = 0, \ldots, N$ is the grid of the scale interval $[0, T]$. We use a backward difference for the scale derivative. Concerning the space discretization, we use the AOS (Additive Operator Splitting)[13] and finite differences for the speckle reduction equation. By splitting the solution into two separate dimensions and rearrange image (each pixel is then only composed of the two neighboring pixels), in this case we have to only invert diagonal dominant tridiagonal matrices. To this aim, we use the LU factorization without pivoting, specialized for tridiagonal matrices. The accuracy requirement of despeckling (about 2%) allows us to take fully advantage of the efficiency of AOS choosing a step size sufficiently large (we set $\Delta\tau = 2.2$). Linear semi-implicit discretization and finite differences are employed for the optic flow computation. Finally, taking into account that the segmentation model equation is a level set equation where discontinuities in the evolving solution is allowed, following [4], we consider as spatial discretization the complementary volume scheme. Both for optic flow and for segmentation discretization leads to the solution of a linear system where the matrix is block pentadiagonal with tridiagonal blocks along the main diagonal and diagonal blocks along the upper and lower diagonals. We employ the GMRES iterative method equipped with Algebraic Multigrid preconditioner (AMG) with the FALGOUT -CLJP coarse grid selection [2]. Convergence and stability of the numerical solution of (1) can be stated in order to guarantee that (u^n, v^n) is the computed approximation at step n of the motion vector at scale τ_n. This follows from the fact that the system matrix has strong diagonal dominance. This guarantees that $u^{\tau_n}(t)$ is the approximation of the segmentation function at scale level τ_n.

Automatic Stopping Criterion of the Segmentation Surface Evolution
For each time $t \in J$, an automatic stopping criterion of the semi-implicit scheme for the segmentation model (1) has been employed by using a set of *marker* points and their motion trajectory. These points are used to determine the approximation of the segmentation surface as $\tau_n \to +\infty$. *Let* $P_1 = (x_1(t), y_1(t)), \ldots, P_m = (x_m(t), y_m(t)) \in \Omega$ *a set of* m *marker points. Let* $\delta_i = \{P = (x(t), y(t)) \in \Omega : |P - P_i| \leq r\}$ *be* m *neighborhoods of* P_i. *The approximation of the segmentation surface* $z = u(\tau_n, x(t), y(t), t)$ as $n \to +\infty$, *stops at step* n *if* $\exists(\bar{x}(t), \bar{y}(t)) \in \delta_i : u(\tau_n, \bar{x}(t), \bar{y}(t), t) < c, \quad i = 1, \ldots, m$ where $c = 0.5 \cdot [max(u(\tau_n)) - min(u(\tau_n))]$.

As shown in Fig. 1, marker points are manually selected on the first frame ($t = t_0$). They are used as prior knowledge about the shape of the LVA chamber. On subsequent frames, i.e. $\forall t > t_0$, their position is automatically updated using their motion trajectory: $\forall i = 1, \ldots, m, \quad P_i(t_i + \Delta t) = P_i + (u(t_i)\Delta t, v(t_i)\Delta t)$ where $\Delta t = t_{i+1} - t_i$, and $t_{i+1} > t_i > t_0$.

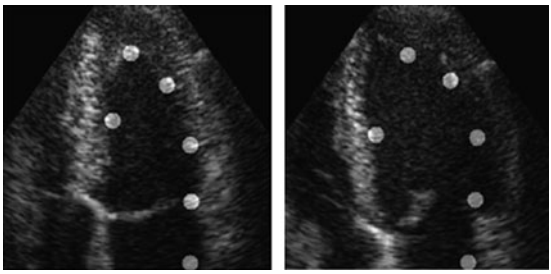


Fig. 1 On the *left*, initial position of the markers on frame 0. On the *right*, position of the markers on frame 14, as given by the motion field

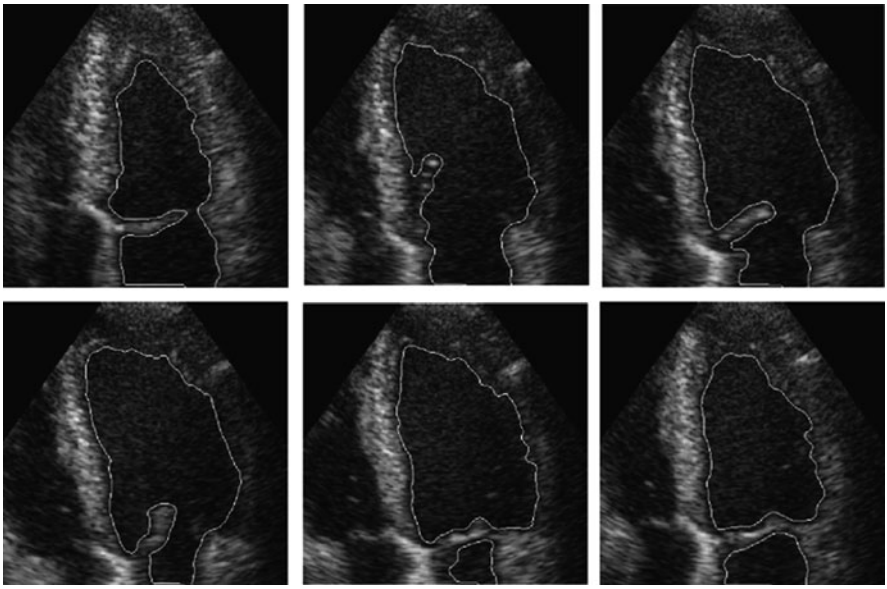


Fig. 2 From *upper left* to *bottom right* the LVA chamber segmentation on frames 3, 9, 15, 19, 22, 25. Despeckling: $\epsilon = 5.0e-2$, $\Delta\tau = 2.2$, $Nscale = 7$. Segmentation: $\epsilon = 1.0$, $\Delta\tau = 0.1$, $Nscale = 46$. Optic flow: $\epsilon = 1.0e-1$, $\Delta\tau = 5.0e-2$, $Nscale = 1$

4 Results

Experiments have been carried out using *PETSc* software library [6] integrated with the package *BoomerAMG* (of Hypre software library) implementing AMG preconditioners [5]. The computing platform is made of 16 blades (1 blade consists of 2 quad core Intel Xeon E5410@2.33GHz) Dell PowerEdge M600. Our experiments confirm that AMG is algorithmically scalable: both for P1 and P3 the convergence factor (per cycle) is very stable at approximately 0.04 for all scales. The setup time averages roughly 1% of cycle time. Finally, the computational work is O(1) per scale step. We found that for segmentation and optic flow, the number of steps

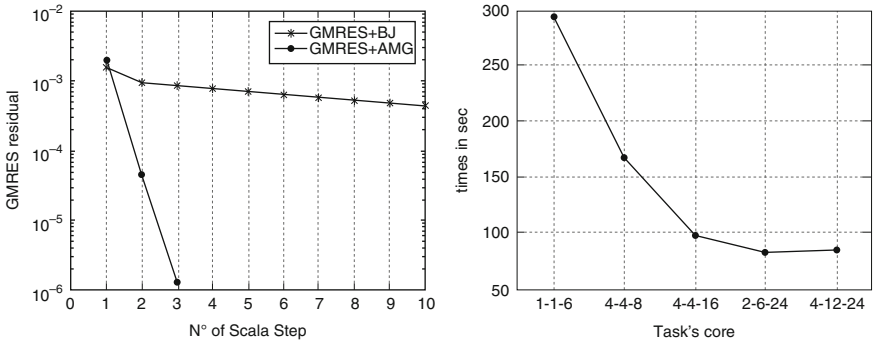


Fig. 3 On the *left*, the convergence rate of GMRES with AMG preconditioner ("...") compared with that obtained using Block Jacobi preconditioner ("- - -") at Nscale = 10. On the *right*, the execution time of the application code versus the core number

of preconditioned GMRES is 3–4. The sequence consists of 26 frames of size 300×300, in Fig. 2 we show those frames related to the aperture of the cardiac valve. The sequential software runs in about 13 min while the parallel software requires 82 s running on 4 blades (32 cores, i.e., 2 for despleckling, 6 for optic flow computation and 24 for segmentation task) of the Dell computing platform (see Fig. 3). We obtain a performance gain of 88, 21% compared to sequential time, with a frame rate per second is 0.034.

References

1. K.Z. Bbbd-Elmoniem, A.M. Youssef, Y.M Kadah, *Real-time speckle reduction and coherence enhancement in ultrasound imaging via nonlinear anisotropic diffusion*. IEEE Trans. Biomed. Eng., Vol. 49(9), pp. 997–1012, 2002
2. A. Clery, R. Falgout, V.E. Henson, J.E. Jones, T. A. Manteuffel, S.F. Mccornick, G.N. Miranda, J.W. Ruge, *Robustness and scalability of algebraic multigrid*, SIAM J. Sci.Comp., Vol. 21(5), pp. 1886–1908, 1999
3. C. Corsi, M. Borsari, F. Consegnati, A. Sarti, C. Lamberti, A. Travaglini, T. Shiota, J.D. Thomas, *Left ventricular endocardial surface detection based on real-time 3D echocardiographic data*, Eur. J. Ultrasound, 13, pp. 41–51, 2001
4. A. Handlovivcov, K. Mikula, F. Sgallari, *Semi-implicit complementary volume scheme for solving level set like equations in image processing and curve evolution*, Numerische Mathematik, 93(4), pp. 675–695, 2003
5. V.E. Henson, U.M. Yang, *BoomerAMG: A parallel algebraic multigrid solver and preconditioner*, Applied Numerical Mathematics, Vol. 41, pp. 155–177, 2002. https://computation.llnl.gov/casc/linear-solvers/sls-hypre.html
6. http://www.mcs.anl.gov/petsc/petsc-as/index.html
7. R. Malladi, J.A. Sethian, *Level Set Methods for Curvature Flow, Image Enchancement, and Shape Recovery in Medical Images, in Visualization and Mathematics*, Eds. H.C. Hege, K. Polthier, pp. 329–345, Springer, Heidelberg, 1997
8. J.A. Noble, "Ultrasound image segmentation: A survey", IEEE Trans. Med. Imaging, Vol. 25(8), pp. 987–1010, 2006

9. G. Sapiro, R. Kimmel, V. Caselles, *Object Detection and Measurements in Medical Images via Geodesic Active Contours*, Proc. SPIE-Vision Geometry, San Diego, July 1995
10. A. Sarti, G. Citti, *Subjective surface and Riemannian mean curvature flow graphs*, Acta Math. Univ. Comeniamae, 70(1), pp. 85–104, 2001
11. J.A. Sethian, *Level Set Methods and Fast Marching Methods Evolving Interfaces in Computational Geometry, Fluid Mechanics, Computer Vision, and Materials Science*, Cambridge University Press, Cambridge, 1999
12. X.C. Tai, E. Hodneland, J. Weickert, N.V. Bukoreshtliev, A. Lundervold, H.H. Gerdes, *Level Set Methods for Watershed Image Segmentation*, Scale Space and Variational Methods in Computer Vision, Eds. F. Sgallari, A. Murli, N. Paragios, LNCS 4485, pp. 178–190, 2007
13. J. Weickert, *Recusrsive Separable Schemes for Nonlinear Diffusion Filters*, Scale Space Theory in Computer Vision, Eds. B. Romey, L. Florack, J. Koendrick, M. Viergever, LNCS 1252, Springer, Heidelberg, pp. 260–271, 1997
14. J. Weickert, *On Discontinuity-Preserving Optic Flow*, Proc. Computer Vision and Mobile Robotics Workshop, pp. 115–122, 1998
15. J. Weickert, C. Shnorr, *A Theoretical Framework for Convex Regularized in PDE-Based Computation of Image Motion*, International Journal of Computer Vision. Vol. 45(3), pp. 2001, 2001
16. X.S. Zhou, D. Comaniciu, A. Gupta, *Information fusion framework for robust shape tracking*, IEEE Trans. Pattern Anal. Mach. Intell., 27(1), pp. 115–129, 2005

A High Order Finite Volume Numerical Scheme for Shallow Water System: An Efficient Implementation on GPUs

M.J. Castro Díaz, M. Lastra, J.M. Mantas, and S. Ortega

Abstract In this work we present a high order finite volume numerical scheme for solving the one layer shallow-water system. The numerical solution of this model is useful for several applications related to geophysical flows, and they impose a great demand of computing power. As a consequence, extremely efficient high performance solvers are required. In this work we perform a GPU implementation of the proposed numerical scheme and some computations are made to test the performance of the implementation.

1 Introduction

Our goal is to efficiently simulate one layer fluids that can be modelled by using a shallow water system, formulated under the form of a conservation law with source terms. The numerical solution of this model is useful for several applications related to geophysical flows: simulation of rivers, channels, dambreak problems, etc. These simulations impose a great demand of computing power due to the dimensions of the domain (space and time). As a consequence, extremely efficient high performance solvers are required to solve and analyze these problems in reasonable execution times. An high order numerical scheme to simulate shallow water system has been presented in [2].

Currently, a cost effective emerging architecture exists which is specially indicated to accelerate considerably computationally intensive tasks like the one

M.J. Castro (✉) and S. Ortega
Facultad de Ciencas, University of Málaga, 29071 Málaga, Spain
e-mail: castro@anamat.cie.uma.es, sergio@anamat.cie.uma.es

M. Lastra and J.M. Mantas
E.T.S. Ing. Informática y Telecomunicaciones, University of Granada, 18071 Granada, Spain
e-mail: mlastra@ugr.es, jmmantas@ugr.es

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_23,

considered in this paper. Modern Graphics Processing Units (GPUs) are not only used to render 3D graphics but can also be a cost effective way to speedup the numerical solution of several mathematical models in science and engineering (see [10, 11] for a revision of the topic). Modern GPUs offer over 100 processing units optimized for performing massively floating point operations in parallel [6]. As a consequence, for several algorithmic structures, these architectures are able to obtain a substantially higher performance than a powerful CPU.

In [5], a explicit central-upwind scheme is implemented on a NVIDIA GeForce 7800 GTX card to simulate the one-layer shallow-water system and a speedup from 15 to 30 is achieved with respect an implementation on an Intel Xeon processor. In [8], a first order path conservative Roe type solver has been implemented on several NVIDIA GeForce cards to simulate the one-layer shallow water system and a speedup of two orders of magnitude faster than a SSE-optimized CPU version of the solver for medium-size problems is achieved.

Here, we follow the strategy described in [8] to design an efficient implementation of the numerical scheme presented in [2] using OpenGL and Cg [3]. We use an utility library described in [8] which facilitates the mapping from CPU to GPU and simplifies the description of the GPU program as sequential composition of data parallel modules (fragment shaders).

2 Mathematical Model: One-Layer Shallow-Water System

The one-layer shallow-system is a system of conservation laws with source terms which models the flow of a shallow layer of homogeneous fluid that occupies a bounded subdomain $D \subset \mathbb{R}^2$ under the influence of a gravitational acceleration g. The system has the following form:

$$\begin{cases} \dfrac{\partial h}{\partial t} + \dfrac{\partial q_x}{\partial x} + \dfrac{\partial q_y}{\partial y} = 0 \\ \dfrac{\partial q_x}{\partial t} + \dfrac{\partial}{\partial x}\left(\dfrac{q_x^2}{h} + \dfrac{g}{2}h^2\right) + \dfrac{\partial}{\partial y}\left(\dfrac{q_x q_y}{h}\right) = gh\dfrac{\partial H}{\partial x} \\ \dfrac{\partial q_y}{\partial t} + \dfrac{\partial}{\partial x}\left(\dfrac{q_x q_y}{h}\right) + \dfrac{\partial}{\partial y}\left(\dfrac{q_y^2}{h} + \dfrac{g}{2}h^2\right) = gh\dfrac{\partial H}{\partial y} \end{cases} \tag{1}$$

where $h(x, y, t) \in \mathbb{R}$ denotes the thickness of the water layer at point (x, y) at time t, $H(x, y)$ is the depth function measured from a fixed level of reference and $q(x, y, t) = (q_x(x, y, t), q_y(x, y, t)) \in \mathbb{R}^2$ is the mass-flow of the water layer at point (x, y) at time t.

Let us denote by

$$F_1(W) = \begin{bmatrix} q_x \\ \frac{q_x^2}{h} + \frac{1}{2}gh^2 \\ \frac{q_x q_y}{h} \end{bmatrix}, \; F_2(W) = \begin{bmatrix} q_y \\ \frac{q_x q_y}{h} \\ \frac{q_y^2}{h} + \frac{1}{2}gh^2 \end{bmatrix},$$

$$S_1(W) = \begin{bmatrix} 0 \\ gh \\ 0 \end{bmatrix}, \; S_2(W) = \begin{bmatrix} 0 \\ 0 \\ gh \end{bmatrix}.$$

Let $J_i(W) = \dfrac{\partial F_i}{\partial W}(W)$, $i = 1, 2$ denote the Jacobians of the fluxes F_i, $i = 1, 2$. Given an unit vector $\boldsymbol{\eta} = (\eta_x, \eta_y) \in \mathbb{R}^2$, we define the matrix $A(W, \boldsymbol{\eta}) = J_1(W)\eta_x + J_2(W)\eta_y$, and the vectors $F(W, \boldsymbol{\eta}) = F_1(W)\eta_x + F_2(W)\eta_y$, $S_{\boldsymbol{\eta}}(W) = \eta_x S_1(W) + \eta_y S_2(W)$.

3 High Order Finite Volume Schemes

Let us consider the computational domain D is divided into M discretization cells or finite volumes, $V_i \subset \mathbb{R}^2$, which are supposed to be closed polygons. Let us denote by $\mathscr{T}$ the set of cells. Hereafter we will use the following notation: given a finite volume V_i, $N_i \in \mathbb{R}^2$ is the center of V_i, $\mathscr{N}_i$ is the set of indexes j such that V_j is a neighbor of V_i; Γ_{ij} is the common edge of two neighbor cells V_i and V_j, and $|\Gamma_{ij}|$ its length; $\boldsymbol{\eta}_{ij} = (\eta_{ij,x}, \eta_{ij,y})$ is the unit vector which is normal to the edge Γ_{ij} and points toward the cell V_j. Let us denote by $|V_i|$ the area of cell V_i.

In order to obtain a high order numerical scheme for system (1) we consider a reconstruction operator, i.e., an operator that associates to a given family $\{W_i\}_{i=1}^{M}$ of values at the cells, two families of functions defined at the edges $\gamma \in \Gamma_{ij} \to W_{ij}^{\pm}(\gamma)$, in such a way that, whenever

$$W_i = \frac{1}{|V_i|} \int_{V_i} W(\mathbf{x})\, d\mathbf{x} \tag{2}$$

for some smooth function W, then $W_{ij}^{\pm}(\gamma) = W(\gamma) + O(\Delta^p), \quad \forall \gamma \in \Gamma_{ij}$.

We will assume that the reconstructions are calculated as follows: given the family $\{W_i\}_{i=1}^{M}$ of values at the cells, first an approximation function is constructed at every cell V_i, based on the values of W_j at some of the cells close to V_i (the *stencil*): $P_i(\mathbf{x}) = P_i\left(\mathbf{x}; \{W_j\}_{j \in \mathscr{B}_i}\right)$, for some set of indexes $\mathscr{B}_i$. If, for instance, the reconstruction only depends on the neighbor cells of V_i, then $\mathscr{B}_i = \mathscr{N}_i \cup \{i\}$. These approximations functions are calculated usually by means

of an interpolation or approximation procedure. Once these functions have been constructed, the reconstructions at $\gamma \in \Gamma_{ij}$ are defined as follows:

$$W_{ij}^-(\gamma) = \lim_{\mathbf{x}\to\gamma} P_i(\mathbf{x}), \quad W_{ij}^+(\gamma) = \lim_{\mathbf{x}\to\gamma} P_j(\mathbf{x}). \tag{3}$$

Clearly, for any $\gamma \in \Gamma_{ij}$ the following equalities are satisfied: $W_{ij}^-(\gamma) = W_{ji}^+(\gamma)$ and $W_{ij}^+(\gamma) = W_{ji}^-(\gamma)$.

We suppose that the reconstruction operator satisfies the following properties:

(HP1) It is conservative, i.e., the following equality holds for any cell V_i:

$$W_i = \frac{1}{|V_i|}\int_{V_i} P_i(\mathbf{x})d\mathbf{x}. \tag{4}$$

(HP2) It is of order p, verifying $W(\gamma) - W_{ij}^{\pm}(\gamma) = \Delta^p g_{ij}(\gamma) + O(\Delta^{p+1})$, for any $\gamma \in \Gamma_{ij}$, being g_{ij} a regular function.

(HP3) It is of order q in the interior of the cells, i.e., if the operator is applied to a sequence $\{W_i\}$ satisfying (2) for some smooth function $W(\mathbf{x})$, then $P_i(\mathbf{x}) = W(\mathbf{x}) + O(\Delta^q)$, $\forall \mathbf{x} \in int(V_i)$.

(HP4) The gradient of P_i provides an approximation of order m of the gradient of W, $\nabla P_i(\mathbf{x}) = \nabla W(\mathbf{x}) + O(\Delta^m)$, $\forall \mathbf{x} \in int(V_i)$.

Once the reconstruction operator has been chosen, the general expression of a semi-discrete scheme high order Roe Scheme is the following:

$$\begin{aligned} W_i'(t) = -\frac{1}{|V_i|}\Bigg[& \sum_{j\in\mathcal{N}_i}\int_{\Gamma_{ij}} \mathscr{D}^-(W_{ij}^-(\gamma,t), W_{ij}^+(\gamma,t), H_{ij}^-(\gamma), H_{ij}^+(\gamma), \boldsymbol{\eta}_{ij})\, d\gamma \\ & + \int_{V_i}\left(S_1(P_i^t(\mathbf{x}))\frac{\partial P_i^H}{\partial x}(\mathbf{x}) + S_2(P_i^t(\mathbf{x}))\frac{\partial P_i^H}{\partial y}(\mathbf{x})\right) dx\Bigg], \end{aligned} \tag{5}$$

where P_i^t are the approximation functions corresponding to the cell values $W_i(t)$, and P_i^H is the approximation function corresponding to the cell values of the given bathimetry. $W_{ij}^{\pm}(\gamma,t)$, $H_{ij}^{\pm}(\gamma)$ are given, respectively, by

$$W_{ij}^-(\gamma,t) = \lim_{\mathbf{x}\to\gamma} P_i^t(\mathbf{x}), \quad W_{ij}^+(\gamma,t) = \lim_{\mathbf{x}\to\gamma} P_j^t(\mathbf{x}),$$

$$H_{ij}^-(\gamma) = \lim_{\mathbf{x}\to\gamma} P_i^H(\mathbf{x}), \quad H_{ij}^+(\gamma) = \lim_{\mathbf{x}\to\gamma} P_j^H(\mathbf{x}).$$

$$\begin{aligned} \mathscr{D}^-(W_L, W_R, H_L, H_R, \boldsymbol{\eta}) = {} & F(W_L, \boldsymbol{\eta}) + P_{LR}^-\left(\mathscr{A}_{LR}(W_R - W_L)\right. \\ & \left. -\mathscr{S}_{LR}(H_R - H_L)\right). \end{aligned} \tag{6}$$

where in the particular case of system (1), we define

$$\mathscr{A}_{LR} = \begin{bmatrix} 0 & \eta_x & \eta_y \\ (-\bar{u}_x^2 + \bar{c}^2)\eta_x - \bar{u}_x\bar{u}_y\eta_y & 2\bar{u}_x\eta_x + \bar{u}_y\eta_y & \bar{u}_x\eta_y \\ -\bar{u}_x\bar{u}_y\eta_x + (-\bar{u}_y^2 + \bar{c}^2)\eta_y & \bar{u}_y\eta_x & \bar{u}_x\eta_x + 2\bar{u}_y\eta_y \end{bmatrix},$$

$$\mathscr{S}_{LR} = \left[0,\, \bar{c}^2\eta_x,\, \bar{c}^2\eta_y\right]^T.$$

Here:

$$\bar{c} = \sqrt{g\bar{h}};\ \bar{u}_\alpha = \frac{\sqrt{h_L}u_{L,\alpha} + \sqrt{h_R}u_{R,\alpha}}{\sqrt{h_L} + \sqrt{h_R}},\ \alpha = x, y;\ \bar{h} = \frac{h_L + h_R}{2}. \tag{7}$$

$$P_{LR}^{-} = \frac{1}{2}\mathscr{K}_{LR} \cdot (I - \mathrm{sgn}(\mathscr{L}_{LR})) \cdot \mathscr{K}_{LR}^{-1}, \tag{8}$$

where $\mathscr{L}_{LR}$ is the diagonal matrix whose coefficients are the eigenvalues of $\mathscr{A}_{LR}$, and $\mathscr{K}_{LR}$ is a matrix whose columns are associated eigenvectors. Finally $\mathrm{sgn}(\mathscr{L}_{LR})$ is the diagonal matrix whose coefficients are the sign of the eigenvalues of the matrix $\mathscr{A}_{LR}$.

The previous numerical scheme provides an approximation of order at least $\alpha = \min(p, q, m)$ for regular solutions of system (1) (see [2] for more details).

The high order extension considered in this work is based on the third order bi-hyperbolic reconstruction introduced in [12] that generalizes the 1d reconstruction presented in [9] (see also [13]). The time-stepping used for the third order scheme is based on an optimal TVD Runge–Kutta method (see [4]). The integral terms have been approximated by means of a Gaussian quadrature of order three. Finally, in order to obtain a high order well-balanced numerical scheme for the one-layer shallow water system, the reconstruction procedure is applied to the variables $\varphi = h - H$, q_x, q_y and H, recovering h by setting $h = \varphi + H$. Due to the explicit character of the numerical scheme, the usual CFL condition has to be imposed (see [2] for details).

4 Obtaining of a GPU Implementation

We have designed a data parallel numerical algorithm from the mathematical description of the numerical scheme. Initially, the finite volume mesh must be constructed from the input data with the appropriate setting of initial and boundary conditions.Then the time stepping is performed by applying a third order Runge–Kutta TVD method, consisting on three steps. At each step, the spatial discretization described in (5) must be performed as follows:

1. *Reconstrucion and integral volume computation*: First a reconstruction procedure at each cell must be performed to define the functions $P_i^t()$ and $P_i^H()$ at each cell. Also, the integral volume

$$\int_{V_i} \left(S_1(P_i^t(\mathbf{x})) \frac{\partial P_i^H}{\partial x}(\mathbf{x}) + S_2(P_i^t(\mathbf{x})) \frac{\partial P_i^H}{\partial y}(\mathbf{x}) \right) dx,$$

   is computed using a third-order gaussian quadrature formula.
2. *Edge-based calculations:* The integral boundary term

$$\int_{\Gamma_{ij}} \mathscr{D}^-(W_{ij}^-(\gamma,t), W_{ij}^+(\gamma,t), H_{ij}^-(\gamma), H_{ij}^+(\gamma), \boldsymbol{\eta}_{ij})\, d\gamma,$$

   must be computed at each edge of the mesh. A third-order gaussian quadrature formula is also used. The vector $\mathscr{D}^-(W_{ij,l}^-, W_{ij,l}^+, H_{ij,l}^-, H_{ij,l}^+, \boldsymbol{\eta}_{ij}) \in \mathbb{R}^3$, $l = 1, \cdots, r$ must be computed at each quadrature point ($r = 2$ if the third order guassian quadrature formula is used). The computation of these contributions can be computed independently for each edge and it is the most costly calculation in the numerical algorithm because it includes several 3×3 matrix computations. Moreover, we only need the data corresponding to the reconstructions of the variables at the volumes V_i and V_j, therefore these computations present a high arithmetic intensity and locality. The value $\Delta t_{ij,l} = \omega_l |\Gamma_{ij}| \parallel \mathscr{L}_{ij,l}^n \parallel_\infty$ must be computed and added to the partial sums associated to each cell (Δt_i and Δt_j).
3. *Computation of the local Δt for each volume*: For each volume V_i, the value of Δt_i is modified to compute the local Δt per volume. In the same way, the computation for each volume can be performed in parallel.
4. *Computation of Δt^n*: The minimum of all the local Δt values previously obtained for each volume must be computed. This phase can also be parallelized if the minimum is calculated following a recursive decomposition approach [7].
5. *Computation of $W_i^{n+1,s}$*: The $n+1, s$th state of each volume must be approximated from the nth and the $n+1, s-1$th states using the data computed at the previous phases. This phase can also be completed in parallel.

We can make the following remarks from the description of the parallel algorithm: the computation steps required by the problem addressed here can be classified into two groups: computations associated to edges and computations associated to volumes. The scheme presents a high arithmetic intensity and the computation exhibits a high degree of locality. The scheme exhibits a high degree of data parallelism because the computation at each edge/volume is independent with respect to the computation peformed to the rest of edges/volumes. The remarks indicate that this problem seems suitable for being implemented on modern graphics processing units. In the numerical scheme presented, the volume state is represented by a 3-tuple and all the operations involve operations between 3-tuples and 3×3 matrices which makes it even more suited for a GPU based computing platform. The only drawback of using GPUs is the need to adapt the computational process

to the graphics processing pipeline and make some mappings between the problem domain and this pipeline.

5 Numerical Test

In order to test the solver, we have considered a circular dam-break problem in a square domain $[-1, 1]\times[-1, 1]$ with a depth function $H(x, y) = 1.0-0.5e^{-(x^2+y^2)}$. As initial condition we set:

$$h(x, y, 0) = \begin{cases} H(x, y) + 0.3 \text{ if } x^2 + y^2 \leq \frac{1}{8} \\ \\ H(x, y) \qquad \text{otherwise,} \end{cases} \qquad q_x(x, y, 0) = q_y(x, y, 0) = 0.$$

Table 1 CPU time (in seconds)

N. Cells	OpenMp	9800 GTX	GTX260
2,500	3.00	1.08	1.05
10,000	10.78	2.13	1.93
40,000	74.50	6.44	4.77
160,000	589.36	36.66	20.21
640,000	4967.2	277.66	142.94
2,560,000	400,10	2179.6	1107.7

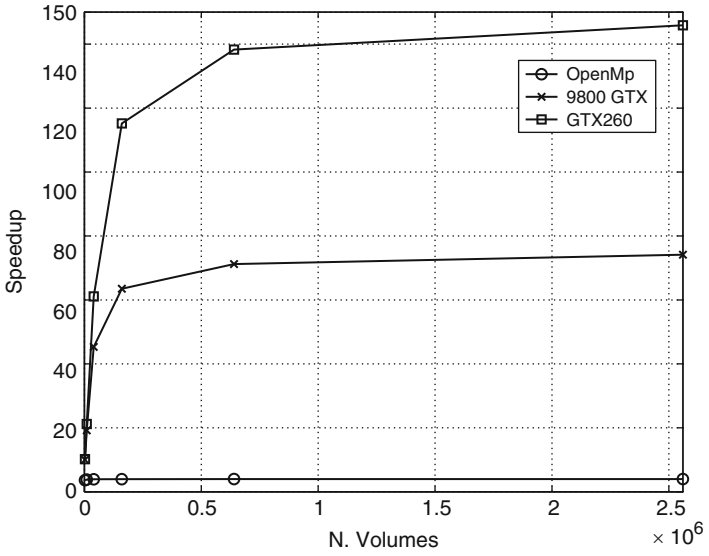


Fig. 1 Speedup vs. number of cells for an OpenMp parallel implementation (four cores), GPU implementation using 9800 GTX and GTX260 graphics cards

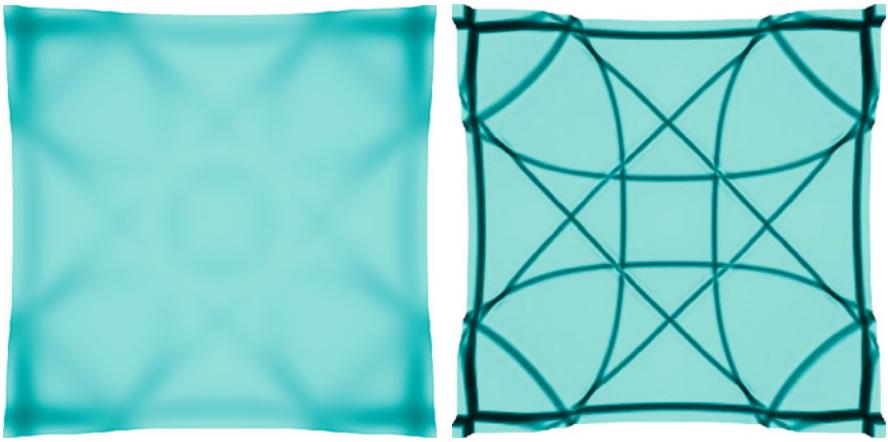


Fig. 2 Circular dam-break problem over a non-flat bottom topography: free surface at t = 1 s. First order Roe method (*left*). High order Roe method (*right*). 400 × 400 mesh

Six uniform meshes of the domain, $Q_k, k = 0, \ldots, 5$, are constructed such that the number of volumes of mesh Q_k is given by $2^{2k} \cdot 2.5 \cdot 10^3, k = 0, \ldots, 5$. The numerical scheme is run in the time interval $[0, 1]$. The CFL parameter is $\gamma = 0.9$ and wall boundary conditions are considered ($\mathbf{q} \cdot \boldsymbol{\eta} = 0$). Table 1 shows the CPU times for an optimized CPU implementation of a Quad-core Intel Xeon Nocona 2.66 Ghz with emt64 extensions, using the SSE CPU units through the use of the Intel Performance Primitives 4.1 (see [1]), for the GPU implementation on a Nvidia GeForce 9800 GTX and on a Nvidia GeForce GTX260. The CPU reduction has been dramatically reduced (see Table 1). In fact a speedup of approximately 140 is achieved for meshes of practical interest using GTX260 card (see Fig. 1). Moreover, we have checked that the use of single precision arithmetic of GPU does not affect in a essential manner to the quality of the numerical solution (see Fig. 2)

Acknowledgements J. Mantas acknowledges partial support from DGI-MEC project MTM2005-08024. J. Mantas, M. Lastra also acknowledge partial support from DGI-MEC project TIN2004-07672-c03-02. M. Castro and S. Ortega acknowledge partial support from DGI-MEC project MTM2006-08075.

References

1. Castro, M.J., García-Rodríguez, J.A., González-Vida, J.M., Parés, C., Solving shallow-water systems in 2D domains using Finite Volume methods and multimedia SSE instructions, J. Comput. App. Math., 221: 16–32, 2008
2. Castro, M.J., Fernández-Nieto, E.D., Ferreiro, A.M, García-Rodríguez, J.A., Parés, C., *High order extensions of Roe schemes for two dimensional nonconservative hyperbolic systems*, J. Sci. Comput., 39: 67–114, 2009

3. Fernando, R., Kilgard, M.J., *The Cg Tutorial: The Definitive Guide to Programmable Real-Time Graphics*, Addison-Wesley, MA, 2003
4. Gottlieb, S., Shu, C.W. *Total variation diminishing Runge–Kutta schemes.* Mat. Comp., 67: 73–85, 1998
5. Hagen, T.R., Hjelmervik, J.M., Lie, K.-A., Natvig, J.R., Ofstad Henriksen M., *Visual simulation of shallow-water waves*, Sim. Modelling Pract. Th., 13: 716–726, 2005
6. http://www.nvidia.com
7. Kumar, V., Grama, A., Gupta, A., Karypis, G., *Introduction to Parallel Computing*, Benjamin, MA, 2003
8. Lastra, M., Mantas, J.M., Ureña, C., Castro, M.J., Garca, J.A., *Simulation of shallow-water systems using graphics processing units.* Accepted on Math. Comp. Simul., 80(3): 598–618, 2009
9. Marquina, A. *Local piecewise hyperbolic reconstruction of numerical fluxes for non linear scalar conservation laws.* SIAM, J. Sci. Comput., 15(4): 892–915, 1994
10. Owens, J.D, Luebke, D., Govindaraju, N., Harris, M., Krüger, J., Lefohn, A.E., Purcell, T., *A Survey of General-Purpose Computation on Graphics Hardware*, Eurographics 2005 State of the Art Report, 2005
11. Rumpf, M., Strzodka, R., *Graphics processor units: New prospects for parallel computing*, L. N. in Computational Science and Engineering, 51: 89–121, 2006
12. Schroll, H.J., Svensson, F. *A bihyperbolic fiinite volume method for quadrilateral meshes.* SIAM: J. Sci. Comput., 26(2): 237–260, 2006
13. Serna, S. *A class of extended limiters applied to piecewise hyperbolic methods.* SIAM: J. Sci. Comput., 28(1): 123–140, 2006

Spectral Analysis for Radial Basis Function Collocation Matrices

R. Cavoretto, A. De Rossi, M. Donatelli, and S. Serra-Capizzano

1 Abstract and Outline of the Paper

The aim of this paper is to provide tools and results for the analysis of the linear systems arising from radial basis function (RBF) approximations of partial differential equations (PDEs), see e.g., [1,9]. Informally, a radial function $\phi(x) : \mathbb{R}^n \to \mathbb{R}$ is a function of the Euclidean norm $||x||$ of x, i.e., $\phi(x) = \eta(||x||)$, for $\eta(t) : \mathbb{R} \to \mathbb{R}$. Examples are functions of the following form $\sqrt{t^2 + c^2}$, multiquadric (MQ), $1/\sqrt{t^2 + c^2}$, inverse multiquadric (IMQ), $e^{-\frac{t^2}{c^2}}$, Gaussian. In this context c is the *shape parameter*, whose value plays a role in modeling problems with various specific features. At least numerically, it is evident that the precision of the approximation procedures based on RBFs is very high. In fact, if h denotes the maximal step size, then the approximation error behaves like $O(\lambda^{c/h})$ for the MQ and like $O(\lambda^{\sqrt{c}/h})$ for the IMQ and Gaussian, where λ is a positive parameter, strictly less than one, and independent of h. The price that has to be paid concerns the increasing ill-conditioning of the related linear systems in which a growth of the order of $e^{\theta c/h}$ is observed at least for large values of c/h, with θ being a positive constant independent of h and also of the shape parameter c. We are interested in the spectral behavior of the resulting matrices, and especially in the extremal behavior (conditioning) and in the global distribution results: such a study is crucial for designing fast and accurate solution methods. A first important step in understanding the spectral behavior of the considered matrices was done in [3], where the remarkable link with Toeplitz sequences generated by a symbol was exploited. Here we give a more precise analysis than in [3], by showing that for some choices of RBF, e.g., IMQ, Gaussian, and for some values of the parameter c/h, the conditioning is not

R. Cavoretto (✉) and A. De Rossi
Dipartimento di Matematica, Università di Torino, Via Carlo Alberto 10, 10123 Torino, Italy
e-mail: roberto.cavoretto@unito.it, alessandra.derossi@unito.it

M. Donatelli and S. Serra-Capizzano
Dipartimento di Fisica e Matematica, Università dell'Insubria – Sede di Como, Via Valleggio 11, 22100 Como, Italy
e-mail: marco.donatelli@uninsubria.it, stefano.serrac@uninsubria.it

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_24,

exponential, but grows mildly as n^2. Furthermore, the spectral analysis is extended from the Toeplitz component to the whole matrix-sequence, by including the boundary conditions term, and some insights on the multidimensional setting are given. The paper is organized as follows. In Sect. 2 we state the collocation technique for the Poisson problem by emphasizing the linear algebra viewpoint. In Sect. 3 we recall spectral properties of Toeplitz matrix-sequences generated by a symbol. In Sect. 4 we give a rigorous explanation of some numerics reported in [3] and we refine the previous results studying the behavior of extremal eigenvalues (conditioning) also in the two-dimensional case. In Sect. 5 we obtain global spectral distribution results for the complete matrix-sequence, taking into account also the boundary conditions.

2 The Linear Algebra Problem from RBF

For n positive integer, let $x_0 = 0 < x_1 < \cdots < x_n < x_{n+1} = 1$ and define $\phi(x) = \eta(|x|)$, where $\eta(t)$ is any function in the class considered in the first section. We are looking for an approximation to the solution in the vector space spanned by the functions $\phi(x - x_i)$, $i = 0, 1, \ldots, n + 1$. This yields a linear system whose coefficient matrix $A_{n+2} = (a_{i,j})_{i,j=0,n+1} \in \mathbb{R}^{(n+2)\times(n+2)}$ is such that $a_{0,j} = \phi(x_0 - x_j)$, $a_{n+1,j} = \phi(x_{n+1} - x_j)$ for $j = 0, \ldots, n + 1$ and $a_{i,j} = \phi''(x_i - x_j)$ for $i = 1, \ldots, n$, $j = 0, \ldots, n + 1$. Let us denote by $T_n = (\phi''(x_i - x_j))_{i,j=1,n}$ the submatrix of A_{n+2} obtained by removing its first and last row and column. When the set $x_i = ih$, $i = 0, \ldots, n + 1$, for $h = 1/(n + 1)$, forms a grid of equally spaced points in the interval $[0, 1]$ the matrix $T_n = (\phi''((i - j)h))$ is a symmetric Toeplitz matrix, i.e., its entries are function of $i - j$. Moreover A_{n+2} is a rank-2 correction to a symmetric Toeplitz matrix. In [3] the authors provided explicit asymptotic estimates, as function of c/h to the condition number $\mu(T_n)$. According to [3], there exists a function $\rho(g)$ of g such that for any n it holds $\mu(T_n) < \rho(g)$, where equality is reached only for $n \to \infty$. Furthermore, an interesting asymptotical estimate $\gamma(g)$ of $\rho(g)$, for $g \to \infty$, was proved:

$$\rho(g) \approx \gamma(g) = \begin{cases} (e^{\pi})^g/(\pi\sqrt{2g}) & \text{for the MQ,} \\ (e^{2\pi})^g/(2e^2\pi^{3/2}g^{3/2}) & \text{for the IMQ,} \\ (e^{\pi^2})^{g^2}/(2e\pi^2 g^2) & \text{for the Gaussian.} \end{cases} \tag{1}$$

Here $\phi(x) \approx \psi(x)$ if $\lim_{x\to\infty} \phi(x)/\psi(x) = 1$, while $\phi(x) \sim \psi(x)$ means asymptotical equivalence that is the existence of two positive constants r and R independent of x, such that $r\phi(x) \le \psi(x) \le R\phi(x)$ for every x large enough. From the numerical experiments performed in [3], the above asymptotic bounds are very strict even for small values of g when $\rho(g)$ is small, while the quantity $\rho(g)$ becomes an extremely pessimistic upper-bound when $\rho(g)$ is moderate (say e.g., $g = 3, 4$). For larger values of g the picture changes again since the conditioning

becomes exponential in g, but almost independent of n. However in [3], an important question was not rigorously answered: in which cases the quantity $\rho(g)$ is a faithful approximation (and not only a mere upperbound) of the condition number $\mu(T_n)$? In which case, if not numerically, $\rho(g)$ captures at least the asymptotic order of the conditioning? A rigorous explanation of some of these phenomena is given in Sect. 4. In the two-dimensional case, it can be easily verified that the series associated with the IMQ and the Gaussian functions are convergent so that the Toeplitz matrix machinery can be in principle applied for these two classes of radial functions, by taking into consideration the associated continuous symbol. The challenge concerns the case of MQ radial functions where the symbol is discontinuous at $x = 0$, $y = 0$ (it diverges to $+\infty$), but it seems to be a smooth function in the rest of the domain. It should be recalled that the asymptotical spectral behavior of Toeplitz sequences is well understood far beyond the continuous setting, since the symbol is required to be simply Lebesgue integrable.

3 Toeplitz Matrices and Spectral Properties

Let f be a Lebesgue integrable function defined on $(0,1)^d$ and taking values in $\mathbb{C}$. Then, for d-indices $r = (r_1, \ldots, r_d)$, $j = (j_1, \ldots, j_d)$, $n = (n_1, \ldots, n_d)$, $e = (1, \ldots, 1)$, $\underline{0} = (0, \ldots, 0)$, the Toeplitz matrix $T_n(s)$ of size $\hat{n} \times \hat{n}$, $\hat{n} = n_1 \cdot n_2 \cdots n_d$, is defined as $T_n(s) = [\hat{s}_{r-j}]_{r,j=\underline{0}}^{n-e}$, where $\hat{s}_k$ are the Fourier coefficients of s defined by equation $\hat{s}_j = \hat{s}_{(j_1,\ldots,j_d)}(s) = \int_{[0,1]^d} s(t_1, \ldots, t_d) e^{-\mathrm{i}2\pi(j_1 t_1 + \cdots + j_d t_d)}\, dt_1 \cdots dt_d$, with $\mathrm{i}^2 = -1$ and integers j_ℓ such that $-\infty < j_\ell < \infty$ for $1 \le \ell \le d$. The function $s(x)$ is called *symbol*. If $s(x)$ is real valued then $\hat{s}_{-k}$ is the conjugate of $\hat{s}_k$ so that T_n is Hermitian; if, in addition, $s(x)$ is symmetric around the axis $y = 1/2$, then T_n is real symmetric. For the following result see e.g., [4] and references therein.

Theorem 1. *If $s(x) \ge 0$ almost everywhere and not identically constant over $[0,1]$ then T_n is positive definite for any n and its eigenvalues belong to $(\inf s(x), \sup s(x))$, where* inf *and* sup *are intended up to zero Lebesgue measure sets. Moreover, $\lambda_1^{(n)}$ is a decreasing sequence converging to $\inf s(x)$, and $\lambda_n^{(n)}$ is an increasing sequence converging to $\sup s(x)$, where $\lambda_1^{(n)}$ and $\lambda_n^{(n)}$ are the minimal and the maximal eigenvalues of T_n, respectively. Furthermore, if $s(x)$ is locally twice differentiable around its infimum points with positive second derivative in at least one of them, then $\lambda_1^{(n)} - \inf s(x) \sim cn^{-2}$, with c positive constant independent of n. Analogously, if $s(x)$ is locally twice differentiable around its supremum points with negative second derivative in at least one of them, then $\sup s(x) - \lambda_n^{(n)} \sim cn^{-2}$, with c positive constant independent of n. Therefore the spectral condition number $\mu(T_n) = \lambda_n^{(n)}/\lambda_1^{(n)}$ is an increasing sequence converging to $\sup s(x)/\inf s(x)$.*

The above result is true also for $n \times n$ block Toeplitz matrices $T_{n,m}$ with $m \times m$ Toeplitz blocks associated with a symbol $s(x, y) \in L^1([0,1]^2)$. Concerning the

case of matrix-sequences an important notion is that of spectral distribution in the eigenvalue sense, linking the collective behavior of the eigenvalues of all the matrices in the sequence to a given function (or to a measure). The notion goes back to Weyl and has been investigated by many authors in the Toeplitz and Locally Toeplitz context (see [4, 8]). For any function F defined on $\mathbb{C}$ and for any $m \times m$ matrix A, the symbol $\Sigma_\lambda(F, A)$ stands for the mean $\Sigma_\lambda(F, A) := \frac{1}{m}\sum_{j=1}^{m} F\left(\lambda_j(A)\right)$ with $\lambda_j(A)$, $j = 1, \ldots, m$, denoting the eigenvalues of A. Let $\mathscr{C}_0(\mathbb{C})$ be the set of continuous functions with bounded support defined over the complex numbers, d a positive integer, and θ a complex-valued measurable function defined on a set $G \subset \mathbb{R}^d$ of finite and positive Lebesgue measure $\mu(G)$. Here G will be often equal to $(0, 1)^d$ so that $e^{i2\pi\overline{G}} = \mathbb{T}^d$ with $\mathbb{T}$ denoting the complex unit circle. A matrix sequence $\{A_k\}$ is said to be *distributed* (*in the sense of the eigenvalues*) *as the pair* (θ, G), or to *have the distribution function* θ, which is defined by $\{A_k\} \sim_\lambda (\theta, G)$, if, $\forall F \in \mathscr{C}_0(\mathbb{C})$, the limit relation $\lim_{k\to\infty} \Sigma_\lambda(F, A_k) = \frac{1}{\mu(G)} \int_G F(\theta(t))\, dt, t = (t_1, \ldots, t_d)$ is satisfied. For multilevel Toeplitz sequences $\{T_n(s)\}$, with s integrable d variate symbol, the eigenvalues are not explicitly known, but we know the distribution at least when s is real valued, see [8], that is $\{T_n(s)\} \sim_\lambda (s, Q^d)$, $Q = (0, 1)$.

4 A Rigorous Interpretation of Some Numerics

With reference to Sect. 2, we observe that for the considered RBFs $s(x)$ is a smooth positive function (see [3]) so that, when applying Theorem 1, instead of inf and sup we use min and max. In [3] the ratio $\rho(g) = \max s(x)/\min s(x)$ and its asymptotic estimate $\gamma(g)$ have been computed (see (1)), for $g = 1, 2, 3, 4$, in the case of MQ, IMQ and Gaussian function, respectively. The results of this computation are reported in Table 1.

The asymptotic estimates are very precise even for small values of g. However the main interest relies in the evaluation of the actual condition number of $T_n = T_n(s)$. Therefore, we have to compare the values of $\gamma(g)$ with the actual condition numbers of the Toeplitz matrices T_n for several values of n. Table 2 reports the spectral condition number $\mu = \mu(T_n)$ of the Toeplitz matrix T_n, for different values of n in the case $g = 1$, $g = 2$, respectively. It is interesting to point out that in the MQ case with $g = 1, 2$ and IMQ case with $g = 1$, the asymptotic bounds are

Table 1 Values of $\rho(g) = \max s(x)/\min s(x)$ and of its asymptotic estimate $\gamma(g)$

	MQ		IMQ		Gaussian	
g	$\rho(g)$	$\gamma(g)$	$\rho(g)$	$\gamma(g)$	$\rho(g)$	$\gamma(g)$
1	4.73	5.21	8.4	6.5	4.1e2	3.6e2
2	8.1e1	8.5e1	1.2e3	1.2e3	6.5e14	6.5e14
3	1.5e3	1.6e3	3.4e5	3.6e5	7.8e35	7.8e35
4	3.1e4	3.2e4	1.2 e8	1.2 e8	4.4 e65	4.4e65

Table 2 Values of the spectral condition number $\mu(T_n)$ for different values of n and g

n	MQ		IMQ		Gaussian	
	g = 1	g = 2	g = 1	g = 2	g = 1	g = 2
20	4.6	72.6	5.7	6.3	66	246
50	4.7	78.8	7.5	19.0	217	337
100	4.7	80.1	8.1	52.3	338	400
200	4.7	80.4	8.3	147	395	1543
400	4.7	81.0	8.3	375	413	6069

roughly reached for relatively small values of n whereas for the Gaussian case and for the IMQ with $g = 2$ the values of $\mu(T_n)$ are far from the asymptotic value even for moderately large values of n. The explanation relies completely in Theorem 1. Indeed the conditioning is given by

$$\frac{\max s(x) - c_1/n^2}{\min s(x) + c_2/n^2}, \tag{2}$$

with c_1 and c_2 positive constants independent of n. Hence the approximation $\max s(x)/\min s(x)$ is numerically accurate when $\min s(x)$ is far away from zero, but it is not correct when, for larger values of g, the minimum of $s(x)$ exponentially approaches zero. In that case a more reasonable approximation is given by $n^2 \cdot \max s(x)/c_2$, i.e., by approximating $\min s(x)$ with zero and by neglecting the term c_1/n^2 since $\max s(x)$ is positive and dominating. In reality the columns $g = 2$ in Table 2 for IMQ and Gaussian show exactly the predicted growth: when the size n doubles, the value of the conditioning grows by a factor 4, which is coherent with the given guess of an asymptotic growth proportional to n^2. When g becomes larger than 2, the surprise is that we observe another change in the picture. The conditioning becomes extreme: we really appreciate the exponential growth of $\frac{\max s(x)}{\min s(x)}$, but there is no longer dependency on n. How to explain this phenomenon? The reason relies again in formula (2). We recall that Kac, Murdoch, and Szegö gave the expression of c_2 as the second derivative of s in the minimum point which has to be positive by local convexity. Therefore the explanation of the latter phenomenon could be given in terms of $c_2 = c_2(g)$: if $c_2(g)$ is positive but rapidly converging to zero as a function of g, then the quantity $\frac{\max s(x)}{\min s(x)}$ really captures the conditioning of T_n. In other words, the observed behavior can be explained again by formula (2) but we need to show that at the minimum point not only the first derivative is zero but the second derivative is a very small positive number. Unfortunately, as reported in Table 3, the latter statement is completely false in the IMQ and Gaussian setting, while the desired behavior is observed for the MQ radial basis functions (see also Fig. 1). However, concerning IMQ and Gaussian RBFs, for $g \geq 3$ it becomes clear from Fig. 1 that $x = 0$ is not the only minimal point, at least numerically. A further minimal point shows up at $x = 0.5$ and the function has locally the expected behavior. In fact, around $x = 0.5$, the graph of the function becomes flatter and flatter as g increases. This visual evidence is supported also by the numerical values

Table 3 Case 1D – Values of $s''(x)$ for the MQ, IMQ and Gaussian functions

RBF	$s''(x)$	$g = 1$	$g = 2$	$g = 3$	$g = 4$
MQ	$s''(0.5)$	$1.2211 \times 10^{+1}$	$3.3383 \times 10^{+0}$	4.1208×10^{-1}	3.7203×10^{-2}
IMQ	$s''(0)$	$+\infty$	$+\infty$	$+\infty$	$+\infty$
Gaussian	$s''(0)$	$7.1034 \times 10^{+1}$	$1.3995 \times 10^{+2}$	$2.0992 \times 10^{+2}$	$2.7989 \times 10^{+2}$

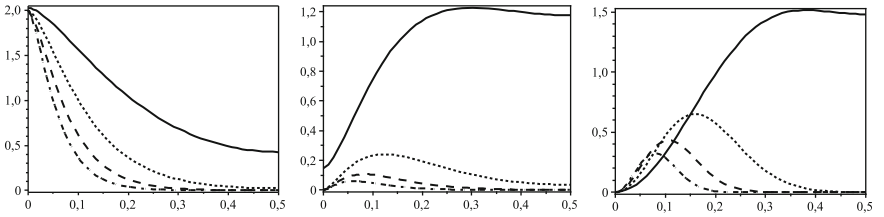


Fig. 1 Case 1D – Graph of $s(x)$ for $g = 1, 2, 3, 4$

Table 4 Case 1D – Values of $s(x)$, $s'(x)$ and $s''(x)$ for the IMQ function

IMQ	$g = 1$	$g = 2$	$g = 3$	$g = 4$
$s(0.5) - s(0)$	$1.0163 \times 10^{+0}$	3.6406×10^{-2}	1.4040×10^{-3}	7.2263×10^{-5}
$s'(0.5)$	0	0	0	0
$s''(0.5)$	$5.1521 \times 10^{+0}$	$3.0658 \times 10^{+0}$	3.1383×10^{-1}	2.3285×10^{-2}

Table 5 Case 1D – Values of $s(x)$, $s'(x)$ and $s''(x)$ for the Gaussian function

Gaussian	$g = 1$	$g = 2$	$g = 3$	$g = 4$
$s(0.5) - s(0)$	$1.4799 \times 10^{+0}$	1.8096×10^{-3}	1.1907×10^{-8}	5.0081×10^{-16}
$s'(0.5)$	0	0	0	0
$s''(0.5)$	$9.9590 \times 10^{+0}$	$2.1205 \times 10^{+0}$	8.3464×10^{-5}	1.1702×10^{-11}

of $s(0.5) - s(0)$, $s'(0.5)$, and $s''(0.5)$ reported in Tables 4 and 5, where it becomes evident that $x = 0.5$ is a further point of local minimum (numerically global) with the desired features.

The value of the second derivative at $x = 0.5$ in our setting shows that $c_2(g)$, as a function of g, rapidly collapses to zero exponentially and therefore for larger g, the expression in the denominator $\min s(x) + c_2/n^2$ can be approximated by $\min s(x)$ since c_2 collapses to zero as $\min s(x)$ but the division by n^2 makes it negligible. Furthermore, the quantity $\max s(x) - c_1/n^2$ can be approximated as usual by $\max s(x)$ since this maximum is always well separated from zero. In conclusion in this setting, for $g \geq 3$, the true approximation of the conditioning is given by $\frac{\max s(x)}{\min s(x)}$ which is extremely high with respect to g, but essentially constant with respect to n. Both in 1D and 2D, as n tends to infinity, the values of minimal and maximal eigenvalues tend to some fixed values. These quantities can

be considered an accurate approximation of the infimum and supremum of the symbol. The only exception is observed in the 2D case for MQ radial basis functions where, as n doubles also the value of the maximal eigenvalue approximately doubles. This exceptional behavior is not a surprise since the symbol defined in this case is unbounded at $x = 0$ or $y = 0$. Indeed for $n = m$ and setting the real size of the matrix T_{mn} is $N = mn$, $\lambda_{\max}(T_{mn})$ grows as $N^{0.52}$. The latter fully agrees with the Riemann-Lebesgue lemma for which we expect $\lambda_{\max}(T_{mn}) = o(N)$. Therefore we have an indication that the symbol $s \in L^1([0,1]^2$ and that the singularity at $(0,0)$ is of the type $(x^2 + y^2)^{-\alpha}$ with α close to 0.52.

5 Spectral Distribution of the Complete Matrix-Sequence

By the Szegö distribution result we know that $\{T_n(s)\} \sim_\lambda (s, Q)$ and $\{T_{mn}(s)\} \sim_\lambda (s, Q^2)$ with $Q = (0,1)$ and we would like to deduce the same distribution result for the real sequences $\{A_n\}$ and $\{A_{mn}\}$ arising in dD RBF collocation, where the real collocation matrix is a rank 2 correction of $T_n(s)$ in the case of $d = 1$ and is a rank $2n$ correction when $d = 2$ with $n = m$. We introduce the notion of *approximating class of sequences* (a.c.s.) and we give a theorem for dealing with this concept (see [5–7]). Suppose a sequence of matrices $\{A_k\}$ of increasing size d_k is given. We say that $\{\{B_{k,m}\} : m \in \mathbb{N}^+\}$, $B_{k,m}$ of size d_k, is an approximating class of sequences (a.c.s.) for $\{A_k\}$ if, for all sufficiently large $m \in \mathbb{N}$, the following splitting holds: $A_k = B_{k,m} + R_{k,m} + N_{k,m}$ for all $k > k_m$, with rank $R_{k,m} \le d_k\, c(m)$, $\|N_{k,m}\| \le \omega(m)$, where $\|\cdot\|$ is the spectral norm (maximal singular value), k_m, $c(m)$ and $\omega(m)$ depend only on m and, moreover, $\lim_{m\to\infty} \omega(m) + c(m) = 0$.

Theorem 2. [7] *Let $\{\{B_{k,m}\}, m \in \mathbb{N}^+\}$ be an a.c.s. for $\{A_k\}$ ($A_k \in M_{d_k}(\mathbb{C})$) such that $E_{k,m} = N_{k,m} + R_{n,m}$, $B_{k,m}$ are Hermitian, d_k is increasing with k, and $\{B_{k,m}\} \sim_\lambda (h_m, G)$, $0 < \mu(G) < \infty$, $\lim_{m\to\infty} h_m = h$ in measure on G, with $\sup_m \sup_k \|B_{k,m}\| = \widetilde{C}$, $\sup_m \sup_k \|E_{k,m}\| = \widehat{C}$, where $\widetilde{C}$, $\widehat{C}$ are positive constants. Moreover, $\|E_{k,m}\|_1 \le c(m) d_k$ with $c(m) \xrightarrow[m\to\infty]{} 0$ ($\|\cdot\|_1$ being the trace norm, see [2]). Then h is real valued and $\{A_k\} \sim_\lambda (h, G)$.*

Denoting by B_n the rank 2 correction matrix associated with the boundary conditioning, we find $\|B_n\|_1 = \sqrt{\lambda_1} + \sqrt{\lambda_2}$, where λ_1, λ_2 are the only nonzero eigenvalues of $B_n B_n^T$. For all the RBF here considered, it is possible to prove that the spectral norm of B_n is infinitesimal with respect to n. Moreover, the numerics inform us that $\|B_n\|_1 \sim 1/n^\alpha$ where $\alpha = 1/2$ for the MQ RBF and $\alpha = 2$ when considering IMQ and Gaussian RBF. We just observe that the very same calculations give the desired bounds also for $d = 2$. Therefore Theorem 2 leads to $\{A_n\} \sim_\lambda (s, Q)$ and $\{A_{mn}\} \sim_\lambda (s, Q^2)$, respectively, that is the same spectral distribution of the Toeplitz counterparts.

References

1. Belytschko T., Krongauz Y., Organ D., Fleming M., Krysl P.: Meshless methods: An overview and recent developments. Comput. Methods Appl. Mech. Eng. **139**, 3–47 (1996)
2. Bhatia R.: Matrix Analysis. Springer, New York (1997)
3. Bini D.A., De Rossi A., Gabutti B.: On certain (block) Toeplitz matrices related to radial functions. Linear Algebra Appl. **428**, 508–519 (2008)
4. Böttcher A., Silbermann B.: Introduction to Large Truncated Toeplitz Matrices. Springer, New York (1999)
5. Golinskii L., Serra-Capizzano S.: The asymptotic properties of the spectrum of non symmetrically perturbed Jacobi matrix sequences. J. Approx. Theory **144**, 84–102 (2007)
6. Serra-Capizzano S.: Distribution results on the algebra generated by Toeplitz sequences: A finite dimensional approach. Linear Algebra Appl. **28**, 121–130 (2001)
7. Serra-Capizzano S., Sesana D.: Tools for the eigenvalue distribution in a non-Hermitian setting. Linear Algebra Appl. **430**, 423–437 (2009)
8. Tilli P.: A note on the spectral distribution of Toeplitz matrices. Linear Multilin. Algebra **45**, 147–159 (1998)
9. Wendland H.: Scattered data approximation. Cambridge Monogr. Appl. Comput. Math., vol. 17, Cambridge University Press, Cambridge (2005)

Finite Element Solution of the Primitive Equations of the Ocean by the Orthogonal Sub-Scales Method

Tomás Chacón Rebollo, Macarena Gómez Mármol, and Isabel Sánchez Muñoz

Abstract This paper introduces a method for the numerical solution of steady Primitive Equations of the Ocean. This is an adaptation of the Orthogonal Subscales – Variational Multiscale Method, using conforming finite elements. We choose this method on one hand because it is a stabilized method, thus providing a low-cost and accurate discretization. On another hand, because it also is a LES turbulence model, so no further turbulence modeling is needed. We perform a numerical analysis of stability and convergence by means of representation of stabilizing terms in spaces of bubble finite elements. In particular, we give an original proof of the inf-sup condition to estimate the surface pressure. We present some numerical experiments for 2D flows that confirm the theoretical expectations.

1 Introduction

The turbulent nature of oceanic flows at large space scales makes necessary the derivation of specific turbulence models (Cf. [1]) to perform their numerical simulation. Standard LES (Large-Eddy Simulation) numerical turbulence models are based upon two modeling steps: Derivation of continuous turbulence model, and numerical discretization of these models. The first step is performed by some kind of averaging of turbulence effects, without a clear meaning of the nature of the mean flow modeled. Moreover, many of these models (e.g., two-equations turbulence models, frequently used), are mathematical objects more singular than the Navier–Stokes

T.C. Rebollo (✉) and M.G. Mármol
Dpto. de Ecuaciones Diferenciales y Análisis Numérico, Universidad de Sevilla.
Facultad de Matemáticas. C/ Tarfia, s/n. 41012 Sevilla, Spain
e-mail: chacon@us.es, macarena@us.es

I.S. Muñoz
Dpto. de Matemática Aplicada I, Escuela Universitaria de Ingeniería Técnica de Agrícola. Ctra. de Utrera, Km. 1. 41013 Sevilla, Spain
e-mail: isanchez@us.es

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_25,

equations. So, their numerical simulation is subject to severe stability restrictions, requiring frequent computing tricks (Cf. http://www.gotm.net/index.php).

The Orthogonal Subscales (OSS) Method (Cf. [3]) is a method of the family of Variational Multiscale Method, recently introduced, that is proving to provide a direct numerical modeling of turbulence, with clear conditions of applicability, and clear meaning of the numerical solution provided.

We address in this paper the numerical solution of the Primitive Equations of the Ocean by OSS Method. The Primitive Equations are one of the standard models in Geophysics for oceanic flows at large space scales. This is a first step towards the testing of the abilities of this modeling to simulate the turbulence effects in Oceanic flows.

In this paper we analyze how to apply the OSS modeling to the linearized Primitive Equations, and propose a model. We next analyze the stability and accuracy of this model. We are considering general domains with depth possibly vanishing, so we cannot apply the regularity analysis performed by Ziane, Titi and co-workers (Cf. [4,5]).

The paper is organized as follows: The Primitive Equations are introduced in Sect. 2. The numerical approximation by OSS is derived in Sect. 3, and its stability and convergence analysis is presented in Sect. 4. Finally, some numerical tests for 2D flows, with good agreement with the theoretical expectations, are presented in Sect. 5

2 Continuous Problem

We shall consider the steady linearized Primitive Equations in d space dimensions ($d = 2$ or 3). Let us consider a depth function $D > 0$, defined on a (d-1)-dimensional bounded domain ω. Let us consider the domain

$$\Omega = \{(x,z) \in \mathbb{R}^d, \quad x \in \omega, \quad -D(x) < z < 0\}, \quad \omega \subset \mathbb{R}^{d-1}$$

with boundary $\partial\Omega = \Gamma_s \cup \Gamma_b$, where $\Gamma_s = \omega \times \{0\}$ is the sea surface, and $\Gamma_b = \partial\Omega \setminus \Gamma_s$ is the ocean bottom. Consider a velocity field $a : \bar{\Omega} \to \mathbb{R}^d$ with free divergence. We set the boundary problem

$$\begin{cases} \text{Obtain } (y, y_v) : \overline{\Omega} \mapsto \mathbb{R}^d & \text{Velocity} \\ \text{ and } P : \Omega \mapsto \mathbb{R} & \text{Pressure} \\ a \cdot \nabla y - \mu \Delta y + \nabla_H P = f & \text{in } \Omega \\ \partial_v P = -\rho g & \text{in } \Omega \\ \nabla \cdot (y, y_v) = 0 & \text{in } \Omega \\ y|_{\Gamma_b} = 0, \quad \mu \dfrac{\partial y}{\partial n}|_{\Gamma_s} = \tau_w, \quad y_v \cdot n_v|_{\Gamma_b} = 0, \quad y_v|_{\Gamma_s} = 0. \end{cases}$$

where f is a source term, and τ_w is the surface wind tension. For simplicity, we include in f effects due to Coriolis force and density variations. The full Primitive Equations, including those terms, have the same mathematical nature of the one considered. So, we are considering a simplified model of linearized Primitive Equations, that still include the main mathematical difficulties. In the model above, in addition to the source term, the flow is also forced by the surface wind stress. We include the "rigid-lid" assumption: $y_v = 0$ at surface $z = 0$. This hypothesis is acceptable for large space scales.

We shall consider a reduced formulation, that only include the horizontal velocity and the surface pressure:

$$\begin{cases} \text{Obtain } y : \overline{\Omega} \mapsto \mathbb{R}^{d-1} & \text{Horizontal velocity} \\ \text{ and } p : \omega \mapsto \mathbb{R} & \text{Surface pressure} \\ a \cdot \nabla y - \mu \Delta y + \nabla_H p = f & \text{in } \Omega \\ \nabla_H \cdot < y >= 0 & \text{in } \omega \\ y|_{\Gamma_b} = 0, \quad \mu \dfrac{\partial y}{\partial n}|_{\Gamma_s} = \tau_w & \end{cases} \tag{1}$$

where

$$< y > (x) = \int_{-D(x)}^{0} y(x,s) ds.$$

The condition $\nabla_H \cdot < y >= 0$ is equivalent to $y_v = 0$ at surface. Mathematically, it is a restriction similar to the free divergence, whose associated Lagrange multiplier is the surface pressure. The full pressure and the vertical velocity may be recovered by $P(x,z) = p(x) + g \int_0^z \rho(x,s) ds, \quad y_v(x,z) = \int_z^0 \nabla_H \cdot y(x,s)\, ds.$

When $a \in H^1(\Omega)^{d-1} \times L^2(\Omega)$, $f \in L^2(\Omega)$ and $\tau_w \in H^{1/2}(\Gamma_s)$, problem (1) admits at least a weak solution $(y, p) \in H_b^1(\Omega)^{d-1} \times L_0^{3/2}(\Omega, \partial_3)$, satisfying

$$B(y, p; w, q) = F(v), \quad \forall (w, q) \in \mathbf{W}_b^{1,3}(\Omega)^{d-1} \times L_0^2(\Omega, \partial_3) \tag{2}$$

with $B(y, p; w, q) = (a \cdot \nabla y, w) + \mu(\nabla y, \nabla w) - (\nabla_H \cdot w, p) + (\nabla_H \cdot y, q)$, $F(v) = (f, v) + (\tau_w, w)_{\Gamma_s}$, where the spaces are defined by

$$H_b^1(\Omega) = \{y \in H^1(\Omega), \quad y|_{\Gamma_b} = 0\}, \quad \mathbf{W}_b^{1,3}(\Omega)^{d-1} = \{w \in W^{1,3}(\Omega), w|_{\Gamma_b} = 0\},$$
$$L_0^r(\Omega, \partial_3) = \{q \in L^r(\Omega), \quad \partial_3 q = 0, \quad \int_\Omega q = 0\}.$$

Equation (2) is a Petrov–Galerkin formulation where the test functions have more regularity than the unknown. This is due to the lack of regularity of the vertical velocity. The solution depends continuously on the data:

$$\nu \|\nabla y_h\|_{L^2(\Omega)} + \|p\|_{L^{3/2}(\Omega)} \leq (1 + \|a\|_{L^d(\Omega)})(\|f\|_{L^2(\Omega)} + \|\tau_w\|_{H^{1/2}(\Gamma_s)}),$$

3 Discrete Problem

We derive our discretization by the Variational Multiscale procedure. To describe it, let us introduce the condensed notation $u = (y, p)$, $v = (w, q)$, $U = H_b^1(\Omega)^{d-1} \times L_0^{3/2}(\Omega, \partial_3)$, $V = \mathbf{w}3^{d-1} \times L_0^2(\Omega, \partial_3)$. Consider a discrete subspace U_h of U, and the decomposition $U = U_h \bigoplus \widetilde{U}$, $V = U_h \bigoplus \widetilde{V}$, where $\widetilde{U} = U_h^{\perp} \cap U$ and $\widetilde{V} = U_h^{\perp} \cap V$ are the sub-scale spaces ($U_h^{\perp}$ is the orthogonal space to U_h in $L^2(\Omega)^d$). We decompose $u = u_h + \tilde{u}$ and $v = v_h + \tilde{v}$ with obvious meaning. Formulation (2) is equivalent to

$$(P) \begin{cases} B(u_h, v_h) + B(\tilde{u}, v_h) = L(v_h), & \forall v_h \in U_h, \\ B(\tilde{u}, \tilde{v}) = L(\tilde{v}) - B(u_h, \tilde{v}), & \forall \tilde{v} \in \widetilde{V} \end{cases} \tag{3}$$

The general VMS procedure consists in approximately solving the equation for $\tilde{u}$, and inserting the solution in the first one. This provides a modeled equation for the resolved scales u_h. To approximate the second equation, the standard VMS procedure writes it as $\langle \mathscr{L}(\tilde{u}), \tilde{v} \rangle = \langle R(u_h), \tilde{v} \rangle$, where $\mathscr{L}$ is the linear operator from U onto V' associated to the variational formulation (2), $R(u_h) = f - \mathscr{L} u_h$ is the residual associated to u_h and the notation $< \cdot >$ stands for the duality between $\widetilde{V}'$ and $\widetilde{V}$. If this problems admits a solution, $\tilde{u}$ is a function of $R(u_h)$.

To derive the OSS modeling of subscales to our Petrov-Galerkin framework, consider a triangulation $\mathscr{T}_h$ of Ω. We approximate $\widetilde{U} \simeq \sum_{K \in \mathscr{T}_h} \widetilde{U}_K$, where $\widetilde{U}_K$ is a subspace of $\widetilde{U}$ formed by functions vanishing on ∂K, and similarly approximate $\widetilde{V}$. Further, assume that $\mathscr{L}(\widetilde{u_K})$ (we denote $\tilde{u}_K = \tilde{u}_{|_K}$), and $R(u_h)_{|_K}$ have $L^2(K)$ regularity. Next, approximate $\mathscr{L}$ restricted to $\widetilde{V}_K$ by a diagonal operator, so $\mathscr{L}(\tilde{u}_K) \simeq \lambda_K \tilde{u}_K$, where λ_K is a $d \times d$ diagonal matrix. Then the second equation in (3) yields

$$\left(\tilde{u}_K - \tau_K R(u_h)_{|_K}, \tilde{v}_K\right)_{L^2(K)} = 0, \quad \forall \tilde{v}_K \in \widetilde{V}_K, \quad \text{where} \quad \tau_K = [\lambda_K]^{-1}.$$

Then,

$$(\tilde{u} - \tau R(u_h), \tilde{v})_{L^2(\Omega)} = 0, \quad \forall \tilde{v} \in \hat{V} = \left[\sum_{K \in \mathscr{T}_h} \widetilde{V}_K \right] \bigcap U_h^{\perp},$$

where τ is the piecewise constant function that takes the value τ_K on K of $\mathscr{T}_h$. Space $\hat{V}$ is dense in $U_h^{\perp}$, so we deduce $\Pi_{U_h^{\perp}} (\tilde{u} - \tau R(u_h)) = 0$, where $\Pi_{U_h^{\perp}}$ denotes the $L^2(\Omega)$ orthogonal projection on $U_h^{\perp}$. But $\tilde{u} \in \widetilde{U}$ as $\widetilde{U} = U \cap U_h^{\perp}$. Consequently, $\Pi_{U_h^{\perp}}(\tilde{u}) = \tilde{u}$, and

$$\tilde{u} = \Pi_{U_h^{\perp}} (\tau R(u_h)) .$$

Next observe that $B(\tilde{u}, v_h) = \langle \mathscr{L}^*(v_h), \tilde{u} \rangle$, where $\mathscr{L}^*$ is the adjoint operator of $\mathscr{L}$, and $\langle \cdot \rangle$ now stands for the duality between $\widetilde{U}'$ and $\widetilde{U}$. As $\mathscr{L}^*(v_h)$ is elementwise smooth, we may approximate

$$B(\tilde{u}, v_h) \simeq \sum_K (L^*(v_h), \tau_K \, \Pi_{U_h^\perp}(R_h)).$$

We arrive so to the modeled equation for u_h

$$(P_h) \begin{cases} \text{Obtain } (y_h, p_h) \in Y_h \times N_h \text{ such that} \\ B(y_h, p_h; v_h, q_h) + (a \cdot \nabla v_h - \nu \Delta v_h + \nabla_H q_h, \Pi_{U_h^\perp}(a_h \cdot \nabla y_h - \nu \Delta y_h + \nabla_H p_h))_\tau \\ = F(v_h) + (a_h \cdot \nabla v_h - \nu \Delta v_h + \nabla_H q_h, \Pi_{U_h^\perp}(f))_\tau, \quad \forall (v_h, q_h) \in Y_h \times N_h \end{cases}$$

Here, we have assumed $U_h = Y_h \times N_h$, where Y_h is a FE subspace of $\mathbf{W}_b^{1,3}(\Omega)^{d-1}$ (3D Horizontal Velocities) and N_h is a FE subspace of $L_0^2(\Omega, \partial_3)$ (2D Surface Pressures), and $(\cdot, \cdot)_\tau$ stands for the scalar product defined by

$$(a, b)_\tau = \sum_{K \in \mathscr{T}_h} \tau_K \, (a, b)_{L^2(K)}.$$

The *stabilization coefficients* τ_K may be calculated by dimensional analysis or by Fourier analysis (Cf. [3]).

4 Stability and Convergence Analysis

We next describe the main elements of the stability and convergence analysis of method (P_h) for piecewise affine F.E. The convergence is strong only for smooth enough solutions:

Theorem 1. *Assume that the triangulations $\{\mathscr{T}_h\}_{h>0}$ are uniformly regular, $a \in L^d(\Omega)^d$ and $\tau_K = \mathscr{O}(h_K^2)$, then*

1. *The discrete problem admits a unique solution $(y_h, p_h) \in Y_h \times N_h$ which is bounded in $H_b^1(\Omega)^{d-1} \times L_0^{3/2}(\Omega, \partial_3)$. This solution satisfies the estimates*

$$\nu \|\nabla y_h\|_{L^2(\Omega)} + \|p_h\|_{L_0^{3/2}(\Omega, \partial_3)} \leq C \, (1 + \|a\|_{L^d(\Omega)}) \, (\|f\|_{L^2(\Omega)} + \|\tau_w\|_{H^{1/2}(\Gamma_s)}).$$

2. *The sequence $\{(y_h, p_h)\}_{h>0}$ contains a subsequence which is weakly convergent in $H_b^1(\Omega)^{d-1} \times L_0^{3/2}(\Omega, \partial_3)$ to a solution of the continuous problem (2). If this solution belongs to $W^{1,3}(\Omega)$, then the convergence is strong.*

Proof. (Main elements) The stability follows in a standard way from the well-known inf-sup condition

$$C\,\|p_h\|_{L_0^{3/2}(\Omega,\partial_3)} \leq \sup_{v_h \in Y_h - \{0\}} \frac{(\nabla_H \cdot v_h,\, p_h)}{\|v_h\|_{W_b^{1,3}}} + \|\nabla_H\, p_h\|_\tau, \quad \forall\, p_h \in N_h.$$

The convergence of the terms in problem (P_h) that appear in (2) also is standard. To prove the convergence to zero of the stabilizing terms, we represent them on spaces of bubble functions by means of static condensation operators. We apply then the theory developed in [2]: The boundedness in H^1 norm of bubble functions representing these stabilizing terms implies their weak convergence to zero in H^1. In its turn, the boundedness follows from the estimate

$$\|\Pi_{U_h^\perp}(a \cdot \nabla y_h + \nabla_H\, p_h)\|_\tau \leq C(\|f\|_{L^2(\Omega)} + \|\tau_w\|_{H^{1/2}(\Gamma_s)}).$$

This analysis can be extended in a straightforward manner to the non-linear Primitive Equations if $d = 2$. The extension of it to $d = 3$ is in progress.

5 Numerical Results

We have tested our numerical model for the 2D Primitive Equations. We have solved the discrete problem through an evolution approach, by means of the linearized equations

$$\begin{cases} \frac{1}{\Delta t}(y^{n+1} - y^n) + y^n \cdot \nabla y^{n+1} - \mu \Delta y^{n+1} + \partial_x p^{n+1} = f \quad \text{in } \Omega \subset \mathbb{R}^2 \\ \partial_x < y^{n+1} >= 0 \quad \text{in } \omega \subset \mathbb{R} \\ y^0 = 0, y^{n+1}|\Gamma_b = 0, \quad \mu \frac{\partial y^{n+1}}{\partial n}|_{\Gamma_s} = g. \end{cases}$$

We have solved this problem for piecewise affine finite elements for both velocity and pressure, using the application FreeFem++.

Test 1: Convergence. To test the convergence order of OSS method (P_h), we have set $\Omega = (0,3)\times(-1,0)$, $\mu = 0,5$ and have taken f, g to obtain the continuous solution

$$y = \left((x_2{+}1)(x_1 - 3)(3x_2{+}1)/\exp(x_2), x_2(-5x_1{+}3 + x^2)(x_2{+}1)2/\exp(x_1)\right),$$

$$p = \exp(x_1).$$

In Tables 1 and 2 we present the estimated convergence orders for the horizontal velocity and pressure, and a comparison with the errors obtained with the (P1+Bubble, P1) discretization. We observe a better accuracy with OSS method, and a super-convergence effect, possibly due to the structured nature of the grids used.

Test 2: Convex and non-convex geometries. We finally have tested the overall characteristics of the computed flow, for convex and non-convex domains. The convex case corresponds to $\Omega = (0,5) \times (-1,0)$ and the data $\mu = 0.5$, $f = 0$,

Table 1 Estimated convergence orders for horizontal velocity

	Horizontal velocity		
h	P1b-P1	OSS	Order in H^1-norm
0.072	0.0586436	0.00369733	
0.036	0.0283502	0.00101718	1.81623
0.018	0.013938	0.00314393	1.67336
0.014	0.011539	0.000211092	1.5564

Table 2 Estimated convergence orders for surface pressure

	Pressure		
h	P1b-P1	OSS	Order in L^2-norm
0.072	0.000932524	0.00045671	
0.036	0.000327411	0.00123518	1.84027
0.018	0.000115275	3.7988e-5	1.68045
0.014	7.65232e-5	2.51929e-5	1.60469

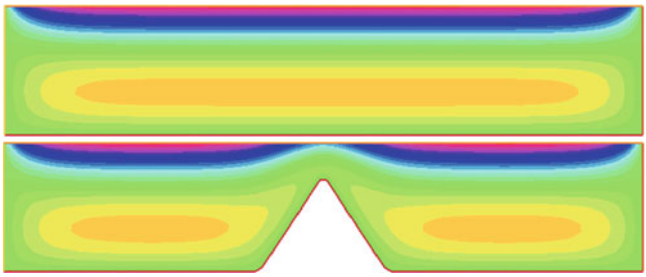


Fig. 1 Horizontal velocity. (**a**) Convex domain (**b**) non-convex domain

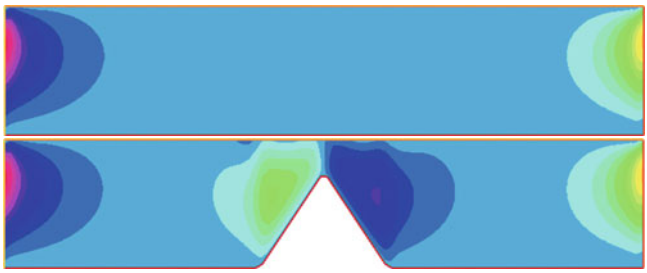


Fig. 2 Vertical velocity. (**a**) Convex domain (**b**) non-convex domain

$g(x) = 1$. The non-convex domain is a deformation of this rectangular Ω, in order to simulate an underwater mountain. We respectively present in Fig. 1a, b the horizontal velocities for the convex and non-convex domains, and in Fig. 2a, b the vertical velocities. In the first case we observe an overall circulation that occurs because the horizontal velocity vanishes in the vertical walls. In the second one we observe a the formation of two large vortex, one on each side of the mountain,

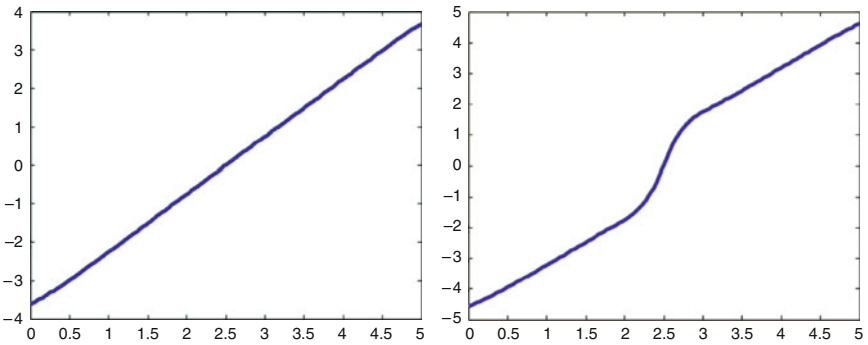


Fig. 3 Pressure for (**a**) convex domain and (**b**) non-convex domain

and a shear layer over the peak of the mountain. Also, in Fig. 3a, b we present the profiles of pressures obtained in both computations. Both pressures are increasing, with a fast increase in the zone near the peak of the mountain. Both pressures are monotonic, without oscillations. Let us remark that the use of (P1+Bubble, P1) discretization for these cases presents some instabilities close to the outflow boundary, where the velocity gradients. These global patterns are physically satisfactory.

Acknowledgements Research partially supported by the Junta de Andaluca Research Project P07-FQM-02538.

References

1. Pond, S., Pickard, G. L. (1993). *Introductory Dinamical Oceanography, 3rd. Edition.* Elsevier, Oxford
2. Chacón Rebollo, T. (1998). A term by term stabilization algorithm for finite element solution of incompressible flow problems. *Numer. Math.* 79(2), 283–319
3. Codina, R. (2008). Analysis of a stabilized finite element approximation of the Oseen equations using orthogonal subscales. *Appl. Numer. Math.* 58(3), 264–283
4. Kukavika, I., Ziane, M. (2007). On the regularity of the primitive equations of the ocean. *Nonlinearity* 20, 2739–2753
5. Cao, C., Titi, I. (2007). Global well-posedness of the three-dimensional viscous primitive equations of large scale ocean and atmosphere dynamics. *Ann. Math.* 166(1), 245–267

Solution of Incompressible Flow Equations by a High-Order Term-by-Term Stabilized Method

Tomás Chacón Rebollo, Macarena Gómez Mármol, and Isabel Sánchez Muñoz

Abstract This paper introduces a high-order easy-to-implement stabilized method for the numerical solution of convection-diffusion and incompressible flows. We obtain stable discretizations of equal-order interpolation of velocity and pressure, as is usual for stabilized methods. The penalty terms have a projection structure that allows to obtain a high order method. We present stability analysis and error estimates results for Oseen equations with Neumann boundary data with general Finite Element discretizations, and for Dirichlet boundary data with **P2–P2** discretization and **P1** projection. We also present some numerical tests that confirm the theoretical expectations.

1 Introduction

The numerical solution of incompressible flow problems faces the un-stabilizing effects originated by the loss of high-frequence resolution. This effect typically appears in the discretization of pressure, but also of several operator terms that could become dominant in the discrete problem: Convection, rotation, reaction, etc. Consistent stabilized methods provide a solution to this problem in the framework of Finite Element discretizations. However, these methods become rather costly for high-order interpolation due to the complexity of the stabilizing term, which is proportional to the element-wise residual (Cf. [3]). Brezzi and Pittkaranta introduced in [1] a first-order penalty method to stabilize the discretization of pressure and allows to use equal-order interpolation of velocity and pressure. This idea was extended

T.C. Rebollo (✉) and M.G. Mármol
Dpto. de Ecuaciones Diferenciales y Análisis Numérico, Universidad de Sevilla. Facultad de Matemáticas. C/ Tarfia, s/n. 41012 Sevilla, Spain
e-mail: chacon@us.es, macarena@us.es

I.S. Muñoz
Dpto. de Matemática Aplicada I, Escuela Universitaria de Ingeniería Técnica de Agrícola. Ctra. de Utrera, Km. 1. 41013 Sevilla
e-mail: isanchez@us.es

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_26,

by Chacón [2] to the stabilization of single operator terms such as those mentioned above. This term-by-term stabilization method, however, still was only first-order accurate.

We introduce in this paper a high-order term-by-term stabilization method, which still is able to stabilize the discretization of pressure and single operator terms, but has the order of accuracy of the interpolation for velocity and pressure.

The structure of the paper is the following: Section 1 introduces the method to stabilize single operator terms in the framework of transport-diffusion equations. Section 2 applies it to Oseen equations, with either Neumann or Dirichlet boundary conditions. We prove stability and optimal error estimates, by means of specific inf-sup condition, that holds virtually for any pair of velocity-pressure spaces in the case of Neumann boundary conditions. We also prove that this condition holds for **P2–P2** discretization and **P1** projection in the case of Dirichlet boundary conditions. In Sect. 3 we present some numerical tests for Oseen equations with smooth solutions, that confirm the theoretical expectations of accuracy.

2 Transport-Diffusion

Let us consider the transport-diffusion equations in a bounded domain $\Omega \subset \mathbf{R}^d$, $d = 2$ or 3 with lipschitz boundary Γ:

$$\begin{cases} \text{Find } y : \Omega \mapsto I\!R & \text{such that} \\ u \cdot \nabla y - \nu \Delta y = f & \text{in } \Omega, \\ y = 0 & \text{on } \Gamma \end{cases}$$

Consider a linear bounded operator B from $H^1(\Omega)$ onto $H^{-1}(\Omega)$. This operator may represent convection: $By = u \cdot \nabla y$, rotation: $By = \nabla \times y$, reaction: $By = \alpha\, y$, etc. The discretization of By may originate spurious oscillations due to the lack of resolution in the discrete space. Our purpose is to devise a method that stabilizes some high-frequency components of the discrete By with high accuracy and low computational cost.

We consider the standard Finite Element space, on a triangulation $\mathcal{T}_h$ of Ω

$$V_h^{(l)} = \{r \in C^0(\overline{\Omega}) \text{ such that } r_{|K} \in P_l(K), \forall K \in \mathcal{T}_h\},$$

and set the discretization

$$\begin{cases} \text{Obtain } y_h \in V_h^{(l)}, & \text{such that} \\ L_h(y_h, v_h) =< f, v_h >, & \forall v_h \in V_h^{(l)}; \end{cases} \tag{1}$$

where $L_h(y, v) = a(y, v) + c_h(y, v)$,

$$a(y, v) = \int_\Omega u \cdot \nabla y\, v\, dx + \nu \int_\Omega \nabla y \cdot \nabla v dx;$$

$$c_h(y,v) = \int_\Omega \tau\,(I-\pi_h)(By)(I-\pi_h)(Bv)\,dx. \tag{2}$$

Here, π_h is a stable interpolation or projection (local or global) operator on the Finite Element space $Z_h = V_h^{(l-1)}$ (one degree less of interpolation than the space of unknowns), and $\tau_{|_K} = const. = \tau_K$. The $\tau_K \simeq h_K^2$ are the stabilization coefficients, assumed to be of order h_K^2. These coefficients are obtained by dimensional analysis, from local relevant parameters to the physical effect to be stabilized. Discretization (1) makes sense if $(Bv_h)_{|_K} \in L^2(K)$ for any $K \in V$, as we assume.

The stabilizing properties of scheme (1) is given by the following result:

Lemma 1. Stability *Problem (1) admits a unique solution that satisfies*

$$\|y_h\|_{H^1(\Omega)} + h\,\|By_h\|_{L^2(\Omega)} \le C\,\|f\|_{H^{-1}(\Omega)} \tag{3}$$

Proof. The proof of this lemma follows from two facts: On one hand, $\|\pi_h(By_h)\|_{L^2(\Omega)}$ is bounded because y_h is bounded and B is a bounded operator. On another hand, $\|(I-\pi_h)(By_h)\|_{L^2(\Omega)}$ is directly bounded from the structure of the discrete problem. The remaining of the proof standard and we shall omit it for brevity.

To explain the stabilizing effect of this procedure, observe that as B is bounded,

$$\|By_h\|_{H^{-1}(\Omega)} \le \|B\|\,\|f\|_{H^{-1}(\Omega)}$$

Then, when $\|B\|$ is large, this estimate degenerates. Opposite, estimate (3) yields

$$h\,\|By_h\|_{L^2(\Omega)} \le C\,\|f\|_{H^{-1}(\Omega)},$$

which is uniform with respect to $\|B\|$.

The high-order accuracy of scheme (1) is given by the

Theorem 1. Error estimates. *Assume that the family of triangulations $\mathcal{T}_h$ is regular. Then, if $y \in H^{l+1}(\Omega)$, then*

$$\|y-y_h\|_{H^1(\Omega)} + h\,\|(I-\pi_h)(By_h)\|_{L^2(\Omega)} \le C\,|u|_{H^{l+1}(\Omega)}\,h^l$$

This proof is again standard, using that the stabilizing coefficients are of order h^2. Notice that the high-frequency components of By_h are small, one order below the order of the interpolation on the finite element space Y_h.

3 Oseen Equations

We next consider the Oseen equations after time discretization as a model problem for incompressible flows, where we consider either Neumann or Dirichlet boundary

conditions:

$$\begin{cases} \text{Obtain } y : \overline{\Omega} \mapsto R^d & \text{Velocity field} \\ \text{ and } p : \Omega \mapsto R & \text{Pressure, such that} \\ u \cdot \nabla y - \mu \Delta y + \alpha y + \nabla p = f & \text{in } \Omega \\ \nabla \cdot y = 0 & \text{in } \Omega \\ -\mu \partial_n y + p\, n = g & \text{on } \Gamma \quad \textbf{or} \\ y = 0 & \text{on } \Gamma; \end{cases} \tag{4}$$

where $u \in H^1(\Omega)^d$ is a given velocity field with free divergence. For this problem, we not only intend to stabilize the discretization of the operator term By, but also, we need to stabilize the discretization of the pressure.

Consider the Finite Element spaces

$$V_h^{(l)} = \{r \in C^0(\overline{\Omega}) \text{ such that } r_{|K} \in P_l(K), \forall K \in \mathcal{T}_h\},$$
$$Y_h = (V_h^{(l)} \cap H_0^1(\Omega))^d, \text{ (for Dirichlet b. c.) } \textbf{or } Y_h = V_h^{(l)} \text{ (for Neumann b. c.);}$$
$$M_h = V^{(l)} \cap L_0^2(\Omega), \text{ (for Dirichlet b. c.) } \textbf{or } M_h = V^{(l)} \text{ (for Neumann b. c.).}$$

We discretize Oseen equations by:

$$\begin{cases} \text{Obtain } (y_h, p_h) \in Y_h \times M_h, & \text{such that} \\ L_h(y_h, p_h; v_h, q_h) =< f, v_h >, & \forall (v_h, q_h) \in Y_h \times M_h; \end{cases} \tag{5}$$

where $L_h(y, p; v, q) = a(y, v) - (p, \nabla \cdot v) - (\nabla \cdot y, q) + c_h(y, v) + d_h(p, q)$; where $a(y, v) = (u \cdot \nabla y, v) + \mu(\nabla y, \nabla v) + (\alpha y, v)$; c_h is a stabilizing term for By_h, with the same structure as (2), defined by

$$c_h(y, v) = \sum_K \tau_{1K}(\sigma_{1h}(By), \sigma_{1h}(Bv))_K,$$

with $\sigma_{1h} = I - \pi_{1h}$, where π_{1h} is a stable $O(h^{l-1})$ interpolation operator on Y_h; and d_h is a stabilizing term for the pressure gradient, defined by

$$d_h(p, q) = \sum_K \tau_{2K}(\sigma_{2h}(\nabla p), \sigma_{2h}(\nabla q))_K,$$

with $\sigma_{2h} = I - \pi_{2h}$, where π_{2h} is the $L^2(\Omega)$ orthogonal projection on some space Z_h. Again, τ_{1K} and τ_{2K} are the stabilizing coefficients, built by dimensional analysis. Concretely, τ_{2K} depends on h_K, and τ_{1K} on h_K, y_K, μ.

These terms must be compared to those corresponding to the stabilizing terms of the pure penalty stabilized method (Cf. [2]), given by $c_h(y, v) = \sum_K \tau_{1K}(By, Bv)_K$,

$d_h(y,v) = \sum_K \tau_{2K}(\nabla p, \nabla q)_K$. The gain in precision is due to the introduction of the residual interpolation operators σ_{1h} and σ_{2h}.

The Neumann boundary conditions are simpler to analyze. In this case, we obtain the following stability result:

Theorem 2. *(***Neumann boundary conditions***) Assume that the triangulations* $\{\mathcal{T}_h\}_{h>0}$ *are uniformly regular. Set* $Z_h = \left[V_h^{(l)}\right]^d$ *(Same interpolation as velocities and pressures). Then, problem* (P_h) *admits a unique solution that satisfies*

$$\|y_h\|_{H^1(\Omega)^d} + \|p_h\|_{L^2(\Omega)} + h\,\|By_h\|_{L^2(\Omega)^d} \leq C\,\|f\|_{H^{-1}(\Omega)}. \tag{6}$$

Proof. (**Sketch**) The main technical difficulty in this proof is the derivation of the inf-sup condition to estimate the pressure. We start from the Verfurth inf-sup condition: There exists $\beta > 0$ such that

$$\beta\,\|q_h\|_{L^2(\Omega)} \leq \sup_{v_h \in Y_h} \frac{(\nabla \cdot v_h, q_h)}{\|v_h\|_{H^1(\Omega)^d}} + h\,\|\nabla q_h\|_{L^2(\Omega)}, \quad \forall q_h \in M_h.$$

To estimate of the second summand, we split

$$\|\nabla q_h\|_{L^2(\Omega)} \leq \|\sigma_{2h}(\nabla q_h)\|_{L^2(\Omega)} + \|\pi_{2h}(\nabla q_h)\|_{L^2(\Omega)}. \tag{7}$$

The first summand is directly bounded by the discrete method (5). To estimate the second one, taking $v_h = \pi_{2h}(\nabla q_h)$ we obtain

$$\|\pi_{2h}(\nabla q_h)\|_{L^2(\Omega)} \leq C \sup_{v_h \in Y_h} \frac{(\nabla \cdot v_h, q_h)}{\|v_h\|_{H^1(\Omega)^d}}.$$

The uniform regularity of the triangulations is needed here, to estimate $\|\pi_{2h}(\nabla q_h)\|_{H^1(\Omega)^d}$ in terms of $\|\pi_{2h}(\nabla q_h)\|_{L^2(\Omega)^d}$ Then, there exists $\beta' > 0$ such that

$$\beta'\,\|q_h\|_{L^2(\Omega)} \leq \sup_{v_h \in Y_h} \frac{(\nabla \cdot v_h, q_h)}{\|v_h\|_{H^1(\Omega)^d}} + h\,\|\sigma_{2h}(\nabla q_h)\|_{L^2(\Omega)}. \tag{8}$$

The two summands in the r.h.s of this inequality are directly bounded by the discrete method (5). This allows to estimate the pressure. The remaining terms in estimate (6) are obtained in a standard way.

This stability result allows to obtain the following error estimates, again in a standard way:

Theorem 3. *Assume that the triangulations* $\{\mathcal{T}_h\}_{h>0}$ *are uniformly regular. Assume* $l = 2$*, and set* $Z_h = \left[V_h^{(l-1)}\right]^d$*. Assume also* $y \in H^{l+1}(\Omega)^d$*,* $p \in H^l(\Omega)$*,* $B \in \mathcal{L}\left(H^{l+1}(\Omega), H^{l-1}(\Omega)\right)$*. Then, the following error estimates hold*

$$\|y - y_h\|_{H^1(\Omega)^d} + \|p - p_h\|_{L^2(\Omega)} + \|By_h\|_{L^2_\tau(\Omega)^d} \leq C\, h^l\, (|y|_{H^{l+1}} + |p|_{H^l}).$$

The Dirichlet boundary conditions case is more complex, because now the derivation of the inf-sup condition is limited by the homogeneous boundary values of the velocity. Indeed, to obtain the estimate

$$\|\nabla q_h\|_{L^2(\Omega)} \leq \|\sigma_{2h}(\nabla q_h)\|_{L^2(\Omega)} + \|\pi_{2h}(\nabla q_h)\|_{L^2(\Omega)},$$

we can no longer take $v_h = \pi_{2h}(\nabla q_h)$. However, we have been able to obtain a similar result when we use **P2–P2** interpolation for velocity-pressure:

Theorem 4. *Assume that the triangulations $\{\mathcal{T}_h\}_{h>0}$ are uniformly regular. Assume $l = 2$, and set $Z_h = \left[V_h^{(1)}\right]^d$. Assume also $y \in H^3(\Omega)^d$, $p \in H^2(\Omega)$. Then, the following error estimates hold*

$$\|y - y_h\|_{H^1(\Omega)^d} + \|p - p_h\|_{L^2(\Omega)} + \|By_h\|_{L^2_\tau(\Omega)^d} \leq C\, h^2\, (|y|_{H^3} + |p|_{H^2}).$$

Proof. In this case, the inf-sup condition (8) still holds, but its deduction is more involved. To do it, we first prove the reduced inf-sup condition that follows: there exists $\gamma > 0$ such that

$$\gamma\, \|Q_h\|_{L^2(\Omega)^d} \leq \sup_{v_h \in Y_h} \frac{(v_h, Q_h)}{\|v_h\|_{L^2(\Omega)^d}}, \quad \forall Q_h \in \left[V_h^{(1)}\right]^d. \tag{9}$$

Then, for any $q_h \in M_h$,

$$\begin{aligned} h\,\gamma\, \|\pi_{2h}(\nabla q_h)\|_{L^2(\Omega)^d} &\leq h \sup_{v_h \in Y_h} \frac{(v_h, \pi_{2h}(\nabla q_h))}{\|v_h\|_{L^2(\Omega)^d}} \\ &\leq h \left[\sup_{v_h \in Y_h} \frac{(v_h, \nabla q_h)}{\|v_h\|_{L^2(\Omega)^d}} + \sup_{v_h \in Y_h} \frac{(v_h, \sigma_{2h}(\nabla q_h))}{\|v_h\|_{L^2(\Omega)^d}} \right] \\ &\leq C \sup_{v_h \in Y_h} \frac{(\nabla \cdot v_h, q_h)}{\|v_h\|_{H^1(\Omega)^d}} + h\, \|\sigma_{2h}(\nabla q_h)\|_{L^2(\Omega)}. \end{aligned}$$

In the last estimate we have again used the uniform regularity of the grids to estimate $\|v_h\|_{H^1(\Omega)^d}$ in terms of $\|v_h\|_{L^2(\Omega)^d}$. This estimate, combined with (7), yields (8). Again, the remaining of the proof is standard.

4 Numerical Tests

We have tested the order of accuracy of our method for Stokes equations with Dirichlet boundary conditions, with smooth solutions in the unit square of $\mathbf{R}^2$, using structured grids. The order of accuracy has been estimated by means of the

Table 1 Computed convergence orders in natural norms for P_2-P_2-P_1 discretization

(N1,N2)	Velocity	Pressure
(40,60)	2.927	2.688
(60,80)	2.928	2.226
(80,120)	2.893	1.97

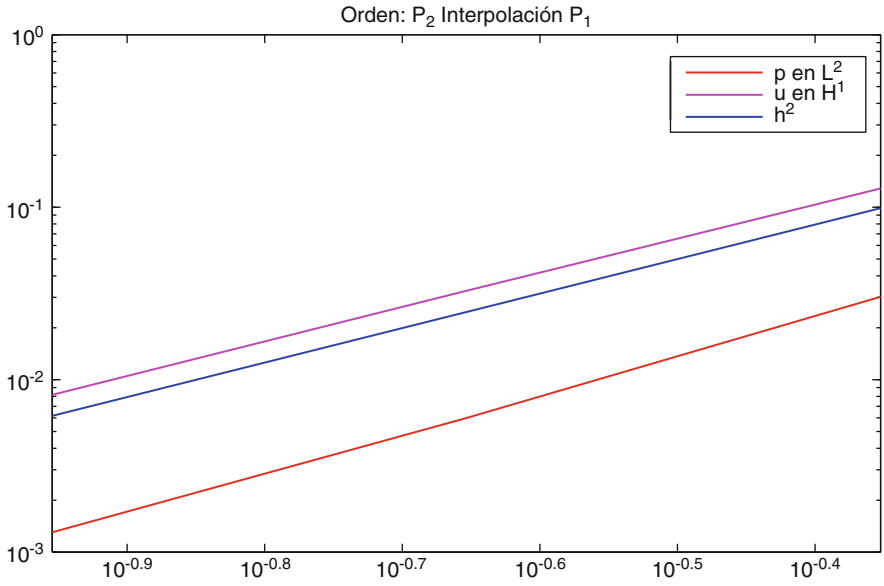


Fig. 1 Estimated convergence order for Oseen problem with P_2–P_2–P_1 discretization

results of two grids. Table 1 presents the estimated convergence orders for P_2-P_2-P_1 discretization when the interpolation operator is the orthogonal L^2 projection. We estimate the convergence orders in the natural norms: H^1 for velocity and L^2 for the pressure. We may observe the expected accuracy for the pressure, but some superconvergence for velocity, possibly due to the structured nature of the grids. We also present in Fig. 1 the estimated convergence orders, in log-log coordinates, for the same test on non-structured grids. We observe here that the second order accuracy is clearly reached. We finally present in Fig. 2 the estimated convergence orders of the same test for P_3-P_3-P_2 discretization, on non-structured grids. In this case we recover a third order convergence, although by now we do not have a theoretical support for these results.

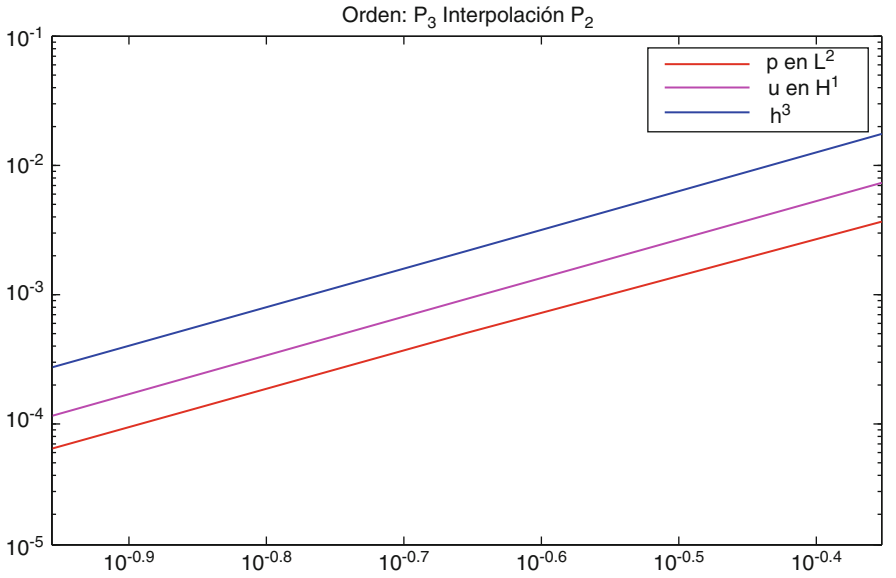


Fig. 2 Estimated convergence order for Oseen problem with P_3-P_3-P_2 discretization

Acknowledgements Research partially funded by the Spanish Ministerio de Educación y Ciencia Research Project MTM 2009-07719.

References

1. Brezzi, F.; Pitkaranta, J. (1984). On the stabilization of finite element approximations of the Stokes equations, in *Efficient solutions of elliptic systems.* Kiel: Vieweg
2. Chacón Rebollo, T. (1998). A term by term stabilization algorithm for finite element solution of incompressible flow problems. *Numer. Math.* Volume 79, Number 2, 283–319
3. Codina, R. (2008). Analysis of a stabilized finite element approximation of the Oseen equations using orthogonal subscales. *Appl. Numer. Math.* Volume 58, Issue 3, 264–283

Solving Large Sparse Linear Systems Efficiently on Grid Computers Using an Asynchronous Iterative Method as a Preconditioner

T.P. Collignon and M.B. van Gijzen

Abstract This paper describes an efficient iterative algorithm for solving large sparse linear systems on Grid computers. The algorithm is a combination of a synchronous flexible outer iterative method and a coarse-grain asynchronous inner iterative method as a preconditioner. The preconditioning iteration is performed on heterogeneous computing hardware. We present experimental results on a heterogeneous computing grid of a complete implementation using GridSolve as middleware for a 3D convection–diffusion problem.

1 Introduction

In this paper we present an efficient iterative method for solving large linear systems that is designed to exploit the characteristics of Grid computing. The algorithm is a combination of the *flexible* iterative method GMRESR [7] and an asynchronous iterative method [2] as *preconditioner*. The preconditioning iteration is performed on heterogeneous computational hardware and as a result, the preconditioner varies in each outer iteration step. We therefore use a flexible method such as GMRESR, which can handle a varying preconditioner.

Since asynchronous iterative methods are fault-tolerant, can adapt to the computational environment, and lack global synchronisation points, they are naturally suited to Grid computing. However, the slow block Jacobi-like convergence rate of these methods limits the practical applicability [2, 4]. By using an asynchronous method as a coarse-grain preconditioner in a flexible iterative method, we can improve overall convergence rates and extend the range of applications.

The inner–outer algorithm is implemented using the Grid middleware GridSolve [5, 8], which allows for a decoupling of the two iteration processes. The

T. P. Collignon (✉) and M.B. van Gijzen
Delft University of Technology, Delft Institute of Applied Mathematics, Mekelweg 4, 2628 CD, Delft, The Netherlands
e-mail: t.p.collignon@tudelft.nl, m.b.vangijzen@tudelft.nl

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_27,

outer iteration is performed sequentially on the (stable) client machine, while the inner preconditioning iteration is performed on (unstable) heterogeneous computing hardware. Since global synchronisation is a highly expensive operation, the bulk of the computational work is performed by the asynchronous preconditioning iteration. In this way, efficient use is made of the available computational resources. For completeness, we have also evaluated a parallel implementation of the outer iteration.

Algorithm 1 GMRESR (truncated version)

Require: Parameters $m, \epsilon_{\text{in}}, T_{\max}$; Initial guess x_0; Set $r_0 = b - Ax_0$.
Ensure: Approximate solution to $Ax = b$.

1: **for** $k = 0, 1, \ldots,$ until convergence **do**
2: Evaluate $u = M(r_k, \epsilon_{\text{in}}, T_{\max})$; {Preconditioning step: see Alg. 3}
3: Compute $c = Au$; {Matrix–vector multiplication}
4: Compute $[c_k, u_k] = \text{orthonorm}(c, u, c_i, u_i, k, m)$; {Orthogonalisation step}
5: Compute $\gamma = c_k^\top r_k$;
6: Update $x_{k+1} = x_k + \gamma u_k$;
7: Update $r_{k+1} = r_k - \gamma c_k$;
8: **end for**

Algorithm 2 (A–)synchronous block Jacobi iteration on p processors

1: Initialize $u^{(0)}$;
2: **for** $k = 1, 2, \ldots,$ until convergence **do**
3: **for** $i = 1, 2, \ldots, p$ **do**
4: (i.) Solve $A_{ii} u_i^{(k)} = r_i - \sum_{j=1, j\neq i}^{p} A_{ij} u_j^{(k-1)}$; {synchronous iterations}
5: (ii.) Solve $A_{ii} u_i^{\text{new}} = r_i - \sum_{j=1, j\neq i}^{p} A_{ij} u_j^{\text{old}}$; {a-synchronous iterations}
6: **end for**
7: **end for**

Numerical experiments for a large 3D convection–diffusion problem demonstrate the effectiveness of the algorithm.

2 Sparse Linear Solvers in Grid Environments

We are interested in designing efficient iterative methods for solving large sparse linear systems,

$$Ax = b, \quad \text{with a non-symmetric, non-singular matrix } A, \tag{1}$$

on heterogeneous networks of computers. Parallel asynchronous iterative methods possess several characteristics that are perfectly suited for Grid computing, such as lack of synchronisation points [1]. Unfortunately, they also have significant drawbacks, such as slow convergence rates [2]. We propose to use an asynchronous method as a coarse-grain preconditioner in a flexible iterative method. By using a slowly converging asynchronous method as a preconditioner in a fast converging flexible method we expect to achieve overall fast convergence.

We choose GMRESR as a flexible method, partly because the orthogonalisation process can be easily truncated, which is essential for practical implementations. The truncated variant of GMRESR is shown in Alg. 1. The preconditioning step in the second line consists of computing some approximate solution to $Au = r_k$ using an asynchronous iterative method. The obtained search direction is then orthogonalised in line 4 against m previous search directions.

Asynchronous algorithms generalise simple iterative methods such as the classical block Jacobi iteration. To compute an approximation to $Au = r$ using p processors, the coefficient matrix, the solution vector, and the right-hand side are partitioned into blocks as follows:

$$A = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1p} \\ A_{21} & A_{22} & \cdots & A_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ A_{p1} & A_{p2} & \cdots & A_{pp} \end{bmatrix}, \quad u = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_p \end{bmatrix}, \quad r = \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_p \end{bmatrix}. \tag{2}$$

In the standard synchronous Jacobi iteration process (see line 4 of Alg. 2), the processors operate in parallel on their part of the vector $u^{(k)}$, followed by a synchronisation point at each iteration step k. In our asynchronous algorithm (see line 5 of Alg. 2), a processor computes u^{new} using information u^{old} that is available on the process at that particular time. As a result, each separate block Jacobi iteration process may use out-of-date information, but the lack of synchronisation points and the reduction of communication can potentially result in improved parallel performance. Note that in practical implementations, the inner systems of Alg. 2 are often solved (approximately) by some other iterative method.

3 Parallel Implementation Details

3.1 Brief Description of GridSolve

GridSolve is a distributed programming system which uses a client-server model for solving complex problems remotely on global networks. The middleware consists of the following components.

1. The client, which can remotely execute tasks on computational servers using information provided by the agent.
2. The agent, which actively monitors server properties such as CPU speed, memory size, computational services, workload, and availability.
3. The computational servers, which can run predefined tasks. Any data that are read or generated locally during the execution of a task are lost after the task completes, unless the data are stored on the IBP data depot.
4. The IBP data depot, which acts as a storage device and is accessible by the client and the servers. The client uploads (downloads) data to (from) the IBP data depot which is in close proximity to the computational servers and tasks can then read (write) data from (to) the depot. Therefore, using the IBP data depot induces bridge communication between the client and the servers.

3.2 Decoupled Iterations

The coarse-grain nature of the asynchronous preconditioning iteration makes this operation naturally suited for distributed computing. Moreover, the preconditioning step can be performed on unreliable hardware. Stalling of one of the preconditioning servers will result in a less effective preconditioning operation, but the main solution method will not break down.

Algorithm 3 Asynchronous block Jacobi iteration task for each server i

Ensure: $u_i = M(r_i, \epsilon_{\text{in}}, T_{\max})$

1: Read r_i from IBP depot; Set $u_i = 0$;
2: Perform ILU decomposition of A_{ii};
3: **while** $t_{\text{elapsed}} < T_{\max}$ **do**
4: Read relevant part of u from IBP depot;
5: Compute $v_i = r_i - \sum_j A_{ij} u_j$;
6: Solve $A_{ii} p_i = v_i$ approximately with accuracy ϵ_{in};
7: Update $u_i \leftarrow u_i + p_i$;
8: Write u_i to IBP depot;
9: **end while**

However, the other operations (i.e. the matrix–vector multiplication, orthogonalisation, and vector operations) are relatively fine-grain and need to be performed on stable hardware. It may therefore be natural to perform the outer iteration on the (reliable) client machine. This approach has an obvious limitation. Depending on the problem size and the number of servers, the outer iteration may become a computational bottleneck. We have therefore also implemented a parallel outer iteration using techniques described in [3].

Currently, the matrix is partitioned using a homogeneous one-dimensional block-row distribution, both in the preconditioning iteration and in the outer iteration. The vectors are distributed accordingly. What follows are various implementation

issues pertaining to performing the outer iteration in sequential or parallel. Note that in both cases the inner preconditioning is performed in parallel on heterogeneous computing hardware.

3.2.1 Sequential Outer Loop

All of the operations – with exception of the preconditioning iteration – are performed on the client machine. There is a single GridSolve task for the preconditioning step, which implies that there is a single global synchronisation point in each outer iteration step. The client machine begins by updating the complete residual on the IBP data depot. Algorithm 3 shows the specific steps performed by each server i in the preconditioning phase.

At the beginning of task i, the appropriate portion of the residual is read and the task starts iterating on its portion of u. At the end of each block Jacobi iteration step, the server updates the relevant portion(s) of u (i.e. the overlap) on the IBP depot. This process continues until some appropriate criterion is met, which is currently related to a simple time limit. Each process then writes its part of u to the IBP depot and the complete vector u is read by the client machine. The obtained search direction is then used to compute the new iterate and residual. This procedure is repeated until convergence.

3.2.2 Parallel Outer Iteration

In this case, the only data that is communicated between the client and the computational nodes are the results of the (partial) inner products. The classical Gram–Schmidt algorithm (CGS) was chosen for the orthogonalisation step, since it has favourable parallel properties.

By combining operations as much as possible, three distinct GridSolve tasks can be constructed, giving three synchronisation points per outer iteration step. The first task consists of two main operations: updating the iterate and residual and performing the asynchronous Jacobi iterations. The second GridSolve task has two operations: computing the local matrix–vector product and performing the first phase of the CGS algorithm. The third and last GridSolve task performs the second phase of CGS and stores the newly computed search directions.

A disadvantage of this approach is that every GridSolve task should be performed on reliable hardware. That is, should any of the tasks fail, it is likely that important intermediate information is lost, halting the entire outer iteration process.

4 Numerical Simulations

We have conducted several experiments solving the following 3D convection–diffusion problem,

$$\begin{cases} -\nabla^2 u + (2\mathbf{p} \cdot \nabla)u = f, & u \in \Omega, \\ u = g, & u \in \partial\Omega, \end{cases} \tag{3}$$

where $\mathbf{p} = (1, 2, 3)$, Ω is the domain, and f, g are given vectors. Discretisation by the finite difference scheme with a seven point stencil on a uniform $n_x \times n_y \times n_z$ mesh results in a sparse linear system of equations $Ax = b$ where A is of order $n = n_x n_y n_z$. Centered differences are used for the first derivatives. The grid points are numbered using the standard (lexicographic) ordering, resulting in a heptadiagonal coefficient matrix. The right-hand side vector b is generated from the constant solution $x = 1$.

4.1 Experimental Setup

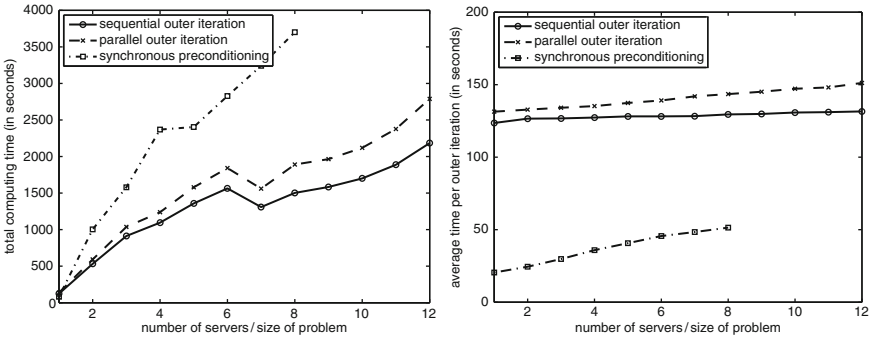
The experiments are performed using a local cluster, which is a multi-user system. It is moderately heterogeneous in design, consisting of twelve nodes: six Intel 2.20 GHz machines, two Intel 2.66 GHz machines, and four AMD Athlon 2.20 GHz machines. The nodes are equipped with memory in the range 2–4 GB and the cluster is interconnected through 100 MB s^{-1} Ethernet links. The experiments are performed on a typical work day, while other users perform their computations.

The IBP depot is started on one of the nodes in the cluster. The Jacobi sweeps are performed for a fixed number of seconds $T_{\max} = 120$s and we use matrix-free storage. The inner iterations are solved inaccurately with relative tolerance $\epsilon_{\text{in}} = 10^{-4}$ using the recent Krylov method IDR(s) [6] taking $s = 4$ and preconditioned with ILU. In the context of Grid computing, it is natural to fix the problem size per server and investigate the scalability of the algorithm by adding more servers in order to solve bigger problems. For each experiment, we take $n_x = n_y = n_z$ such that the number of equations of unknowns per server is approximately 250,000.

The outer iteration is performed either sequentially on the client machine or in parallel. The complete linear system is solved with relative tolerance $\epsilon = 10^{-8}$. To limit memory requirements, the truncation parameter is kept small ($m = 5$).

4.2 Experimental Results

Five executions of the algorithm are performed, each time using a different and random set of servers. Figure 1 shows experimental results obtained using up to



Total execution time Time per outer iteration step

Fig. 1 Large heterogeneous cluster with 250,000 equations per server

twelve servers (i.e. for problem sizes between 250,000 and three million), using both the sequential and parallel outer iteration. For comparison, results for standard (synchronous) block Jacobi preconditioning with inexact subdomain solves (using the parallel outer iteration) are also included.

Figure 1a shows the average execution times of the total iteration processes. The results show that the execution time of the asynchronous method increases when servers are added. This can be attributed almost completely to the increase in the number of outer iterations, which is approximately 4 for 2 servers and 11 for 8 servers. This increase in outer iterations is a result of the following two effects:

1. The coefficient matrix becomes increasingly ill-conditioned due to the increase of the problem size; and
2. The number of subdomains in asynchronous block Jacobi increases, which makes the preconditioner less effective.

Factors that could also have a large impact on the effectiveness of the preconditioner are the heterogeneity of the hardware, the differences in workload, and fluctuations in network load. In the current computational environment and using the aforementioned parameters, the number of Jacobi sweeps during a single preconditioning step ranged between approximately 120 on a fully dedicated server and 30 on a fully occupied server. However, because the subdomains are assigned to different servers in each outer iteration step, these effects are averaged out and the spread in execution time remained within 10%.

Keeping the problem size per server fixed implies that – in the ideal case where overhead is negligible – the execution time per outer iteration remains constant. This is demonstrated in Fig. 1, were we show the average times per outer iteration step. The results indicate that for the sequential outer loop the overhead is rather small. Also, for the parallel outer loop the increase in overhead due to the additional work and the (GridSolve) communication overhead is quite limited. In this case, the overhead grows more rapidly with increasing number of servers – compared to the sequential outer iteration.

For synchronous block Jacobi preconditioning, the total execution time grows significantly faster than for the asynchronous preconditioning if the number of servers is increased. This is a combination of two effects. Firstly, the number of iterations is higher for the synchronous preconditioner. That is, from 41 on 2 servers to 72 on 8 servers. A possible explanation is that, in contrast to asynchronous preconditioning, there is no exchange of information between the subdomains for synchronous preconditioning. Secondly, the time per iteration grows faster for the synchronous preconditioner. Since one synchronous preconditioning step requires much less computations than an asynchronous preconditioning step, the computation-to-synchronisation ratio is more favourable for asynchronous preconditioning. As a result, the asynchronous method outperforms the synchronous preconditioning technique. Moreover, this difference in performance becomes increasingly more significant for higher number of servers.

5 Conclusions

We have described in detail an iterative algorithm for solving in parallel large sparse linear systems on Grid computers. The method is designed to exploit the characteristics of heterogeneous networks of computers. By using an asynchronous iteration as a preconditioner in a flexible outer iteration, a method is obtained that adapts to a volatile computational environment. Furthermore, we have presented a fully working implementation using mature Grid middleware, applied to a realistic test problem. Also, valuable numerical experiments were performed under real-world conditions and we believe that the obtained results are promising in the context of sparse iterative solvers and Grid computing.

References

1. Bahi, J.M., Contassot-Vivier, S., Couturier, R.: Evaluation of the asynchronous iterative algorithms in the context of distant heterogeneous clusters. Parallel Comput. **31**(5), 439–461 (2005)
2. Bertsekas, D.P., Tsitsiklis, J.N.: Parallel and Distributed Computation: Numerical Methods. Prentice-Hall, NJ (1989). Republished by Athena Scientific, 1997
3. Collignon, T.P., van Gijzen, M.B.: Two implementations of the preconditioned Conjugate Gradient method on heterogeneous computing grids. Int. J. Appl. Math. Comp. Sci. (AMCS) **20**(1), 109–121 (2010)
4. Couturier, R., Denis, C., Jézéquel, F.: GREMLINS: a large sparse linear solver for grid environment. Parallel Comput. **34**, 380–391 (2008)
5. Dongarra, J., Li, Y., Shi, Z., Fike, D., Seymour, K., YarKhan, A.: Homepage of NetSolve/GridSolve (2007). http://icl.cs.utk.edu/netsolve/
6. Sonneveld, P., van Gijzen, M.B.: IDR(s): A family of simple and fast algorithms for solving large nonsymmetric linear systems. SIAM J. Sci. Comput. **31**(2), 1035–1062 (2008)
7. van der Vorst, H., Vuik, C.: GMRESR: A family of nested GMRES methods. Num. Lin. Alg. Appl. **1**(4), 369–386 (1994)
8. YarKhan, A., Seymour, K., Sagi, K., Shi, Z., Dongarra, J.: Recent developments in GridSolve. Int. J. High Perform. Comput. Appl. (IJHPCA) **20**(1), 131–141 (2006)

Hierarchical High Order Finite Element Approximation Spaces for H(div) and H(curl)

Denise De Siqueira, Philippe R.B. Devloo, and Sônia M. Gomes

Abstract The aim of this paper is to present a systematic procedure for the construction of a hierarchy of high order finite element approximations for H(div) and H(curl) spaces based on quadrilateral and triangular elements with rectilinear edges. The principle is to chose appropriate vector fields, based on the geometry of each element, which are multiplied by an available set of H^1 hierarchical scalar basic functions. We show that the resulting local vector bases can be combined to obtain continuous normal or tangent components on the elements interfaces, properties that characterize piecewise polynomial functions in H(div) or H(curl), respectively.

1 Introduction

In applications of mixed methods, the mathematical analysis uses constantly H(div) and H(curl) spaces, and approximations of them are required [1]. The main characteristic of piecewise polynomial H(div) functions is the continuity of the normal components over the interface of the elements, while H(curl) functions require continuous tangencial components. There are several papers in the literature where the techniques employed in the construction of finite element spaces for H(div) and H(curl) are based on De Rham Diagram (e.g., [2, 4, 5]).

In the present paper we present a different approach. Instead of De Rham Diagram, we use the geometry of the elements to construct appropriate vector fields which are multiplied by hierarchical H^1 conforming scalar functions developed in [3]. Using this systematic procedure, hierarchical vector bases are defined for quadrilateral and triangular elements. There are those basic functions that are

D. De Siqueira (✉) and S.M. Gomes
Unicamp, IMECC, Brasil
e-mail: dsiqueira@ime.unicamp.br, soniag@ime.unicamp.br

P.R.B. Devloo
Unicamp, FEC, Brasil
e-mail: phil@fec.unicamp.br

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_28,

associated to the edges, whose normal (or tangencial) components on the edges of the element are expressed in terms of the H^1 scalar basis functions corresponding to them. There are also other basis vector functions which are internal to the element, whose normal (or tangencial) components vanish over all edges. Therefore, H(div) (or H(curl)) conforming spaces can be created by simply imposing that the sum of the multiplying coefficients associated with the edge vector functions of neighboring elements is zero.

The finite element spaces obtained by our proposed procedure differs from the ones derived via De Rham diagrams. For instance, in a quadrilateral element, the current approximating spaces have dimension $(p+1)^2$, while using De Rham diagrams [5] the dimension is $2(p+2)(p+1)$. The main difference occurs in the number of internal basic functions, which sum p^2-1 in the current proposal, and $2p(p+1)$ in finite element spaces constructed in [5].

The outline of the paper is the following. Section 2 is dedicated to the construction of the H(div) approximation spaces of any degree p. We present the H^1 hierarchical scalar basic functions, both for quadrilateral and triangular functions. For both cases, the appropriate vector fields are presented and the resulting hierarchical vector bases are defined, and their principal properties are described. In Sect. 3 the H(curl) case is briefly considered, since in bidimensional regions this setting is derived from the H(div) case by simply rotating the corresponding vector field by $\pi/2$. The conclusions of the present paper are presented in Sect. 4.

2 H(div) Approximation Spaces

Let $\Omega \subset \mathbb{R}^2$ be a bounded domain with Lipschitz boundary $\partial\Omega$. The purpose in this section is to construct approximations of the H(div) space

$$H\,(div;\Omega) = \left\{ \overrightarrow{\varphi} \in L^2\,(\Omega)^2 : div\left(\overrightarrow{\varphi}\right) \in L^2\,(\Omega) \right\}. \tag{1}$$

by piecewise polynomials of high degree based on a partition $\mathscr{T}_h$ of Ω formed by polygonal elements (triangular or quadrilateral). For each $K \in \mathscr{T}_h$, $\mathscr{V}$ is the set of vertices a_k, $\mathscr{E}$ is the set of edges l_k, and C is the surface element . If $\mathscr{P}_p(K)$ denotes the polynomial space of degree at most p on K, the aim is to construct subspaces $V(\mathscr{T}_h) \subset H\,(div;\Omega)$ of the form

$$V(\mathscr{T}_h) = \left\{ \overrightarrow{\varphi} : \overrightarrow{\varphi}|_K \in \mathscr{P}_p(K) \times \mathscr{P}_p(K), \forall\, K \in \mathscr{T}_h \right\}. \tag{2}$$

In order to built from $V(\mathscr{T}_h)$ an approximation of $H\,(div;\Omega)$, it will be necessary to impose continuity of the normal components $\overrightarrow{\varphi}.\overrightarrow{\eta}$ at the interfaces of the elements.

2.1 H(div): Quadrilateral Elements

Let $\hat{K} = \{(\xi,\eta) : -1 \le \xi, \eta \le 1\}$ be the master element with vertices $a_0 = (-1,-1)$, $a_1 = (1,-1)$, $a_2 = (1,1)$ and $a_3 = (-1,1)$. The edges $l_k, k = 0,\cdots 3$ correspond to the sides linking the vertices a_k to $a_{k+1(mod\,4)}$.

In [3], a hierarchy of finite element subspaces in $H^1(\Omega)$ is constructed, where the basic functions in $\hat{K}$ are classified by:

- 4 vertex functions

$$\varphi^{a_0}(\xi,\eta) = \frac{(1-\xi)}{2}\frac{(1-\eta)}{2}, \quad \varphi^{a_1}(\xi,\eta) = \frac{(1+\xi)}{2}\frac{(1-\eta)}{2} \tag{3}$$

$$\varphi^{a_2}(\xi,\eta) = \frac{(1+\xi)}{2}\frac{(1+\eta)}{2}, \quad \varphi^{a_3}(\xi,\eta) = \frac{(1-\xi)}{2}\frac{(1+\eta)}{2} \tag{4}$$

  Note that the value of φ^{a_k} is one at a_k and zero at the other vertices.
- $4(p-1)$ edge functions

$$\begin{aligned}
\varphi^{l_0,n}(\xi,\eta) &= \varphi^{a_0}(\xi,\eta)[\varphi^{a_1}(\xi,\eta) + \varphi^{a_2}(\xi,\eta)]\, f_n(\xi),\\
\varphi^{l_1,n}(\xi,\eta) &= \varphi^{a_1}(\xi,\eta)[\varphi^{a_2}(\xi,\eta) + \varphi^{a_3}(\xi,\eta)]\, f_n(\eta),\\
\varphi^{l_2,n}(\xi,\eta) &= \varphi^{a_2}(\xi,\eta)[\varphi^{a_3}(\xi,\eta) + \varphi^{a_0}(\xi,\eta)]\, f_n(-\xi),\\
\varphi^{l_3,n}(\xi,\eta) &= \varphi^{a_3}(\xi,\eta)[\varphi^{a_0}(\xi,\eta) + \varphi^{a_1}(\xi,\eta)]\, f_n(-\eta),
\end{aligned}$$

  where f_n are the Chebychev polynomials of degree n, $n = 0, 1, \cdots, p-2$. The edge functions $\varphi^{l_k,n}$ vanish on all edges l_m, $m \neq k$;
- $(p-1)^2$ surface functions

$$\varphi^{C,n_0,n_1}(\xi,\eta) = \varphi^{a_0}(\xi,\eta)\varphi^{a_2}(\xi,\eta)\, f_{n_0}(\xi)\, f_{n_1}(\eta), \tag{5}$$

  with $0 \le n_0, n_1 \le p-2$. These functions are zero on all edges.

Let us consider a set of eighteen vectors $\overrightarrow{v}_m$, as indicated in Fig. 1, satisfying the properties

1. $\overrightarrow{v}_{2+3k} = \overrightarrow{\eta}_k$ is the outward unit normal, and $\overrightarrow{v}_{k+12}$ is tangent to l_k.
2. for $m = 3k$, $\overrightarrow{v}_m \cdot \overrightarrow{v}_{m+1} = \overrightarrow{v}_m \cdot \overrightarrow{v}_{m+2} = \overrightarrow{v}_{m+1} \cdot \overrightarrow{v}_{m+2} = 1$.
3. on the surface element, v_{16} and $\overrightarrow{v}_{17}$ are orthogonal vectors $\overrightarrow{v}_{16} \perp \overrightarrow{v}_{17}$.

We propose the construction of a family of vector functions by multiplication this vector field by the hierarchical scalar basis according to the following procedure:

$4(p+1)$ **edge vector functions**

$$k=0: \quad \overrightarrow{\varphi}^{l_0,a_0} = \varphi^{a_0}\overrightarrow{v}_0, \quad \overrightarrow{\varphi}^{l_0,a_1} = \varphi^{a_1}\overrightarrow{v}_1, \quad \overrightarrow{\varphi}^{l_0,n} = \varphi^{l_0,n}\overrightarrow{v}_2 \tag{6}$$

$$k=1: \quad \overrightarrow{\varphi}^{l_1,a_1} = \varphi^{a_1}\overrightarrow{v}_3, \quad \overrightarrow{\varphi}^{l_1,a_2} = \varphi^{a_2}\overrightarrow{v}_4, \quad \overrightarrow{\varphi}^{l_1,n} = \varphi^{l_1,n}\overrightarrow{v}_5 \tag{7}$$

$$k=2: \quad \overrightarrow{\varphi}^{l_2,a_2} = \varphi^{a_2}\overrightarrow{v}_6, \quad \overrightarrow{\varphi}^{l_2,a_3} = \varphi^{a_3}\overrightarrow{v}_7, \quad \overrightarrow{\varphi}^{l_2,n} = \varphi^{l_2,n}\overrightarrow{v}_8 \tag{8}$$

$$k=3: \quad \overrightarrow{\varphi}^{l_3,a_3} = \varphi^{a_3}\overrightarrow{v}_9, \quad \overrightarrow{\varphi}^{l_3,a_0} = \varphi^{a_0}\overrightarrow{v}_{10}, \quad \overrightarrow{\varphi}^{l_3,n} = \varphi^{l_3,n}\overrightarrow{v}_{11} \tag{9}$$

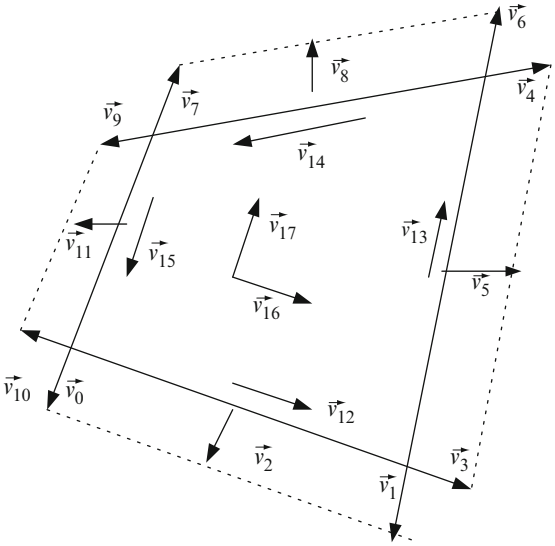


Fig. 1 Vector field for H(div)-quadrilateral elements

Observe that the vector functions associated to the edge l_0 satisfy

$$\overrightarrow{\varphi}^{l_0,a_0} \cdot \overrightarrow{\eta_0} = \varphi^{a_0} \in \mathscr{P}_1(K), \quad \overrightarrow{\varphi}^{l_0,a_1} \cdot \overrightarrow{\eta_0} = \varphi^{a_1} \in \mathscr{P}_1(K), \quad \overrightarrow{\varphi}^{l_0,n} \cdot \overrightarrow{\eta}_0 = \varphi^{l_0,n} \in \mathscr{P}_n(K). \tag{10}$$

Similar results hold for the vectors functions associated to l_k, $k = 1, 2$ and 3

$$\overrightarrow{\varphi}^{l_k,a_j} \cdot \overrightarrow{\eta_k} = \varphi^{a_j} \in \mathscr{P}_1(K), \quad \text{for } j = k, k+1 (mod 4), \quad \overrightarrow{\varphi}^{l_k,n} \cdot \overrightarrow{\eta}_k = \varphi^{l_k,n} \in \mathscr{P}_n(K). \tag{11}$$

$2(p^2-1)$**internal vector functions**

To complete the space, we add three types of functions

$$\overrightarrow{\varphi}_1^{C,n_0,n_1} = \varphi^{C,n_0,n_1} \overrightarrow{v}_{16}, \quad \overrightarrow{\varphi}_2^{C,n_0,n_1} = \varphi^{C,n_0,n_1} \overrightarrow{v}_{17}, \quad \text{and} \quad \overrightarrow{\varphi}_3^{l_k,n} = \varphi^{l_k,n} \overrightarrow{v}_{k+12}. \tag{12}$$

The normal components of these internal vector functions vanishes at all edges.

The numbers of edge and internal vector functions sums $2(p+1)^2$, coinciding with the dimension of $V_K = \mathscr{P}_p(K) \times \mathscr{P}_p(K)$.

2.2 H(div): Triangular Elements

Consider the master triangular element $\hat{K} = \{(\xi, \eta) : 0 \le \xi \le 1, 0 \le \eta \le 1 - \xi\}$, with vertices $a_0 = (0,0)$, $a_1 = (1,0)$ and $a_2 = (0,1)$, and edges $l_k, k = 0, 1, 2$

linking the vertex a_k to $a_{k+1(mod\,3)}$. For the hierarchy of finite element subspaces in $H^1(\Omega)$ constructed in [3], the basic functions are classified by:

- 3 vertex functions

$$\varphi^{a_0}(\xi,\eta) = 1-\xi-\eta, \quad \varphi^{a_1}(\xi,\eta) = \xi, \quad \varphi^{a_2}(\xi,\eta) = \eta \tag{13}$$

  that have unit value on the corresponding vertex and zero on the other ones;
- $3(p-1)$ edge functions

$$\varphi^{l_0,n}(\xi,\eta) = \varphi^{a_0}(\xi,\eta)\varphi^{a_1}(\xi,\eta) f_n(\eta+2\xi-1), \tag{14}$$
$$\varphi^{l_1,n}(\xi,\eta) = \varphi^{a_1}(\xi,\eta)\varphi^{a_2}(\xi,\eta) f_n(\eta-\xi), \tag{15}$$
$$\varphi^{l_2,n}(\xi,\eta) = \varphi^{a_2}(\xi,\eta)\varphi^{a_0}(\xi,\eta) f_n(1-\xi-2\eta) \tag{16}$$

- $\frac{(p-2)(p-1)}{2}$ surface functions

$$\varphi^{C,n_0,n_1}(\xi,\eta) = \varphi^{a_0}(\xi,\eta)\varphi^{a_1}(\xi,\eta)\varphi^{a_2}(\xi,\eta) f_{n_0}(2\xi-1) f_{n_1}(2\eta-1) \tag{17}$$

  with $0 \le n_0 + n_1 \le p-3$.

Consider a field of 14 vectors associated to a triangular element, as illustrated in Fig. 2. These vectors satisfy the properties

1. $\overrightarrow{v}_{2+3k} = \overrightarrow{\eta}_k$ is the outward unit normal, and $\overrightarrow{v}_{k+9}$ is tangent to the edge l_k.
2. for $m = 3k$, $\overrightarrow{v}_m \cdot \overrightarrow{v}_{m+1} = \overrightarrow{v}_m \cdot \overrightarrow{v}_{m+2} = \overrightarrow{v}_{m+1} \cdot \overrightarrow{v}_{m+2} = 1$.
3. $\overrightarrow{v}_{12} \perp \overrightarrow{v}_{13}$.

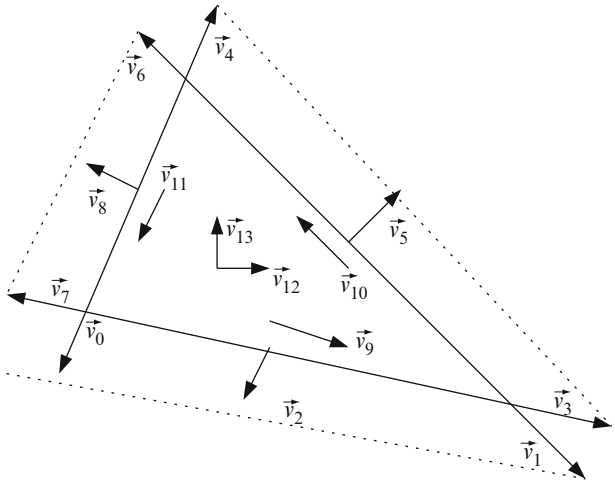


Fig. 2 Vector field for H(div)-triangular elements

As in the quadrilateral case, we introduce the vector functions associated to the edges

$$k=0: \quad \overrightarrow{\varphi}^{l_0,a_0}=\varphi^{a_0}\overrightarrow{v}_0, \quad \overrightarrow{\varphi}^{l_0,a_1}=\varphi^{a_1}\overrightarrow{v}_1, \quad \overrightarrow{\varphi}^{l_0,n}=\varphi^{l_0,n}\overrightarrow{v}_2 \tag{18}$$

$$k=1: \quad \overrightarrow{\varphi}^{l_1,a_1}=\varphi^{a_1}\overrightarrow{v}_3, \quad \overrightarrow{\varphi}^{l_1,a_2}=\varphi^{a_2}\overrightarrow{v}_4, \quad \overrightarrow{\varphi}^{l_1,n}=\varphi^{l_1,n}\overrightarrow{v}_5 \tag{19}$$

$$k=2: \quad \overrightarrow{\varphi}^{l_2,a_2}=\varphi^{a_2}\overrightarrow{v}_6, \quad \overrightarrow{\varphi}^{l_2,a_3}=\varphi^{a_3}\overrightarrow{v}_7, \quad \overrightarrow{\varphi}^{l_2,n}=\varphi^{l_2,n}\overrightarrow{v}_8 \tag{20}$$

and internal vector functions

$$\overrightarrow{\varphi}_1^{C,n_0,n_1}=\varphi^{C,n_0,n_1}\overrightarrow{v}_{12} \quad \overrightarrow{\varphi}_2^{C,n_0,n_1}=\varphi^{C,n_0,n_1}\overrightarrow{v}_{13} \quad \overrightarrow{\varphi}_3^{l_k,n}=\varphi^{l_k,n}\overrightarrow{v}_{9+k}. \tag{21}$$

Again, the normal components of the vector functions associated to the edge l_k are given by

$$\overrightarrow{\varphi}^{l_k,a_j}\cdot\overrightarrow{\eta_k}=\varphi^{a_j}\in\mathscr{P}_1(K), \quad \text{for } j=k,k+1(mod\,3), \quad \overrightarrow{\varphi}^{l_k,n}\cdot\overrightarrow{\eta}_k=\varphi^{l_k,n}\in\mathscr{P}_n(K), \tag{22}$$

and the normal components of the internal vector functions vanish at all edges. Furthermore, for triangular elements the total number of edge and internal vector functions is $(p+1)(p+2)$, also coinciding with the dimension of $V_K=\mathscr{P}_p(K)\times\mathscr{P}_p(K)$.

Having defined the two set of hierarchical vector functions in V_K, both for quadrilateral and triangular elements, it remains to verify that they indeed form bases for V_K. Furthermore, if they are supposed to be combined to span H(div) spaces $V(\mathscr{T}_h)$, we need to show that the normal components on the elements interfaces are continuous.

Theorem 1. *The edge and internal vector functions defined in (6–12), for quadrilateral elements, and in formulae (18–21) for triangular elements, form a hierarchical basis for V_K.*

Proof. Since the cardinality of such bases coincide with the dimension of V_K, it only remains to prove their linear independency. Let us consider the linear combination

$$\sum_{l_k\in\mathscr{E}}\sum_{j=k}^{k+1(mod\,4)}\alpha_j^k\overrightarrow{\varphi}^{l_k,a_j}+\sum_{l_k\in\mathscr{E}}\sum_{n=0}^{p-2}\beta_n^k\overrightarrow{\varphi}^{l_k,n}+\sum_{m=1}^{2}\sum_{n_0=0}^{p-2}\sum_{n_1=0}^{p-2}\gamma_m^{n_0,n_1}\overrightarrow{\varphi}_m^{C,n_0,n_1}+\sum_{l_k\in\mathscr{E}}\sum_{n=0}^{p-2}\mu_n^k\overrightarrow{\varphi}_3^{l_k,n}=0.$$

Restricting to the edge l_k and doing the inner product with $\overrightarrow{\eta}_k$ we obtain

$$\sum_{j=k}^{k+1(mod\,4)}\alpha_j^k\varphi^{a_j}+\sum_{n=0}^{p-2}\beta_n^k\varphi^{l_k,n}=0.$$

Using the linear independency of φ^{a_j} and $\varphi^{l_k,n}$, we conclude that $\alpha_j^k=\beta_j^k=0$. Next, considering the tangencial component associate to l_k we obtain that, restricted to this edge,

$$\sum_{n=0}^{p-2} \mu_n^k \varphi^{l_k,n} = 0,$$

implying that $\mu_n^k = 0$. Finally, doing the inner product with ν_{16}, and then with ν_{17}, we obtain

$$\sum_{n_0=0}^{p-2} \sum_{n_1=0}^{p-2} \gamma_1^{n_0,n_1} \varphi_1^{C_{n_0,n_1}} = \sum_{n_0=0}^{p-2} \sum_{n_1=0}^{p-2} \gamma_2^{n_0,n_1} \varphi_2^{C_{n_0,n_1}} = 0$$

to get $\gamma_m^{n_0,n_1} = 0$, and conclude the proof. □

Theorem 2. *Using the hierarchical vector bases defined by (6–12), for quadrilateral elements, and in formulae (18–21) for triangular elements, H(div)-conforming spaces $V(\mathscr{T}_h)$ can be created by imposing that the sum of the multiplying coefficients associated with the edge vector functions of neighboring elements is zero .*

Proof. Let $\overrightarrow{\varphi} \in V(\mathscr{T}_h)$, and K_i and K_j be two elements that share a commom edge l_k. To verify that the quantity $\overrightarrow{\varphi} \cdot \overrightarrow{\eta}_k$ is continuous across the edge l_k, taking into account the the internal vector functions have vanishing normal components on all edges, it is only necessary to verify whether the contributions of the normal components of the edge vector functions associated to l_k can be made compatible. But for these functions we have already seen that $\overrightarrow{\varphi}^{l_k,a_j} \cdot \overrightarrow{\eta_k} = \varphi^{a_j}$, and $\overrightarrow{\varphi}^{l_k,n} \cdot \overrightarrow{\eta_k} = \varphi^{l_k}$. Therefore, since the outward unit norm changes its sign from K_i to K_j , for the normal component to be continuous it is sufficient that the sum of the coefficients multiplying the edge functions is zero. □

3 H(curl) Approximation Spaces

Now we turn to the question concerning the construction of approximations of the H(curl) space

$$H(\mathbf{curl}; \Omega) = \left\{ \overrightarrow{v} \in L^2(\Omega)^2 : \mathbf{curl}\left(\overrightarrow{v}\right) \in L^2(\Omega)^2 \right\} \tag{23}$$

In order to construct piecewise polynomial vector subspaces $V(\mathscr{T}_h) \subset H(\mathbf{curl}; \Omega)$ of the form (2), it will be necessary to impose continuity of the tangencial components at the interfaces of the elements K. Similar to the H(div) case, a systematic procedure consists in first choosing an appropriate vector field, based on the geometry of the elements, and then multiply it by the set of H^1 hierarchical scalar basic functions. In order to guarantee the continuity of the tangencial components of the functions $V(\mathscr{T}_h)$ on the interfaces of the elements, such H(curl) vector field can be obtained by a $\pi/2$ rotation of the H(div) vector field, as shown in Fig. 3.

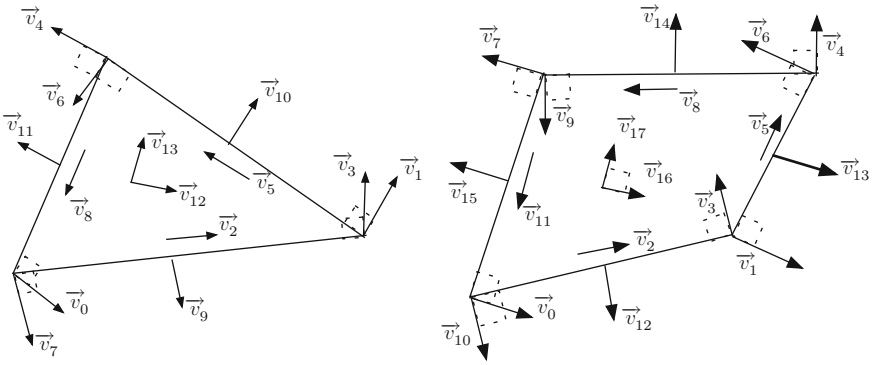


Fig. 3 Vector field for H(curl) 2D elements

4 Conclusion

We present a systematic procedure to construct hierarchical bases for H(div) and H(curl) approximation spaces which are consistent with the proprieties of these spaces. The geometrical properties of the elements are strongly used for the definition of vector fields, which are multiplied by consistent H^1 scalar functions already developed, to obtain continuity of the normal or tangencial components of the resulting vector functions. As perspectives, we plan to extend the construction of H(div) and H(curl) approximation space for elements with curvilinear boundaries and to study the stability of the mixed finite element method using these bases.

Acknowledgements The authors thankfully acknowledges financial support from the Brazilian National Agency of Petroleum, Natural Gas and Biofuels(ANP - PETROBRAS). P. Devloo and S. Gomes thankfully acknowledges financial support from CNPq – the Brazilian Research Council.

References

1. Brezzi, F., Fortin, M., *Mixed and Hybrid Finite Element Methods*, Springer, Berlin, 1991
2. Demkowicz, L. F., *Polynomial Exact Sequence and Projection-Based Interpolation with Application Maxwell Equations*, Lectures CIME Summer School, Italy, 2006
3. Devloo, P. R. B., Rylo, E. C., Bravo, C. M. A. A., *Systematic and generic construction of shape functions for p-adaptive meshes of multidimensional finite elements*, Computer Methods in Applied Mechanics and Engineering, 198, 1716–1725, 2009
4. Solin Pavel, Karel Segeth, Ivo Dolezel, *Higher-Order Finite Element Methods*, Chapman-Hall, London, 2004
5. Zaglamayr Sabine, *Hight Order Finite Element Methods for Electromagnetic Field Computation*, PhD Thesis, Johannes Keppler Universität Linz, 2006

Some Theoretical Results About Stability for IMEX Schemes Applied to Hyperbolic Equations with Stiff Reaction Terms

Rosa Donat, Inmaculada Higueras, and Anna Martinez-Gavara

Abstract In this work we are concerned with certain numerical difficulties associated to the use of high order Implicit–Explicit Runge–Kutta (IMEX-RK) schemes in a direct discretization of balance laws with stiff source terms. We consider a simple model problem, introduced by LeVeque and Yee in [J. Comput. Phys 86 (1990)], as the basic test case to explore the ability of IMEX-RK schemes to produce and maintain non-oscillatory reaction fronts.

1 Introduction

For convection–diffusion problems, linear multistep Implicit Explicit (IMEX) methods were proposed and analyzed as far back as the late 1970s [11]. Instances of these methods have been successfully applied to the incompressible Navier–Stokes equations [6] and in environmental modeling studies [12]. A systematic, comparative study for PDEs of convection–diffusion type was carried out in [1], and a corresponding study for reaction-diffusion problems arising in morphology is reported in [9].

In the context of hyperbolic conservation laws with stiff source terms, a common alternative is to use semi-implicit schemes, in which the convective derivative is treated in an explicit fashion, while the source term is handled implicitly. A second order semi-implicit scheme of this type was used in [7] to analyze the pathological

A.M. Gavara (✉)
Universidad de Sevilla, Avda. Reina Mercedes 2, 41012 Sevilla, Spain
e-mail: gavara@us.es

R. Donat
Universitat de València, Doctor Moliner, 50 46100 Burjassot, Spain
e-mail: Rosa.M.Donat@uv.es

I. Higueras
Universidad Pública de Navarra, Edificio Departamental de las Encinas, 31006 Pamplona, Spain
e-mail: higueras@unavarra.es

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_29,

behavior, of numerical nature, that can occur in certain cases involving hyperbolic conservation laws with stiff source terms. In [7], the simple model problem

$$u_t + u_x = -\mu u(u-1)(u-\frac{1}{2}), \qquad 0 < x < 1, \quad t > 0, \qquad u(x,0) = u_0(x), \quad (1)$$

is used as the basic test problem in the study of a numerical pathology that can occur when solving hyperbolic PDEs with stiff source terms: the occurrence of numerical fronts propagating at non-physical speeds.

As a general technique, IMEX methods are more amenable than the specific (second order) schemes proposed in [7] for the model problem, which are hard to generalize to more complicated situations. The ability to treat the convective part in an explicit fashion, while still maintaining an implicit handling of the source terms, gives a distinct advantage when designing a general purpose high order, high resolution numerical scheme. In fact, IMEX Runge–Kutta schemes have been used by Pareschi and Russo in [8] for hyperbolic systems of conservation laws with relaxation.

In [3], we have concentrated on first order explicit, implicit and IMEX time stepping schemes in order to quantify the source of this pathological behavior. A unified analysis for these schemes was possible by considering the discrete wave speed, and the difference between this quantity and the true speed of the reaction front. This pathology is studied and analyzed in [3] for first order schemes and in [3] for IMEX-RK schemes. An important issue that had to be considered in our analysis, was the need to maintain non-oscillatory reaction fronts, which gave rise to the *weak stability* concept analyzed in [3] for the first order case.

In this work, we analyze this *weak stability* concept for higher order time discretizations, in particular for Diagonally Implicit Explicit Runge Kutta (D-IMEX) schemes used e.g., by Pareschi and Russo in [8]. Motivated by the theory of Strong-Stability Preserving (SSP) schemes for homogeneous conservation laws [4], we seek to obtain conditions that guarantee the preservation of our *weak stability* concept, i.e., the preservation of $[0, 1]$ as an invariant region for the numerical scheme, provided that this interval is an invariant region for certain Euler-type time discretizations of each of the operators involved. For all practical purposes, we have observed that this Weak-Stability Preservation (WSP) property leads to non-oscillatory schemes.

2 Method of Lines Discretizations

The application of the method of lines to the model problem of the previous section reduces the PDE to an initial value problem for a system of ordinary differential equations (ODEs),

$$\frac{\partial U}{\partial t} = D(U(t)) + S(U(t)), \qquad U(0) = (u(x_1,0), u(x_2,0), \ldots, u(x_N,0))^T, \quad (2)$$

for the vector $U(t) = (U_1(t), U_2(t), \ldots, U_N(t))^T$ with components $U_i(t) \approx u(x_i, t)$. Due to the nature of the problem, different operators are assigned to the convective derivative and the source term.

The term $D(U)$ in (2) is the spatial discretization operator of the convective derivative term, $-f(u)_x$. It is well known that when shock computations are involved, conservative formulations,

$$D_i(U) = -\frac{F_{i+1/2} - F_{i-1/2}}{\triangle x} \tag{3}$$

must be considered. Here $F_{i+1/2} = F(U_{i-q+1}, \ldots, U_{i+q})$ is a numerical flux function consistent with the convective flux $f(u)$, i.e., $F(U, \ldots, U) = f(U)$, and F is Lipschitz continuous with respect to its arguments. The term $S(U)$ represents the discrete approximation of the source term, which will always be defined in this work as $S(U)_i = s(U_i)$.

It is widely accepted that stiff source terms should be handled in an implicit fashion, in order to avoid stability problems related with the fast scales. In [3], we examine the limitations that are encountered when using the simplest Euler-type first order schemes for the time discretization of the MOL system (2). On the other hand, there are a number of robust, and rather specialized, numerical flux functions that can be used if discontinuous, or nearly discontinuous, solutions need to be computed. These observations lead, in a rather natural way, to consider IMplicit–EXplicit (IMEX) Runge–Kutta (RK) schemes for the time integration of MOL discretizations (see [8]). The general form of an s-stage D-IMEX scheme system (2) is as follow

$$\begin{aligned} U^{(i)} &= U^n + \triangle t \sum_{j=1}^{i-1} a_{ij}\, D(U^{(j)}) + \triangle t \sum_{j=1}^{i} \tilde{a}_{ij}\, S(U^{(j)}), \qquad 1 \le i \le s, \\ U^{n+1} &= U^n + \triangle t \sum_{i=1}^{s} b_i\, D(U^{(i)}) + \triangle t \sum_{i=1}^{s} \tilde{b}_i\, S(U^{(i)}), \end{aligned} \tag{4}$$

where $U^{(i)}$ represent the internal stages of the method.

In the specialized literature concerning ARK schemes, it is customary to use a compact matrix notation to represent the method. Here, following [5], we denote $\mathscr{A} = (a_{ij})$, $\widetilde{\mathscr{A}} = (\tilde{a}_{ij})$, $b = (b_i)$ and $\tilde{b} = (\tilde{b}_i)$, and define the matrices

$$\mathbb{A} = \begin{pmatrix} \mathscr{A} & 0 \\ b^t & 0 \end{pmatrix}, \quad \widetilde{\mathbb{A}} = \begin{pmatrix} \widetilde{\mathscr{A}} & 0 \\ \tilde{b}^t & 0 \end{pmatrix}.$$

With this notation, the *compact form* of (4) is

$$\mathscr{U} = e \otimes U^n + \triangle t\, (\mathbb{A} \otimes I)\, \mathscr{D}(\mathscr{U}) + \triangle t\, (\widetilde{\mathbb{A}} \otimes I)\, \mathscr{S}(\mathscr{U}), \tag{5}$$

where $e = (1, \ldots, 1)^t \in \mathbb{R}^{s+1}$, $\mathscr{U} = (U^{(1)^t}, \ldots, U^{(s)^t}, (U^{n+1})^t)^t \in \mathbb{R}^{(s+1)N}$, $\mathscr{D}(\mathscr{U}) = (D(U^{(1)})^T, \ldots, D(U^{(s)})^T, 0^T)^T \in \mathbb{R}^{(s+1)N}$, with analogous notation for $\mathscr{S}(\mathscr{U})$. The symbol $\otimes$ denotes the Kronecker product (see, for example, [5] for specific details).

Observe that for a D-IMEX scheme (4), the matrix $\mathbb{A}$, associated to the explicit scheme, is strictly lower triangular and the matrix $\widetilde{\mathbb{A}}$, associated to the implicit one, is lower triangular.

3 Weak Stability Preservation for D-IMEX Schemes

In [3], we pointed out that the solution for the model problem (1) corresponding to the initial profile

$$u(x,0) = \begin{cases} 1, & \text{if } x < x_d\,, \\ 0, & \text{if } x > x_d\,, \end{cases}$$

satisfies $0 \leq u(x,t) \leq 1$. Because of the properties of the source term (see e.g Theorem 2.1 in [3]), it makes sense to require similar inequalities for the numerical solution, that is

$$0 \leq U^n \leq 1 \tag{6}$$

perhaps under certain stepsize restrictions, where in (6), and in the rest of this work, vector and matrix inequalities should be understood component-wise. In [3, Definition 6.1] we introduced the concept of *weakly stable* (WS) methods, defined as schemes such that, when applied to the model problem (1), the numerical solution satisfies property (6) provided $0 \leq U^0 \leq 1$. As we show in [3], the tools used in the context of Strong Stability Preserving SSP (or TVD) RK and SSP (or TVD) additive RK schemes, will serve to establish sufficient conditions that, in practice, guarantee the absence of numerical oscillations. Indeed, we seek to prove a weak stability preservation (WSP) property: assuming weak stability properties – under appropriate stepsize restrictions – for first order time integrators applied to each additive term in (2), similar ones can be ensured for the IMEX scheme, perhaps under different time-step restrictions.

Following the ideas in the SSP theory, the WSP property of an IMEX scheme will rely on the ability to write the internal stages in (7) as a convex combination of Euler steps. The following technical lemma shows that, under rather mild assumptions, the numerical solution and each internal stage in an additive scheme can be written as a linear combination of Euler steps for the convective and source terms. The proof is straightforward and shall be omitted.

Lemma 1. *We consider an additive RK scheme (5) with coefficient matrices* $(\mathbb{A}, \widetilde{\mathbb{A}})$. *Let us consider any splitting of the matrix* $\widetilde{\mathbb{A}}$ *as* $\widetilde{\mathbb{A}} = \widetilde{\mathbb{A}}_+ - \widetilde{\mathbb{A}}_-$. *Let* $r_1, r_2, r_3 \in \mathbb{R}$, *be nonzero numbers such that the matrix* $\mathscr{B} := r_1\,\mathbb{A} + r_2\,\widetilde{\mathbb{A}}_+ + r_3\,\widetilde{\mathbb{A}}_-$ *satisfies that* $(I + \mathscr{B})$ *is invertible. Then scheme (5) can be rewritten as*

$$\mathscr{U} = (I + \mathscr{B})^{-1} e \otimes U^n + r_1((I + \mathscr{B})^{-1}\mathbb{A} \otimes I)(\mathscr{U} + \frac{\Delta t}{r_1}\mathscr{D}(\mathscr{U})) \tag{7}$$

$$+ r_2((I + \mathscr{B})^{-1}\widetilde{\mathbb{A}}_+ \otimes I)(\mathscr{U} + \frac{\Delta t}{r_2}\mathscr{S}(\mathscr{U})) + r_3((I + \mathscr{B})^{-1}\widetilde{\mathbb{A}}_- \otimes I)(\mathscr{U} - \frac{\Delta t}{r_3}\mathscr{S}(\mathscr{U})).$$

Observe that Lemma 1 is valid for general splittings $\widetilde{\mathbb{A}} = \widetilde{\mathbb{A}}_+ - \widetilde{\mathbb{A}}_-$ and real triplets $(r_1, r_2, r_3) \in \mathbb{R}^3$, provided that the matrix $(I + \mathscr{B})$ is invertible. If this is the case, expanding the relation $(I + \mathscr{B})^{-1}(I + \mathscr{B})\, e = e$ we get the following relation

$$(I+\mathscr{B})^{-1}e+r_1\,(I+\mathscr{B})^{-1}\mathbb{A}\,e+r_2\,(I+\mathscr{B})^{-1}\widetilde{\mathbb{A}}_+e+r_3\,(I+\mathscr{B})^{-1}\widetilde{\mathbb{A}}_-\,e = e. \tag{8}$$

Hence, it is easy to see that in order to have a convex combination in (7) we must require also that the following assumption holds true:

ASSUMPTION SP: *Let us assume that, for an additive RK scheme* $(\mathbb{A}, \widetilde{\mathbb{A}})$*, and a given splitting* $\widetilde{\mathbb{A}} = \widetilde{\mathbb{A}}_+ - \widetilde{\mathbb{A}}_-$*, we can find* $r_1, r_2, r_3 \geq 0$ *such that matrix* $I + \mathscr{B}$ *is regular and*

$$(I + \mathscr{B})^{-1} e \geq 0\,, \qquad (I + \mathscr{B})^{-1}\mathbb{A} \geq 0\,, \qquad (I + \mathscr{B})^{-1}\widetilde{\mathbb{A}}_{\pm} \geq 0. \tag{9}$$

The second key ingredient in the proof of our main result is the ability to solve the implicit steps involved in the IMEX process. For D-IMEX schemes (4), there is only one implicit Euler step involving the source term for each internal stage. Hence, we shall assume that the source term $s(u)$ satisfies the following assumption:

ASSUMPTION IES: *If* $0 \leq u^n \leq 1$*, then* u^{n+1} *can be computed from*

$$u^{n+1} = u^n + \tau\, s(u^{n+1})$$

for all $0 < \tau \leq \tau_{IE}$*, and satisfies* $0 \leq u^{n+1} \leq 1$.

Recall that the source term in the model problem (1) satisfies this assumption with $\tau_{IE} = 4/\mu$, [3]. On the other hand, the class of source terms considered in [2] fulfill this assumption for $\tau_{IE} = +\infty$.

The main result in this section establishes sufficient conditions that ensure WSP for a given D-IMEX scheme of the form (4). See [3] for the proof details.

Theorem 1. *Let us consider a D-IMEX method of the form (5), with coefficient matrices* $(\mathbb{A}, \widetilde{\mathbb{A}})$*, for the ODE (2). Assume that:*

1. There exists constants $\tau_D > 0$*,* $\tau_+^s > 0$ *and* $\tau_-^s > 0$ *so that*

$$0 \leq U^n \leq e \qquad \Longrightarrow \qquad 0 \leq U^n + \tau\, D(U^n) \leq e\,, \quad \tau \leq \tau_D\,, \tag{10}$$

$$0 \le U^n \le e \qquad \Longrightarrow \qquad 0 \le U^n + \tau\, S(U^n) \le e\,, \quad \tau \le \tau_+^s\,, \tag{11}$$

$$0 \le U^n \le e \qquad \Longrightarrow \qquad 0 \le U^n - \tau\, S(U^n) \le e\,, \quad \tau \le \tau_-^s\,. \tag{12}$$

2. *The source term $s(u)$ satisfies the assumption IES above.*
3. *$\mathbb{A} \ge 0$, and we have constructed a partition $\widetilde{\mathbb{A}} = \widetilde{\mathbb{A}}_+ - \widetilde{\mathbb{A}}_-$ such that $\widetilde{\mathbb{A}}_+ \ge 0$, $\widetilde{\mathbb{A}}_- \ge 0$ with $\widetilde{\mathbb{A}}_-$ strictly lower triangular for which the assumption SP is satisfied, i.e., there exists $r_1, r_2, r_3 \ge 0$ such that inequalities (9) hold.*

Then $0 \le U^n \le e \Longrightarrow 0 \le U^{n+1} \le e$, under the stepsize restriction

$$\triangle t \le \min\{r_1\, \tau_{_D}\,,\, r_2\, \tau_+^s\,,\, r_3\, \tau_-^s, \gamma\, \tau_{_{IE}}\}\,. \tag{13}$$

where $\gamma = 1/\max\{\tilde{a}_{ii}, 1 \le i \le s\}$.

Theorem 1 provides a useful set of sufficient conditions for WSP in D-IMEX schemes. Since all numerical evidence points out that when numerical oscillations do occur, the values on the numerical wave profile do not lie in [0, 1], non-oscillatory results are expected when these conditions are satisfied.

Clearly, any practical application of Theorem 1 requires, as a previous step, the determination of an appropriate splitting (for more details see [3]). However, we notice that, for practical purposes, the number of variable parameters in the determination of the stepsize restriction (13) can be reduced by imposing that $r_1\tau_{_D} = r_2\tau_+^s = r_3\tau_-^s$. In this case, (13) becomes

$$\frac{\triangle t}{\tau_{_D}} \le r_1\,, \tag{14}$$

where r_1 needs to be determined for each given splitting. Taking into account that τ_D is related to the spatial mesh size $\triangle x$ through the usual CFL restriction for the operator $D(U)$ (in the homogeneous case), the stepsize restriction (14) for WSP takes the form of a CFL-like restriction. Notice that the stepsize restriction $\triangle t \le \gamma\, \tau_{_{IE}}$ in (13) simply ensures the solvability of the nonlinear equations (see assumption IES) involved in the D-IMEX process, and it is otherwise unrelated to the set of restrictions.

4 Numerical Experiments for the LeVeque–Yee Model Problem

In this section, we would like to test the sharpness of the CFL restriction for WSP found in [3]. This CFL restriction is found as the following optimization problem: we set $y = \tau_{_D}/\tau_+^s$, $z = \tau_{_D}/\tau_-^s$,

$$
\begin{aligned}
&max_{\widetilde{\mathbb{A}}_+,\widetilde{\mathbb{A}}_-} r_1 \\
&\text{subject to} \begin{cases}
(I + r_1\,\mathbb{A} + r_1\,y\,\widetilde{\mathbb{A}}_+ + r_1\,z\,\widetilde{\mathbb{A}}_-)^{-1} e \geq 0\,, \\
(I + r_1\,\mathbb{A} + r_1\,y\,\widetilde{\mathbb{A}}_+ + r_1\,z\,\widetilde{\mathbb{A}}_-)^{-1} \mathbb{A} \geq 0\,, \\
(I + r_1\,\mathbb{A} + r_1\,y\,\widetilde{\mathbb{A}}_+ + r_1\,z\,\widetilde{\mathbb{A}}_-)^{-1} \widetilde{\mathbb{A}}_\pm \geq 0\,, \\
\widetilde{\mathbb{A}} = \widetilde{\mathbb{A}}_+ - \widetilde{\mathbb{A}}_-\,, \\
\widetilde{\mathbb{A}}_+, \widetilde{\mathbb{A}}_- \geq 0\,.
\end{cases}
\end{aligned}
\tag{15}
$$

which is solved using the Matlab Optimization Toolbox. To this aim, we present a series of numerical experiments for the model problem (1) and the D-IMEX schemes known as SSP2$(3, 3, 2)$, whose coefficients are

$$
\begin{array}{c|ccc}
0 & 0 & 0 & 0 \\
\frac{1}{2} & \frac{1}{2} & 0 & 0 \\
1 & \frac{1}{2} & \frac{1}{2} & 0 \\
\hline
\mathbb{A} & \frac{1}{3} & \frac{1}{3} & \frac{1}{3}
\end{array}
\qquad
\begin{array}{c|ccc}
\frac{1}{4} & \frac{1}{4} & 0 & 0 \\
\frac{1}{4} & 0 & \frac{1}{4} & 0 \\
1 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\
\hline
\widetilde{\mathbb{A}} & \frac{1}{3} & \frac{1}{3} & \frac{1}{3}
\end{array}
\tag{16}
$$

In the notation SSP$k(s, \sigma, p)$, s is the number of stages of the implicit scheme, σ the number of stages in the explicit scheme and k is the order of the explicit scheme and p is the order of the IMEX scheme. For these IMEX methods, the explicit part is SSP. These IMEX RK schemes are considered in [8] in the context of hyperbolic systems of conservation laws with stiff relaxation terms.

We consider the upwind discretization of the convective derivative, already used for numerical testing in [3]. In this case, it is well known that $\tau_D = \triangle x$. In addition, we also present some numerical tests where the numerical flux function is computed using the standard ENO2 and ENO3 reconstructions (see e.g., [10]). These are examples of nonlinear reconstruction techniques, for which the use of a semi-implicit alternative, such as a D-IMEX scheme, is almost mandatory. The use of the ENO2 numerical flux functions leads to a TVD scheme in the homogeneous case under the usual CFL restrictions. This ensures that $\tau_D = \triangle x$ for the model problem also in this case. On the other hand, the ENO3 numerical flux function is only *Essentially Non Oscillatory*. Our numerical tests indicate that under the CFL restrictions for WSP, the behavior of this scheme is also non-oscillatory. From this point of view, the concepts of WS, and WSP, seem to provide adequate generalizations of the stronger SS equivalents, for those problems where the nature of the solution makes it impossible to enforce them.

In order to obtain the numerical values, we plot the numerical solution with Matlab. When the numerical solution lies in the interval $[0, 1]$, the plot window is automatically set to $[0, 1]$; however, if some value is out of this interval, the plot window is automatically increased. As in many cases the values out of $[0, 1]$ are negligible on the plot, this way of proceeding allows us to easily detect the

Table 1 SSP2(3,3,2): Theoretical and observed CFL for with first order upwind scheme; observed CFL for ENO2 and ENO3

	$\mu\triangle x = 1$	$\mu\triangle x = 2.5$	$\mu\triangle x = 10$
Theoretical	1.1429	0.8017	0.3188
Obser. upwind	$1.3 \leq r_1 \leq 1.4$	$0.9 \leq r_1 \leq 1$	$0.4 \leq r_1 \leq 0.5$
Obser. ENO2	$1.3 \leq r_1 \leq 1.4$	$0.9 \leq r_1 \leq 1$	$0.5 \leq r_1 \leq 0.6$
Obser. ENO3	$1.3 \leq r_1 \leq 1.4$	$0.9 \leq r_1 \leq 1$	$0.6 \leq r_1 \leq 0.7$

loss of WSP. We present the numerical results for the values of $\mu\triangle x = 2.5$ and $\mu\triangle x = 10$ in Table 1, in order to compare the numerical values with the theoretical values obtained with the optimization problem 15. We consider $\triangle x = 10^{-3}$ in all numerical tests in this section.

Here, we have focussed on the ability to maintain non-oscillatory reaction fronts, we have obtained a set of sufficient conditions that ensure the preservation of certain relevant invariant regions, which was defined as a weak stability requirement in [3]. However, it is possible to see more details in this issue in [3], where we also determine the parameters that control the occurrence of reaction fronts traveling at incorrect speeds.

The next step is to test IMEX-RK schemes on more realistic problems. The analysis done for simple hyperbolic equations only allows us to determine which methods are more robust from the point of view of certain qualitative behavior, and which ones behave poorly. Although this behavior cannot be ensured for more complex problems, experience says that it is safer to integrate them with well behaved schemes for test problems. In this context, the results obtained gain relevance.

Acknowledgements The authors acknowledge support from MTM2008-00974, MTM2008-00785 and MTM2006-01275.

References

1. U. M. Ascher, S. J. Ruuth, R. J. Spiteri, *Implicit-explicit Runge-Kutta methods for time-dependent partial differential equations*, Appl. Numer. Math. **25**, pp. 151–167 (1997)
2. A. Chalabi, *On convergence of numerical schemes for hyperbolic conservation laws with stiff source terms*, Math. Comput. **66**, pp. 527–545 (1997)
3. R. Donat, I. Higueras, A. Martinez-Gavara, *On stability issues for IMEX schemes applied to 1d scalar hyperbolic equations with stiff reaction terms*, Mathematics of Computation (accepted)
4. S. Gottlieb, W. C. Shu, E. Tadmor, *Strong stability preserving high-order time discretization methods*, SIAM Rev. **43**(1), pp. 89–112 (2001)
5. I. Higueras, *Strong stability for additive Runge-Kutta methods*, SIAM J. Numer. Anal. **44**, pp. 1735–1758 (2006)
6. G. E. Karniadakis, M. Isracli, S. A. Orszag, *High-order splitting methods for the incompressible Navier–Stokes equations*, J. Comput. Phys. **97**, pp. 414–443 (1991)
7. R. J. LeVeque, H. C. Yee, *A study of numerical methods for hyperbolic conservation laws with stiff source terms*, J. Comput. Phys. **86**, pp. 187–210 (1990)
8. L. Pareschi, G. Russo, *Implicit-explicit Runge-Kutta schemes and applications to hyperbolic systems with relaxation*, J. Sci. Comput. **25**, pp. 129–155 (2005)

9. S. Ruuth, *Implicit-explicit methods for resaction-diffusion problems in pattern formation*, J. Math. Biology. **34**(2), pp. 148–176 (1995)
10. C. W. Shu, *Total-variation-diminishing time discretizations*, SIAM J. Sci. Statist. Comput. **9**, pp. 1073–1084 (1988)
11. J. M. Varah, *Stability restrictions on second order, three level finite difference schemes for parabolic equations*, SIAM J. Numer. Anal. **17**(2), pp. 300–309 (1980)
12. J. G. Verwer, J. G. Blom, W. Hundsdorfer, *An implicit-explicit approach for atmospheric transport-chemistry problems*, Appl. Numer. Math **20**, pp. 191–209 (1996)

Stable Perfectly Matched Layers for the Schrödinger Equations

Kenneth Duru and Gunilla Kreiss

Abstract We present a well-posed and stable perfectly matched layer (PML) for the time-dependent Schrödinger wave equations. The layer consists of the Hamiltonian (H) perturbed by a complex absorbing potential (CAP, $-i\sigma_1$), $H_{cap} = H - i\sigma_1$, and carefully chosen auxiliary functions to ensure a zero reflection coefficient at the interface between the physical domain and the layer. Using standard perturbation techniques, we show that the layer is asymptotically stable. Numerical experiments are presented showing the accuracy of the new PML model. The numerical scheme couples the standard Crank–Nicolson scheme for the modified wave equation to an explicit scheme of the Runge–Kutta type for the auxiliary differential equations.

1 Introduction

Many problems in the engineering and physical sciences are characterized by wave features. The physical propagation of waves occurs in large spatial domains while numerical simulations of such waves are typically confined to smaller computational domains. Special (artificial) boundary conditions are imposed on the boundaries of the computational domains. To ensure the accuracy of numerical simulations, these artificial boundary conditions are such that out-going waves disappear without reflection.

In numerical computations of quantum mechanical systems in open domains, the chemist often adds a complex absorbing potential (CAP) to the Hamiltonian close to the boundary in order to absorb outgoing waves, see [5]. The CAP technique is very popular because of its simplicity. However, CAP generates spurious reflections at the interface between the physical domain and the layer which prevents the

K. Duru (✉) and G. Kreiss
Division of Scientific Computing, Department of Information Technology, Uppsala University, Sweden
e-mail: kenneth.duru@it.uu.se, gunilla.kreiss@it.uu.se

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_30,

convergence of the numerical solution for a finite width layer (i.e., not perfectly matched).

The problem of absorbing boundaries for wave equations was transformed in a seminal paper by Berenger [2]. Berenger derived a (split-field) perfectly matched absorbing layer (PML) such that all waves entering the layer are absorbed without reflection. In general, the PML can be interpreted as a complex coordinate stretching. This enables one to mathematically (algebraically) manipulate the PML equations in order to have a more robust model.

However, PMLs for dispersive waves such as the time-dependent Schrödinger wave equation is less developed, see [1]. In this paper, we propose and analyze new PML equations for the time-dependent Schrödinger wave equations. Our approach is based on the complex coordinate stretching technique [3]. The new layer can be viewed as a modified CAP technique, where the Hamiltonian is perturbed by a CAP and the equations are corrected by accurately and carefully chosen auxiliary variables to ensure that the interface between the physical domain and the absorbing layer does not reflect outgoing waves. Standard perturbation techniques are used to show that our layers are asymptotically stable. The new PML model fits into numerical codes developed for Schrödinger wave equations based on the CAP technique. Note that the auxiliary variables can be discretized and updated independently.

In the next section, we derive the PML equations. Section 3 is devoted to stability analysis. In Sect. 4 some numerical examples are presented. We summarize and make conclusions in Sect. 5.

2 The Schrödinger Wave Equations and the PML

The appropriately scaled linear time dependent Schrödinger wave equation is usually written

$$\begin{aligned} iu_t &= Hu, \qquad \mathbf{x} \in \mathbb{R}^d, \quad t \geq 0, \\ u(\mathbf{x}, 0) &= u_0(\mathbf{x}), \end{aligned} \tag{1}$$

where $u : \mathbb{R}^d \times \mathbb{R}^+ \to \mathbb{C}$, $H = -\triangle + V$, $V : \mathbb{R}^d \times \mathbb{R}^+ \to \mathbb{R}$. Here u is the wave function, H is the Hamiltonian, V is the potential and $\triangle$ is the Laplacian. From now on, we will consider $d = 2$, where

$$H = -\frac{\partial^2}{\partial x^2} + L(\frac{\partial}{\partial y}, y), \quad L(\frac{\partial}{\partial y}, y) = -\frac{\partial^2}{\partial y^2} + V(y).$$

Here L is a linear elliptic operator with strictly positive eigenvalues.

2.1 PML Models

In this section we shall derive the PML equations using the complex coordinate stretching technique. The basic properties of a PML can be found in [7], we will only give an overview of the underlying ideas. For simplicity, we consider $V = const$. We begin by taking Fourier transform in time,

$$i(i\omega\hat{u}) = H\hat{u}. \tag{2}$$

The Cauchy problem (2) admits plane wave solutions,

$$\hat{u}(x, y, \omega) = u_0 e^{-i\omega(\kappa_1 x + \kappa_2 y)}.$$

Here κ_1 and κ_2 are real, and $(\kappa_1, \kappa_2) = (k_x/\omega, k_y/\omega)$, u_0 is a constant amplitude, (k_x, k_y) is the wave vector. Let us consider the case where we want to compute the solution in the half plane $x \leq x_0$, and the PML introduced outside that half plane, that is, in $x > x_0$. We now look for the modification of the wave equation such that it gets exponentially decaying solutions in the PML,

$$\hat{u}(x, y, \omega) = u_0 e^{-\kappa_1 \Gamma(x)} e^{-i\omega(\kappa_1 x + \kappa_2 y)}.$$

Here $\Gamma(x)$ is a real valued non-negative increasing smooth function, which is zero for $x \leq x_0$. We can re-write the decaying solution as

$$\hat{u}(x, y, \omega) = u_0 e^{-i\omega\left(\kappa_1\left(x + \frac{\Gamma(x)}{i\omega}\right) + \kappa_2 y\right)}.$$

This can be seen as a plane wave solution to the wave equation in the transformed variables $(\tilde{x}, y)$, where $\tilde{x} = x + \frac{\Gamma(x)}{i\omega}$. Let $s_1 := \frac{d\tilde{x}}{dx} = 1 + \frac{\sigma_1}{i\omega}$, where $\sigma_1 = \frac{d\Gamma(x)}{dx}$.

The PML can now be viewed as a complex change of spatial variable x in the Fourier transformed wave equation, by changing $\frac{\partial}{\partial x}$ to $\frac{1}{s_1}\frac{\partial}{\partial x}$. By applying this change of variable to (2), we have

$$i(i\omega\hat{u}) = -\frac{1}{s_1}\left(\frac{1}{s_1}\hat{u}_x\right)_x - \hat{u}_{yy} + V\hat{u}. \tag{3}$$

We observe that the plane wave satisfying (3) is

$$\hat{u}(x, y, \omega) = u_0 e^{-i\omega(s_1\kappa_1 x + \kappa_2 y)}. \tag{4}$$

Equation (3) is the PML equation. To obtain a time-dependent problem from (3) we choose the auxiliary variables,

$$\hat{v} = \frac{1}{s_1}\frac{\hat{u}_x}{i\omega}, \quad \hat{w} = -\frac{\hat{u}_y}{i\omega}, \quad \hat{z} = \frac{\hat{u}}{i\omega},$$

then invert the Fourier transform, and we have

$$\begin{aligned} i u_t &= H_{cap} u + F(\Theta, \sigma_1), \\ \Theta_t &= G(u, \Theta, \sigma_1). \end{aligned} \tag{5}$$

Where, $H_{cap} = H - i\sigma_1$, $F(\Theta, \sigma_1) = (\sigma_1 v)_x + (\sigma_1 w)_y + \sigma_1 V z$, $G(u, \Theta, \sigma_1) = [u_x - \sigma_1 v, -u_y, u]^T$ and $\Theta = [v, w, z]^T$. In the 1-D case, there is no variation in the y-direction, the auxiliary variable w vanishes. We note in passing that the original un-damped Schrödinger wave equation (1) is recovered outside the PML, where $\sigma_1 = 0$.

We see that if $F(\Theta, \sigma_1) = 0$, we have the CAP model. The PML model (5) can be viewed as a correction of the CAP technique by the carefully chosen auxiliary function $F(\Theta, \sigma_1)$ to ensure that the equations are perfectly matched. If $V = 0$, the variable z vanishes and we have the potential-free model.

3 Stability Analysis

Without loss of generality we consider $V = 0$. In order to study the stability of the model (5), we assume constant coefficients and plane wave solutions of the form (6). By inserting (6) in (5), we obtain the corresponding dispersion relation (7). From the roots $\eta(k_x, k_y, \sigma_1)$ of the dispersion relation (7) we can decide whether the model (5) is stable.

$$\mathbf{u} = \mathbf{u}_0 e^{i\eta t - i k_x x - i k_y y}, \quad \mathbf{u}_0 \in \mathbb{R}^3, \quad k_x, k_y, x, y, \in \mathbb{R}, \quad t \geq 0. \tag{6}$$

$$(\eta - i\sigma_1)^2 \eta + \eta^2 k_x^2 + (\eta - i\sigma_1)^2 k_y^2 = 0. \tag{7}$$

The system (5) is asymptotically stable if for all $(k_x, k_y) \in \mathbb{R}^2$ the roots $\eta(k_x, k_y, \sigma_1)$ of the dispersion relation have non-negative imaginary parts, i.e., $\Im\eta \geq 0$. If there are multiple roots with $\Im\eta = 0$, there must be a corresponding number of linearly independent vectors $\mathbf{u}_0$.

Theorem 1. *The constant coefficient PML model (5) is asymptotically stable for all* $\sigma_1 \geq 0$.

Proof. We introduce the normalization

$$k_1 = \frac{k_x}{|k|}, \; k_2 = \frac{k_y}{|k|}, \; |k|^2 = k_x^2 + k_y^2, \; \lambda(k_1, k_2, \epsilon) = \frac{\eta(k_x, k_y, \sigma_1)}{|k|^2}, \; \epsilon = \frac{\sigma_1}{|k|^2},$$

and we have

$$(\lambda - i\epsilon)^2 \lambda + \lambda^2 k_1^2 + (\lambda - i\epsilon)^2 k_2^2 = 0. \tag{8}$$

Consider first $k_1 = 0$. Since $k_1^2 + k_2^2 = 1$, from (8) we have

$$(\lambda - i\epsilon)^2(\lambda + 1) = 0.$$

There is a simple root $\lambda_1 = -1$ and a double root $\lambda_{2,3} = i\epsilon$. Clearly these modes are stable.

Next we consider $k_2 = 0$. This corresponds to the 1-D case. Also since $k_1^2 + k_2^2 = 1$, from (8) we have

$$\lambda\Big((\lambda - i\epsilon)^2 + \lambda\Big) = 0. \tag{9}$$

For all $\epsilon > 0$, there are 3 distinct roots

$$\lambda_1 = 0, \quad \lambda_{2,3} = \frac{i2\epsilon - 1 \pm \sqrt{1 - i4\epsilon}}{2}.$$

By evaluating the square root we have

$$\lambda_{2,3} = \frac{i2\epsilon - 1 \pm \left(\sqrt{\frac{1+\sqrt{1+(4\epsilon)^2}}{2}} - i\sqrt{\frac{2}{\frac{1}{4}+\sqrt{(\frac{1}{4})^2+\epsilon^2}}}\epsilon\right)}{2}, \tag{10}$$

with the imaginary parts

$$\Im\lambda_{2,3} = \left(2 \pm \sqrt{\frac{2}{\frac{1}{4} + \sqrt{(\frac{1}{4})^2 + \epsilon^2}}}\right)\frac{\epsilon}{2}.$$

$\Im\lambda_{2,3} > 0$, for all $\epsilon > 0$, and it follows that the mode $(k_1, 0)$ is stable.

For the remaining wave numbers, we use a standard perturbation technique. At $\epsilon = 0$, there are 3 roots

$$\lambda_1 = -(k_1^2 + k_2^2) = -1, \qquad \lambda_{2,3} = 0.$$

The first root, λ_1, is denoted the physical mode and the others, $\lambda_{2,3}$, are called the non-physical modes. In the vicinity of the physical mode, the function $\epsilon \to \lambda(k_1, k_2, \epsilon)$ has an expansion of the form

$$\lambda = -1 + \epsilon\lambda_\epsilon + O(\epsilon^2). \tag{11}$$

Insert (11) in (8) and equate different powers of ϵ to zero separately. The linear term yields

$$\lambda_\epsilon = i2k_1^2,$$

showing the stability of the physical mode for all sufficiently small $\epsilon > 0$. The non-physical modes at $\epsilon = 0$ appear as a double root. Thus we must investigate the possibility of an expansion of the form

$$\bar{\lambda} = \bar{\lambda}_\epsilon \epsilon^r + o(\epsilon^r), \quad r < 1.$$

Assume $r < 1$, and $\bar{\lambda}_\epsilon \neq 0$. From (8) we have the leading order terms $\sim \epsilon^{2r}$,

$$\epsilon^{2r} \bar{\lambda}_\epsilon^2 = 0,$$

implying $\bar{\lambda}_\epsilon = 0$, which is a contradiction. Thus we instead consider $r = 1$, yielding to leading order terms $\sim \epsilon^2$,

$$\bar{\lambda}_\epsilon = i k_2^2 \pm k_1 k_2.$$

Clearly these modes are also stable for all sufficiently small $\epsilon > 0$.

It remains to prove that the modes $k_1 \neq 0$ and $k_2 \neq 0$ are stable for all $\epsilon > 0$. By the perturbation analysis we know that for sufficiently small ϵ the corresponding roots have positive imaginary parts. Since the roots are continuous functions of ϵ instability can only occur if there is a purely real root of (8) for some $\epsilon > 0$.

Consider $k_1 \neq 0$ and $k_2 \neq 0$ and assume $\lambda \in \mathbb{R}$ is a solution of (8) for some $\epsilon > 0$. Considering the imaginary and the real parts of (8) separately yields

$$\begin{aligned} \lambda + k_2^2 &= 0, \\ (\lambda^2 - \epsilon^2)(\lambda + k_2^2) + \lambda^2 k_1^2 &= 0. \end{aligned}$$

These relations can only be satisfied if $k_1 = 0$ or $k_2 = 0$, which is a contradiction. This completes the proof. □

4 Numerical Tests

In this section, we present some numerical examples. We discretize in space using standard fourth order node centered finite difference approximations. After discretizations in space, we have a system of first order ordinary differential equations (ODE)

$$\Theta_t = G_h(u, \Theta, \sigma_1), \tag{12}$$

$$u_t = A_h u + F_h(\Theta, \sigma_1). \tag{13}$$

If $F_h(\Theta, \sigma_1)$ is known, the wave function u can be propagated using any preferred method of choice, and $F_h(\Theta, \sigma_1)$ is simply treated as a forcing. However, $F_h(\Theta, \sigma_1)$ depends on Θ which depends on time t, and is also coupled to the wave equation if $\sigma_1 \neq 0$. We have chosen to update the auxiliary variables Θ with an explicit scheme of the Runge-Kutta type. In the explicit scheme, we only use (13) to

calculate the needed Runge–Kutta stages without updating the wave function u. The calculated auxiliary variables are coupled to the standard Crank–Nicolson scheme to propagate the wave function u. Without the PML the Crank–Nicolson scheme is unconditionally stable. Since we have used an explicit scheme for the auxiliary variable we have a time step restriction. We also comment that the PML model can easily be adapted to existing codes developed for CAP by accurately discretizing the auxili ary variables Θ and the auxiliary functions F, G and appending them accordingly.

In the first experiment, we consider the 1-D Schrödinger wave equation (1) posed on an unbounded domain $-\infty \leq x \leq \infty$, with the initial data

$$u(x,0) = \exp(-x^2 + ik_0x), \quad k_0 = 5. \tag{14}$$

if the potential $V = 0$, the Schrödinger wave equation has an exact solution

$$u(x,t) = \sqrt{\frac{i}{-4t+i}} \exp\Big[\frac{-ix^2 - k_0x + k_0^2t}{-4t+i}\Big]. \tag{15}$$

In order to perform numerical experiments we restrict ourself to a bounded domain $-d-4 \leq x \leq 4+d$, where the physical domain corresponds to $-4 \leq x \leq 4$, and the extra length $d > 0$ is the PML width, simulating unbounded domain. In this experiment the aim is to study the numerical reflections at the interface of the layer and the computational domain (perfect matching). We therefore consider a wide layer $d = 16$, and the final simulation time $T = 1$, such that the reflections at the outer boundaries do not yet affect the solutions in the computational domain. The damping profile is a quadratic monomial $\sigma_1 = 100((|x|-4)/16)^2$. Numerical experiments were performed for the PML model and the CAP model. We also compute a reference solution in a larger domain $-25 \leq x \leq 25$.

The discretization errors were obtained by comparing the numerical solutions to the exact solution (15) in the interior of the domain. By comparing the PML and CAP solutions to the reference solution in the interior of the domain we obtain an accurate measure of the numerical reflections. In Table 1, the second and third columns show the numerical reflections for different resolutions while column 4 through column 6 show the discretization errors. In Table 1, we see that reflections and discretization errors from the PML model are small and they approach zero at

Table 1 Numerical reflections and discretization errors with the 4th-order accurate numerical scheme, for the zero potential 1-D model

Mesh-size	PML(reflection)	CAP(reflection)	PML(error)	CAP(error)	Reference(error)
0.2000	1.2190e-05	2.0226e-04	2.9069e-04	3.2804e-04	2.9612e-04
0.1000	9.7048e-07	1.9830e-04	1.5875e-05	1.9555e-04	1.6119e-05
0.0500	8.2285e-08	1.9540e-04	1.0895e-06	1.9519e-04	1.1010e-06
0.0250	7.2138e-09	1.9391e-04	1.0993e-07	1.9390e-04	1.1053e-07

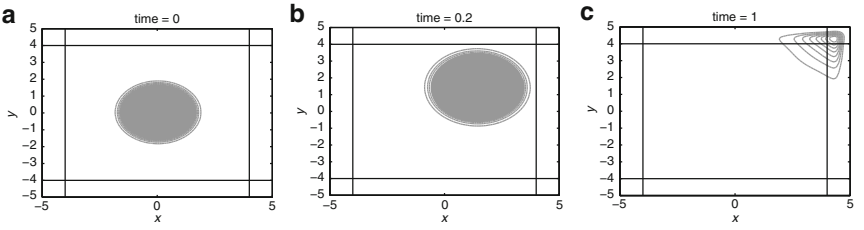


Fig. 1 The dynamics of $|u|^2$. All figures are with contours between -1 and 1 at intervals of 0.001, exempting the zero contour

the rate $O(h^4)$ while numerical reflections and discretization errors from the CAP model do not converge to zero.

In the 2-D experiment, we surround a Cartesian grid with PML and introduce initial data at the center of the domain. The behavior of the 2-D PML model is illustrated in Fig. 1, showing how the probability cloud propagates diagonally into the upper right corner and it is being effectively absorbed.

5 Summary

Perfectly matched layers for the time-dependent Schrödinger wave equations are derived and analyzed. Using plane waves and standard perturbation techniques we showed that the solutions in the PML decay exponentially in time and the direction of increasing damping. Numerical experiments are presented, illustrating the perfect matching and absorption properties of the new model.

The new model can be viewed as a modified CAP technique, where the Hamiltonian is perturbed by a CAP and the equations are corrected by carefully and accurately chosen auxiliary functions to ensure that the interface between the PML and the physical domain has a zero reflection coefficient. Since CAP reflects outgoing waves, a lot of effort has been made to derive optimal CAP parameters. These derived optimal parameters can be used in the new PML model to improve results of numerical simulations.

References

1. Antoine X., Arnold A., Besse C., Ehrhardt M., Schädle A. A Review of Transparent and Artificial Boundary Conditions Techniques for Linear and Non-linear Schrödinger Equations. Commun. Comput. Phys. 4(4) (2008) 729–796
2. Berenger J. P. A perfectly Matched Layer for the Absorption of Electromagnetic Waves. J. Comput. Phys. 114 (1994) 185–200

3. Chew W. C., Weedon W. H. A 3-D Perfectly Matched Medium from Modified Maxwell's Equations with Stretched Coordinates. Micro. Opt. Tech. Lett. 7(13) (1994) 599–604
4. Hagstrom T. New results on absorbing layers and radiation boundary conditions, Topics in computational wave propagation, 142, Lect. Notes Comput. Sci. Eng. 31, Springer, Berlin, 2003
5. Santra R. Why Complex Absorbing Potentials Work: A Discrete-Variable-Representation Perspective. Phys. Rev. A 74 (2006) 034701
6. Sjögreen B., Petersson N. A. Perfectly Matched Layer for Maxwell's Equations in Second Order Formulation. J. Comput. Phys. 209 (2005) 19–46
7. Taflove A. Advances in Computational Electrodynamics, The Finite-Difference Time-Domain. Artec House Inc. 1998
8. Zheng C. A Perfectly Matched Layer Approach to the Non-linear Schrodinger Wave Equations. J. Comput. Phys. 227 (2007) 537–556

Domain Decomposition Schemes for Frictionless Multibody Contact Problems of Elasticity

Ivan I. Dyyak and Ihor I. Prokopyshyn

Abstract The class of parallel Robin (Poincaré) domain decomposition schemes which are based on the penalty method and the simple iteration method for variational equations is proposed for solution of frictionless multibody contact problems of elasticity. The convergence of these schemes is proved. The numerical analysis is made for 2D contact problems using FEM approximations.

1 Introduction

The contact problems of elastic bodies are widely used in many fields of science and engineering. The recent achievements on analytical and numerical methods for solution of these problems are given in work [10].

Nowadays the numerical methods based on the theory of variational inequalities have become very popular for solving contact problems of elasticity. The development of domain decomposition methods (DDM) has given a powerful incentive to this approach, especially for the solution of multibody contact problems. The brief overview of existing DDM for contact problems is given in [4, 10].

Using the penalty variational formulation and the simple iteration method for variational equations we have proposed parallel Neumann and parallel Dirichlet domain decomposition schemes for solution of unilateral frictionless two-body contact problems of elasticity [9]. These schemes have been generalized for multibody contact problems and their convergence has been proved [2].

I.I. Dyyak (✉)
Ivan Franko National University of Lviv, Ukraine
e-mail: dyyak@franko.lviv.ua

I.I. Prokopyshyn
IAPMM, Naukova 3-b, 79060, Lviv, Ukraine
e-mail: ihor84@gmail.com

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_31, © Springer-Verlag Berlin Heidelberg 2010

In this contribution we propose wider class of parallel Robin (Poincaré) domain decomposition methods for frictionless contact problems which includes the schemes we have considered earlier in [2, 9].

2 Formulation of the Frictionless Multibody Contact Problem

Let us consider the frictionless contact problem of N elastic bodies $\Omega_\alpha \subset \mathbf{R}^3$ with sectionally smooth boundaries $\Gamma_\alpha = \partial\Omega_\alpha, \alpha = 1, 2, ..., N, \Omega = \bigcup_{\alpha=1}^{N} \Omega_\alpha$ (Fig. 1).

A stress-strain state of each body Ω_α is described by the vector of displacements $\mathbf{u}_\alpha = \mathbf{u}_{\alpha\, i}\, \mathbf{e}_i$, by the tensor of strains $\hat{\boldsymbol{\epsilon}}_\alpha = \varepsilon_{\alpha\, ij}\, \mathbf{e}_i \mathbf{e}_j$ and by the tensor of stresses $\hat{\boldsymbol{\sigma}}_\alpha = \sigma_{\alpha\, ij}\, \mathbf{e}_i \mathbf{e}_j$. These quantities satisfy Cauchy relations, Hooke's Law and the equilibrium equations:

$$\varepsilon_{\alpha\, ij} = \frac{1}{2}\left(\frac{\partial u_{\alpha\, i}}{\partial x_j} + \frac{\partial u_{\alpha\, j}}{\partial x_i}\right), \quad i, j = 1, 2, 3\,, \tag{1}$$

$$\sigma_{\alpha\ ij} = C_{\alpha\ ijkl}\, \varepsilon_{\alpha\ kl}, \quad i, j = 1, 2, 3\,, \tag{2}$$

$$\frac{\partial \sigma_{\alpha\ ij}}{\partial x_j} + f_{\alpha\, i} = 0, \quad i = 1, 2, 3\,, \tag{3}$$

where $f_{\alpha\, i}\,,\ i = 1, 2, 3$ are the components of the volume forces vector $\mathbf{f}_\alpha = f_{\alpha\, i}\, \mathbf{e}_i$.

Let us introduce the outer unit normal $\mathbf{n}_\alpha$ and the local coordinate system $\boldsymbol{\xi}_\alpha$, $\boldsymbol{\eta}_\alpha$, $\mathbf{n}_\alpha$ on the boundary Γ_α. Then the displacements and the stresses on the surface can be written in the following way:

$$\mathbf{u}_\alpha = u_{\alpha\,\xi}\boldsymbol{\xi}_\alpha + u_{\alpha\,\eta}\boldsymbol{\eta}_\alpha + u_{\alpha\, n}\mathbf{n}_\alpha\,,\ \ \boldsymbol{\sigma}_\alpha = \hat{\boldsymbol{\sigma}}_\alpha \cdot \mathbf{n}_\alpha = \sigma_{\alpha\,\xi}\boldsymbol{\xi}_\alpha + \sigma_{\alpha\,\eta}\boldsymbol{\eta}_\alpha + \sigma_{\alpha\, n}\mathbf{n}_\alpha\,.$$

The boundary Γ_α of every domain consists of three parts: $\Gamma_\alpha = \Gamma_\alpha^u \bigcup \Gamma_\alpha^\sigma \bigcup S_\alpha$, where $S_\alpha = \bigcup_{\beta \in B_\alpha} S_{\alpha\beta}$ is the possible contact area of body Ω_α with other bodies,

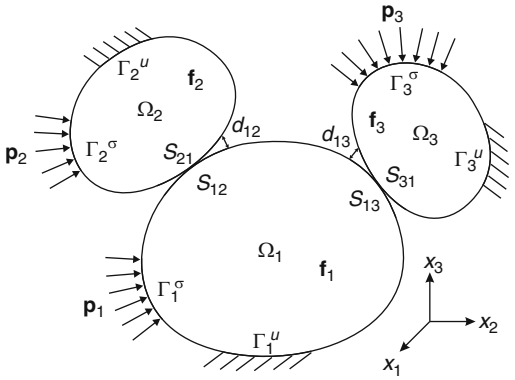


Fig. 1 Contact of several 3D bodies

$S_{\alpha\beta}$ is the possible contact area of body Ω_α with body Ω_β and $B_\alpha \subset \{1, 2, \ldots, N\}$ is the set of the indexes of all bodies which contact with body Ω_α .

We assume that surfaces $S_{\alpha\beta} \subset \Gamma_\alpha$ and $S_{\beta\alpha} \subset \Gamma_\beta$ are sufficiently close in a sense that $\mathbf{n}_\alpha(\mathbf{x}) \approx -\mathbf{n}_\beta(\mathbf{x}')$, $\mathbf{x} \in S_{\alpha\beta}$, $\mathbf{x}' = P(\mathbf{x}) \in S_{\beta\alpha}$ – projection of $\mathbf{x}$ on $S_{\beta\alpha}$ [5,6]. Let $d_{\alpha\beta}(\mathbf{x}) = \|\mathbf{x} - \mathbf{x}'\|$ be a distance between bodies Ω_α and Ω_β before the deformation.

On the part Γ_α^u the kinematical (Dirichlet) boundary conditions are prescribed:

$$\mathbf{u}_\alpha(\mathbf{x}) = 0, \ \ \mathbf{x} \in \Gamma_\alpha^u \ . \tag{4}$$

On the part Γ_α^u we consider the static (Neumann) boundary conditions:

$$\boldsymbol{\sigma}_\alpha(\mathbf{x}) = \mathbf{p}_\alpha(\mathbf{x}), \ \ \mathbf{x} \in \Gamma_\alpha^\sigma \ , \tag{5}$$

where $\mathbf{p}_\alpha$ are prescribed boundary stresses.

On the possible contact areas $S_{\alpha\beta}$ $\left(\mathbf{x} \in S_{\alpha\beta}, \ \ \mathbf{x}' = P(\mathbf{x}) \in S_{\beta\alpha}\right)$ the unilateral frictionless contact conditions hold:

$$\sigma_{\alpha n}(\mathbf{x}) = \sigma_{\beta n}(\mathbf{x}') \le 0 \ , \tag{6}$$

$$\sigma_{\alpha\xi}(\mathbf{x}) = \sigma_{\beta\xi}(\mathbf{x}') = 0, \ \ \sigma_{\alpha\eta}(\mathbf{x}) = \sigma_{\beta\eta}(\mathbf{x}') = 0 \ , \tag{7}$$

$$u_{\alpha n}(\mathbf{x}) + u_{\beta n}(\mathbf{x}') \le d_{\alpha\beta}(\mathbf{x}) \ , \tag{8}$$

$$\left(u_{\alpha n}(\mathbf{x}) + u_{\beta n}(\mathbf{x}') - d_{\alpha\beta}(\mathbf{x})\right)\sigma_{\alpha n}(\mathbf{x}) = 0 \ . \tag{9}$$

3 Variational Formulations of the Problem: The Penalty Method

Let us consider Sobolev space $V_\alpha = \left(H^1(\Omega_\alpha)\right)^3$ at the domain Ω_α and the closed subspace $V_\alpha^0 = \{\mathbf{u}_\alpha : \ \mathbf{u}_\alpha \in V_\alpha; \ \mathbf{u}_\alpha(\mathbf{x}) = 0, \ \mathbf{x} \in \Gamma_\alpha^u\}$. The space $V_0 = V_1^0 \times \ldots \times V_N^0$ is the Hilbert space with the scalar product $(\mathbf{u}, \mathbf{v})_{V_0} = \sum_{\alpha=1}^N (\mathbf{u}_\alpha, \mathbf{v}_\alpha)_{V_\alpha^0}$, $\mathbf{u}, \mathbf{v} \in V_0$.

Let us introduce the closed convex set of all displacements at V_0 which satisfy the non-penetration contact conditions (8):

$$K = \left\{\mathbf{u} : \ \mathbf{u} \in V_0; \ u_{\alpha n}(\mathbf{x}) + u_{\beta n}(\mathbf{x}') \le d_{\alpha\beta}(\mathbf{x}), \ \mathbf{x} \in S_{\alpha\beta}, \ \mathbf{x}' \in S_{\beta\alpha}\right\}, \tag{10}$$

where $\{\alpha, \beta\} \in Q$, $Q = \{\{\alpha, \beta\} : \ \alpha \in \{1, 2, \ldots, N\}, \ \beta \in B_\alpha\}$.

The contact problem (1) – (9) has an alternative formulation as the minimization problem of the quadratic functional at the set K [6,7]:

$$F(\mathbf{u}) = \frac{1}{2} A(\mathbf{u}, \mathbf{u}) - L(\mathbf{u}) \to \min_{\mathbf{u} \in K} \ , \tag{11}$$

where $A(\mathbf{u},\mathbf{v}) = \sum_{\alpha=1}^{N} a_\alpha(\mathbf{u}_\alpha,\mathbf{v}_\alpha)$ is the bilinear form which represents the total deformation energy of the system of bodies and $L(\mathbf{u}) = \sum_{\alpha=1}^{N} l_\alpha(\mathbf{u}_\alpha)$ is the linear form which is equal to the external forces work,

$$a_\alpha(\mathbf{u}_\alpha,\mathbf{v}_\alpha) = \int_{\Omega_\alpha} \hat{\boldsymbol{\sigma}}_\alpha(\mathbf{u}_\alpha) \ :\hat{\boldsymbol{\epsilon}}_\alpha(\mathbf{v}_\alpha)\, d\Omega \ , \quad \mathbf{u}_\alpha,\mathbf{v}_\alpha \in V_\alpha^0 \ ,$$

$$l_\alpha(\mathbf{v}_\alpha) = \int_{\Gamma_\alpha^\sigma} \mathbf{p}_\alpha\cdot\mathbf{v}_\alpha\, dS + \int_{\Omega_\alpha} \mathbf{f}_\alpha\cdot\mathbf{v}_\alpha\, d\Omega \ , \quad \mathbf{v}_\alpha \in V_\alpha^0 \ , \quad \alpha = 1,2,\ldots,N \ .$$

The bilinear form A is symmetric, continuous and coercive, and the linear form L is continuous, i.e.,

$$\forall\, \mathbf{u},\mathbf{v} \in V_0 \quad A(\mathbf{u},\mathbf{v}) = A(\mathbf{v},\mathbf{u}) \ , \tag{12}$$

$$\exists\, M > 0 \ \ \forall\, \mathbf{u},\mathbf{v} \in V_0 \quad |A(\mathbf{u},\mathbf{v})| \le M \|\mathbf{u}\|_{V_0} \|\mathbf{v}\|_{V_0} \ , \tag{13}$$

$$\exists\, B > 0 \ \ \forall\, \mathbf{u} \in V_0 \quad A(\mathbf{u},\mathbf{u}) \ge B \|\mathbf{u}\|_{V_0}^2 \ , \tag{14}$$

$$\exists\, H > 0 \ \ \forall\, \mathbf{v} \in V_0 \quad |L(\mathbf{v})| \le H \|\mathbf{v}\|_{V_0} \ . \tag{15}$$

According to [1, 6, 7], the minimization problem (11) has the unique solution at the set K and is equivalent to the following variational inequality:

$$A(\mathbf{u},\mathbf{v}-\mathbf{u}) - L(\mathbf{v}-\mathbf{u}) \ge 0 \ , \quad \forall\, \mathbf{v} \in K, \ \ \mathbf{u} \in K \ . \tag{16}$$

To satisfy the non-penetration condition (8) and to obtain an unconstrained optimization problem, we have used the penalty method with following penalty [5]:

$$J_\theta(\mathbf{u}) = \frac{1}{2\theta} \sum_{\{\alpha,\,\beta\}\in Q} \int_{S_{\alpha\beta}} \left[\left(d_{\alpha\beta}(\mathbf{x}) - u_{\alpha n}(\mathbf{x}) - u_{\beta n}(\mathbf{x}')\right)^-\right]^2 dS. \tag{17}$$

Here $\theta > 0$ is a penalty parameter, and the quantity $\sigma_{\alpha\beta} = \left(d_{\alpha\beta} - u_{\alpha n} - u_{\beta n}\right)^- / \theta$ has a sense of the contact stress.

Let us consider the following unconstrained minimization problem:

$$F_\theta(\mathbf{u}) = \frac{1}{2} A(\mathbf{u},\mathbf{u}) - L(\mathbf{u}) + J_\theta(\mathbf{u}) \to \min_{\mathbf{u}\in V_0} . \tag{18}$$

The penalty J_θ is two times Gateaux differentiable at V_0 and the Gateaux derivatives satisfy the properties:

$$\forall\, \mathbf{u} \in V_0 \ \ \exists\, R > 0 \ \ \forall\, \mathbf{v} \in V_0 \quad \left|J_\theta'(\mathbf{u},\mathbf{v})\right| \le R \|\mathbf{u}\|_{V_0} \|\mathbf{v}\|_{V_0} \, , \tag{19}$$

$$\exists\, D > 0 \ \ \forall\, \mathbf{u},\mathbf{v},\mathbf{w} \in V_0 \quad \left|J_\theta''(\mathbf{u},\mathbf{v},\mathbf{w})\right| \le D \|\mathbf{v}\|_{V_0} \|\mathbf{w}\|_{V_0} \, , \tag{20}$$

$$\forall\, \mathbf{u},\mathbf{v} \in V_0 \quad J_\theta''(\mathbf{u},\mathbf{v},\mathbf{v}) \ge 0. \tag{21}$$

Due to the properties (12)–(15) and (19)–(21), it can be shown that the functional F_θ is two times Gateaux differentiable, strictly convex, weakly lower semicontinuous and $\lim_{\|\mathbf{u}\|_{V_0}\to\infty} F_\theta(\mathbf{u}) = \infty$. Hence [1], there exists the unique solution of the minimization problem (18) and this problem is equivalent to the following variational equation:

$$A(\mathbf{u},\mathbf{v}) + J'_\theta(\mathbf{u},\mathbf{v}) - L(\mathbf{v}) = 0\,,\ \ \forall\ \mathbf{v}\in V_0\,,\ \ \mathbf{u}\in V_0, \tag{22}$$

where

$$J'_\theta(\mathbf{u},\mathbf{v}) = -\frac{1}{\theta}\sum_{\{\alpha,\beta\}\in Q}\int_{S_{\alpha\beta}} \left(d_{\alpha\beta} - u_{\alpha n} - u_{\beta n}\right)^- \left(v_{\alpha n} + v_{\beta n}\right) dS.$$

Using the results of works [3, 8], we have proved the next theorem about the convergence of this method.

Theorem 1. *If* $\bar{\mathbf{u}}_\theta \in V_0$ *is the solution of the variational equation* (22) *for* $\theta > 0$ *and if* $\bar{\mathbf{u}} \in K$ *is the solution of the variational inequality* (16)*, then* $\|\bar{\mathbf{u}}_\theta - \bar{\mathbf{u}}\|_{V_0} \underset{\theta\to\mathbf{0}}{\to} 0$.

4 The Simple Iteration Method and Domain Decomposition

Let $G(\mathbf{u},\mathbf{v})$ be a bilinear form on $V_0\times V_0$ which satisfies properties:

$$\forall\,\mathbf{u},\mathbf{v}\in V_0\ \ G(\mathbf{u},\mathbf{v}) = G(\mathbf{v},\mathbf{u})\,, \tag{23}$$

$$\exists\,\widetilde{M} > 0\ \ \forall\,\mathbf{u},\mathbf{v}\in V_0\ \ |G(\mathbf{u},\mathbf{v})| \le \widetilde{M}\,\|\mathbf{u}\|_{V_0}\,\|\mathbf{v}\|_{V_0}\,, \tag{24}$$

$$\exists\,\widetilde{B} > 0\ \ \forall\,\mathbf{u}\in V_0\ \ G(\mathbf{u},\mathbf{u}) \ge \widetilde{B}\,\|\mathbf{u}\|^2_{V_0}\,. \tag{25}$$

For the numerical solution of the nonlinear variational equation (22) we use the simple iteration method [3] in the form [2, 9]:

$$G(\mathbf{u}^{k+1},\mathbf{v}) = G(\mathbf{u}^k,\mathbf{v}) - \gamma\left[A(\mathbf{u}^k,\mathbf{v}) + J'_\theta(\mathbf{u}^k,\mathbf{v}) - L(\mathbf{v})\right],\ \ k = 0,1,\ldots, \tag{26}$$

where $\mathbf{u}^k\in V_0$ is the k-th approximation to the exact solution $\bar{\mathbf{u}}\in V_0$ of the problem (22) and $\gamma\in\mathbf{R}$ is an iterative parameter.

We have proved the next theorem [2], which generalize the theorem about the convergence of the simple iteration method for linear variational equations [3].

Theorem 2. *Let conditions (13)–(15), (19)–(21), (23)–(25) hold and the iteration parameter* γ *lies in the interval* $(0;\gamma_2)\,,\ \ \gamma_2 = 2B\widetilde{B}/\,(M+D)^2$.

Then the sequence $\{\mathbf{u}^k\}\subset V_0$ *obtained by the iterative method (26) converges strongly in* V_0 *to the exact solution* $\bar{\mathbf{u}}\in V_0$ *of the variational equation (22),*

i.e. $\left\|\mathbf{u}^k-\bar{\mathbf{u}}\right\|_{V_0}\underset{k\to\infty}{\to}0$. *Moreover, the convergence rate in the norm* $\|\mathbf{u}\|_G=\sqrt{G(\mathbf{u},\mathbf{u})}$ *is linear:* $\left\|\mathbf{u}^{k+1}-\bar{\mathbf{u}}\right\|_G\le q\left\|\mathbf{u}^k-\bar{\mathbf{u}}\right\|_G$, $q=\left[1-\gamma\left(2B-\gamma(M+D)^2/\widetilde{B}\right)/\widetilde{M}\right]^{\frac{1}{2}}<1$.

Let us choose the bilinear form G in the simple iteration method (26) as follows

$$G(\mathbf{u},\mathbf{v})=A(\mathbf{u},\mathbf{v})+X(\mathbf{u},\mathbf{v}),\ \ \mathbf{u},\mathbf{v}\in V_0\ , \tag{27}$$

where

$$X(\mathbf{u},\mathbf{v})=\frac{1}{\theta}\sum_{\{\alpha,\,\beta\}\in Q}\int_{S_{\alpha\beta}}\left(u_{\alpha n}v_{\alpha n}+u_{\beta n}v_{\beta n}\right)\psi_{\alpha\beta}\,dS, \tag{28}$$

$$\psi_{\alpha\beta}(\mathbf{x})=\begin{cases}0, & \mathbf{x}\in S_{\alpha\beta}\backslash S^1_{\alpha\beta}\\ 1, & \mathbf{x}\in S^1_{\alpha\beta}\end{cases}\ ,\ \ S^1_{\alpha\beta}\subseteq S_{\alpha\beta},\ \ \{\alpha,\beta\}\in Q.$$

The surface $S^1_{\alpha\beta}$ is a subset of $S_{\alpha\beta}$, and the function $\psi_{\alpha\beta}$ is the characteristic function of the area $S^1_{\alpha\beta}$.

Thus, the simple iteration method (26) with the bilinear form (27) can be written in the following way:

$$A(\tilde{\mathbf{u}}^{k+1},\mathbf{v})+X(\tilde{\mathbf{u}}^{k+1},\mathbf{v})=L(\mathbf{v})+X(\mathbf{u}^k,\mathbf{v})-J'_\theta(\mathbf{u}^k,\mathbf{v})\ , \tag{29}$$

$$\mathbf{u}^{k+1}=\gamma\tilde{\mathbf{u}}^{k+1}+(1-\gamma)\,\mathbf{u}^k,\ \ k=0,1,\dots\ . \tag{30}$$

The bilinear form X is symmetric, continuous and nonnegative. Therefore, according to Theorem 2, the sequence $\{\mathbf{u}^k\}$ converges strongly to the solution of (22) for $\gamma\in(0;\gamma_2)$.

As the common quantities of the subdomains are known from the previous iteration, then the method (29) and (30) leads to domain decomposition and the variational equation (29) splits into N separate variational equations for each subdomain Ω_α:

$$\begin{aligned}a_\alpha(\tilde{\mathbf{u}}^{k+1}_\alpha,\ \mathbf{v}_\alpha)+\frac{1}{\theta}\sum_{\beta\in B_\alpha}\int_{S_{\alpha\beta}}\psi_{\alpha\beta}\left(\tilde{u}^{k+1}_{\alpha n}-u^k_{\alpha n}\right)\,v_{\alpha n}\,dS=\\ =l_\alpha(\mathbf{v}_\alpha)+\frac{1}{\theta}\sum_{\beta\in B_\alpha}\int_{S_{\alpha\beta}}\left(d_{\alpha\beta}-u^k_{\alpha n}-u^k_{\beta n}\right)^-v_{\alpha n}\,dS,\end{aligned} \tag{31}$$

$$\mathbf{u}^{k+1}_\alpha=\gamma\tilde{\mathbf{u}}^{k+1}_\alpha+(1-\gamma)\,\mathbf{u}^k_\alpha\ ,\ \ \alpha=1,2,\dots,N,\ \ k=0,1,\dots\ . \tag{32}$$

At each iteration k we have to solve N parallel elasticity problems (31) with general Robin (Poincaré) condition on $S_{\alpha\beta}$. Hence, this method refers to the parallel Robin (Poincaré) domain decomposition schemes.

By choosing different characteristic functions $\psi_{\alpha\beta} = \psi_{\alpha\beta}(\mathbf{x})$, $\{\alpha, \beta\} \in Q$ in the method (31) and (32), we can get different domain decomposition schemes. Thus, taking $\psi_{\alpha\beta}(\mathbf{x}) \equiv 0 \ \forall \ \{\alpha, \beta\} \in Q$, we get the parallel Neumann method [2, 9]:

$$\sigma_{\alpha\beta}^k = \left(d_{\alpha\beta} - u_{\alpha\, n}^k - u_{\beta\, n}^k\right)^- / \theta \ , \tag{33}$$

$$a_\alpha(\tilde{\mathbf{u}}_\alpha^{k+1}, \mathbf{v}_\alpha) = l_\alpha(\mathbf{v}_\alpha) + \sum_{\beta \in B_\alpha} \int_{S_{\alpha\beta}} \sigma_{\alpha\beta}^k \, v_{\alpha n} \, dS \ , \tag{34}$$

$$\mathbf{u}_\alpha^{k+1} = \gamma \tilde{\mathbf{u}}_\alpha^{k+1} + (1-\gamma)\, \mathbf{u}_\alpha^k \ , \ \ \alpha = 1, 2, \ldots, N, \ \ k = 0, 1, \ldots . \tag{35}$$

If at each iteration k we choose

$$\psi_{\alpha\beta} = \chi_{\alpha\beta}^k(\mathbf{x}) = \begin{cases} 0, & d_{\alpha\beta}(\mathbf{x}) - u_{\alpha\, n}^k(\mathbf{x}) - u_{\beta\, n}^k(\mathbf{x}') \geq 0 \\ 1, & d_{\alpha\beta}(\mathbf{x}) - u_{\alpha\, n}^k(\mathbf{x}) - u_{\beta\, n}^k(\mathbf{x}') < 0 \end{cases} \ , \ \ \mathbf{x} \in S_{\alpha\beta}, \ \ \mathbf{x}' \in S_{\beta\alpha} \ , \tag{36}$$

then we shall get the method [9]:

$$a_\alpha(\tilde{\mathbf{u}}_\alpha^{k+1}, \ \mathbf{v}_\alpha) + \frac{1}{\theta} \sum_{\beta \in B_\alpha} \int_{S_{\alpha\beta}} \chi_{\alpha\beta}^k \left(\tilde{u}_{\alpha\, n}^{k+1} - (d_{\alpha\beta} - u_{\beta\, n}^k)\right) v_{\alpha n} \, dS = l_\alpha(\mathbf{v}_\alpha) \ , \tag{37}$$

$$\mathbf{u}_\alpha^{k+1} = \gamma \tilde{\mathbf{u}}_\alpha^{k+1} + (1-\gamma)\, \mathbf{u}_\alpha^k, \ \ \alpha = 1, 2, \ldots, N, \ \ k = 0, 1, \ldots \ . \tag{38}$$

At each step k of this scheme we have to solve N parallel elasticity problems (37) with prescribed displacements $d_{\alpha\beta} - u_{\beta\, n}^k$ on a subset of the surface $S_{\alpha\beta}$. Therefore this method refers to nonstationary Dirichlet domain decomposition schemes.

Note, that in the most general case we can choose $\psi_{\alpha\beta}$ (i.e., $S_{\alpha\beta}^1$) differently for each $\{\alpha, \beta\} \in Q$.

The advantages of proposed numerical domain decomposition schemes are the simplicity and the regularization of the contact problem through the penalty method.

5 Numerical Investigations

The numerical analysis of the schemes (31) and (32) has been made for 2D contact problem of two transversally isotropic bodies Ω_1 and Ω_2 with the plane of isotropy parallel to the plane $x_2 = 0$ (Fig. 2). We have used FEM with 15 quadratic triangular elements on the possible contact area.

The material properties of the bodies are: $E_\alpha = 1.0$, $E'_\alpha = 0.5$, $\nu_\alpha = \nu'_\alpha = 0.3$, $G_\alpha \big/ G'_\alpha = 2.0$, where E_α, ν_α and G_α are the elasticity modulus, Poisson's ratio and the shear modulus in the plane of isotropy for the body Ω_α respectively, and E'_α, ν'_α, G'_α are these constants in the orthogonal direction, $\alpha = 1, 2$.

The distance between bodies before the deformation is $d_{12} = 10^{-3} x_1^2$, the compression of the bodies is $\Delta = 2.154434 \cdot 10^{-3}$.

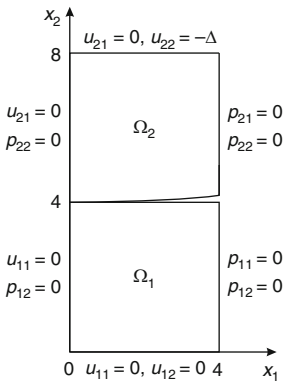


Fig. 2 Plane contact of two bodies

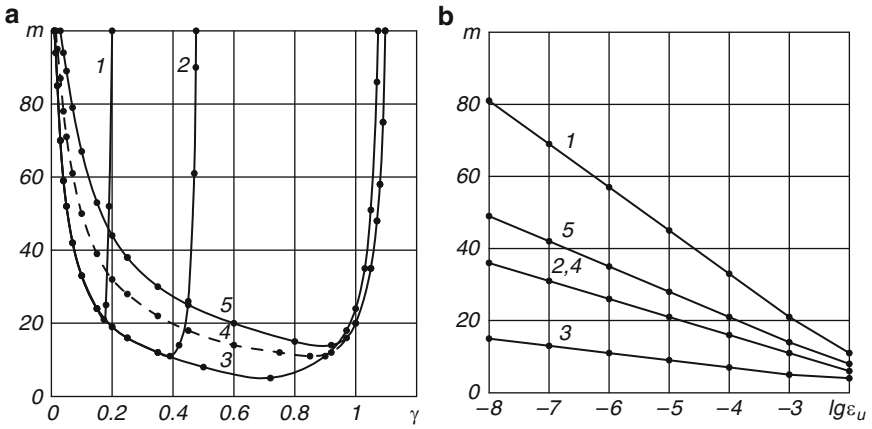


Fig. 3 Total iteration number m dependence: (**a**) on iteration parameter γ for accuracy $\varepsilon_u = 10^{-3}$; (**b**) on logarithmic accuracy $\lg \varepsilon_u$ for optimal iteration parameter $\gamma = \bar{\gamma}$

We have taken the penalty parameter in the form $\theta = c\left(h_1/E_1' + h_2/E_2'\right)$, where h_α is the height of the body Ω_α, $\alpha = 1, 2$, and $c = 0.05$ is the dimensionless penalty parameter.

We have used the following stopping criterion:

$$\left\| u_{\alpha n}^{k+1} - u_{\alpha n}^{k} \right\|_2 \Big/ \left\| u_{\alpha n}^{k+1} \right\|_2 \le \varepsilon_u,\ \alpha = 1, 2\ , \tag{39}$$

where $\|u_{\alpha n}\|_2 = \sqrt{\sum_j \left[u_{\alpha n}(\mathbf{x}^j)\right]^2}$ is the discrete norm, $\mathbf{x}^j \in S_{12}$ are the finite element nodes on the possible contact area, and $\varepsilon_u > 0$ is the relative accuracy.

At the Fig. 3 the convergence rates of different domain decomposition schemes are compared. The first curve corresponds to the parallel Neumann scheme ($S_{12}^1 = \emptyset$), curves 2, 3, 4 and 5 correspond to the parallel Robin (Poincaré) schemes

with S_{12}^1 equal to [0; 0.5], [0; 1], [0; 1.5] and [0; 2] ($S_{12}^1 = S_{12}$) respectively. Curve 3 also represents the nonstationary parallel Dirichlet scheme (37) and (38).

The optimal iteration parameters $\bar{\gamma}$, for the schemes represented at curves 1–5, are 0.173, 0.39, 0.72, 0.85 and 0.92 respectively. For $\gamma = \bar{\gamma}$ and accuracy $\varepsilon_u = 10^{-3}$ these schemes converge in 21, 11, 5, 11 and 14 iterations.

Thus, the convergence rate of all methods is linear. The parallel Robin (Poincaré) scheme with the surface closed to the real contact area ($S_{12}^1 = [0; 1]$), and the nonstationary parallel Dirichlet scheme ($\psi_{12} = \chi_{12}^k$) have the highest convergence rates. These two schemes also have the widest convergence range for γ.

The problem of a priory determination of iterative parameter γ and its upper bound γ_2, as well as development of nonstationary schemes, need additional study.

References

1. Céa, J.: Optimisation. Théorie et algorithmes. Dunod, Paris (1971)
2. Dyyak, I. I., Prokopyshyn, I. I.: The convergence of parallel Neumann domain decomposition scheme for frictionless multibody contact problems of elasticity. Mat. Met. Fiz.-Mekh. Polya. **52**(3), 78–89 (2009) [In Ukrainian]
3. Glowinski, R., Lions, J. L., Trémolières, R.: Analyse numérique des inéquations variationnelles. Dunod, Paris (1976)
4. Hüeber, S., Wohlmuth, B. I.: A primal-dual active set strategy for non-linear multibody contact problems. Comput. Meth. Appl. Mech. Engrg. **194**(27–29), 3147–3166 (2005)
5. Kikuchi, N., Oden, J. T.: Contact Problem in Elasticity: A Study of Variational Inequalities and Finite Element Methods. SIAM, Philadelphia (1988)
6. Kravchuk, A. S.: The formulation of the contact problem for several deformable bodies as the nonlinear programming problem. PMM. **42**(3), 466–474 (1978) [In Russian]
7. Kuzmenko, V. I.: On variational approach to the theory of contact problems for nonlinear elastic multilayer bodies. PMM. **43**(5), 893–901 (1979) [In Russian]
8. Lions, J. L.: Quelques méthodes de résolution des problèmes aux limites non linéaire. Dunod, Gauthier-Villars, Paris (1969)
9. Prokopyshyn, I.: Parallel domain decomposition schemes for frictionless contact problems of elasticity. Visnyk Lviv Univ. Ser. Appl. Math. Comp. Sci. **14**, 123–133 (2008) [In Ukrainian]
10. Wriggers, P.: Computational Contact Mechanics, second ed. Springer, Heidelberg (2006)

Analysis and Acceleration of a Fluid-Structure Interaction Coupling Scheme

Michael R. Dörfel and Bernd Simeon

Abstract The aim of this contribution is to shed new light on the intrinsic properties of fluid-structure coupling methods. By studying the Dirichlet–Neumann coupling for a one-dimensional model problem, it is shown that the masses at the interface can be shifted to accelerate the convergence of the extensively applied Gauss–Seidel-type iteration. Furthermore, the applied time integration as well as the size of the spatial grid adjacent to the interface influence the convergence behavior. Numerical studies confirm the results.

1 Introduction

In recent years, research in fluid-structure interaction has attracted much attention. Not only the problem variety ranges from aeroelasticity to biomedical applications, but also the solver strategies vary from monolithic approaches where both fields are solved simultaneously [1, 6, 10] to partitioned approaches where separate solvers for the fluid and structural subproblems are employed [2, 12]. Finding efficient and robust coupling schemes in the latter turn out to be very challenging in particular in biomedical applications such as hemodynamics due the physical properties of the interaction between blood flow and arterial wall [4, 5, 11]. As discovered by Förster et al. in [4], the mass densities of the fluid and the wall play a crucial role here leading to the so-called "added mass" effect, which represents, in case of weak coupling, an instability. In case of strong Dirichlet–Neumann coupling, on the other hand, iterative schemes in the fashion of the Gauss–Seidel technique and convergence acceleration by Aitken's method are in wide use [5, 11, 13]. As was recently shown by Joosten et al. [9] for a model problem with discrete masses the mass densities are also relevant here.

M.R. Dörfel (✉) and B. Simeon
Center for Mathematical Sciences, Chair for Numerical Analysis, Technische Universität München, Boltzmannstr. 3, 85748 Garching, Germany
e-mail: doerfel@ma.tum.de, simeon@ma.tum.de

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_32,

This contribution aims at shedding further light on the intrinsic properties of partitioned fluid-structure coupling methods. By studying a one-dimensional model problem, which can be viewed as a cross-section perpendicular to a two-dimensional interface, it turns out that the masses at the interface can be shifted to accelerate the convergence. Furthermore, the applied time integration as well as the size of the spatial grid adjacent to the interface influence the convergence behavior.

This contribution is organized as follows. In Sect. 2 the above mentioned one-dimensional model problem is presented. After discretization in space by finite elements, it is combined with typical temporal discretizations such as the θ-method and Newmark method, and the widespread strong Dirichlet–Neumann coupling is analysed. The mass shifting, a further acceleration technique besides the well-established Aitken relaxation, is derived in Sect. 3, and moreover numerical tests are reported that confirm the theoretical results. The article finishes with a summary and an outlook on future work.

2 One-Dimensional Model Problem

As shown by Joosten et al. [9], basic properties of coupling schemes in fluid-structure interaction can already be detected by studying simplified models. Instead of point masses connected by springs and dampers, however, we propose a one-dimensional coupled PDE that corresponds to a cross-section through a given interface and that includes the transport of information from the structure to the fluid and vice versa.

2.1 Problem definition

The model problem consists of a beam of length l_1 and a one-dimensional fluid of length l_2, with the variables of interest being the structural displacement $d : \Omega_S = (0, l_1) \to \mathbb{R}$ and the fluid velocity $u : \Omega_F = (l_1, l_1 + l_2) \to \mathbb{R}$.

The beam is fixed at its left boundary $\Gamma_S^D = \{0\}$ by a Dirichlet boundary condition $d(0) = 0$. In the interior of the structural domain Ω_S the equation of *linear elastodynamics* holds,

$$\rho^S \ddot{d} - k \frac{\partial^2}{\partial x^2} d = 0 \tag{1}$$

with density ρ^S and stiffness parameter k. Note that the coupling interface in this example consists purely of the right boundary of the beam $\Gamma_{\mathrm{FSI}} = \{l_1\}$ where the beam is connected to the fluid both by a Dirichlet boundary condition forcing the equilibrium of the velocities and by a Neumann boundary stating that the structural force $k \frac{\partial}{\partial x} d(l_1) = F_S$ is equal to the fluidal force F_F, i.e.,

$$\dot{d}(l_1) = u(l_1) \qquad \text{and} \qquad F_S = F_F. \tag{2}$$

In the fluid a linear diffusion equation is considered,

$$\rho^F \dot{u} - \mu \frac{\partial^2}{\partial x^2} u = 0 \tag{3}$$

with density ρ^F and viscosity μ. Thus, the fluid force for the coupling is $F_F = \mu \frac{\partial}{\partial x} u(l_1)$. At the right boundary of the domain, i.e., at $\Gamma_F^D = \{l_1 + l_2\}$, a Dirichlet boundary condition fixes the velocity by requiring $u(l_1 + l_2) = 0$.

2.2 Discretization in space and time

This model problem is now discretized by the Finite Element Method (FEM), see e.g., [7]. Hence both (1) and (3) are transformed into their weak form, and after the Galerkin projection onto the subspaces

$$V_S = \{\varphi_i^S \in H^1(0, l_1),\ i = 1 \dots N_S,\ \varphi_i^S(0) = 0\} \text{ and}$$
$$V_F = \{\varphi_i^F \in H^1(l_1, l_1 + l_2),\ i = 1 \dots N_F,\ \varphi_i^F(l_1 + l_2) = 0\}$$

one obtains the ODE systems

$$\mathbf{M}_S \ddot{\mathbf{d}} + \mathbf{K}_S \mathbf{d} = \mathbf{b}_S \qquad \text{and} \qquad \mathbf{M}_F \dot{\mathbf{u}} + \mathbf{K}_F \mathbf{u} = \mathbf{b}_F \tag{4}$$

where $\mathbf{d}$ and $\mathbf{u}$ denote the vectors of the unknown coefficients for the structural displacements and fluid velocitites, respectively. As usual, the mass matrices are given by $\mathbf{M}_{S/F,ij} = \rho_{S/F} \int \varphi_i^{S/F} \varphi_j^{S/F}$, the stiffness matrices by $\mathbf{K}_{S/F,ij} = k/\mu \int \frac{\partial}{\partial x} \varphi_i^{S/F} \frac{\partial}{\partial x} \varphi_j^{S/F}$, and the load vectors by $\mathbf{b}_{S,i} = F_S \varphi_i^S(1)$, $\mathbf{b}_{F,i} = -F_F \varphi_i^F(1)$.

Next, two possibly different but implicit temporal discretization methods are applied to (4). To keep the framework more general, these schemes are written as

$$\ddot{\mathbf{d}}^{n+1} = \frac{\delta^S}{\Delta t^2} \mathbf{d}^{n+1} + \mathbf{f}_{\text{old}} \qquad \text{and} \qquad \dot{\mathbf{u}}^{n+1} = \frac{\delta^F}{\Delta t} \mathbf{u}^{n+1} + \mathbf{f}_{\text{old}}$$

where $\mathbf{f}_{\text{old}}$ contains simply the history data from previous timesteps $t_n, t_{n-1}, \dots$. Examples for these discretizations are the Newmark-beta method for the structure where $\delta^S = \frac{1}{\beta}$ or the one-step-theta method for the fluid, where $\delta^F = \frac{1}{\theta}$ with $\beta, \theta \in (0, 1)$. In this way, one obtains two linear equations

$$(\delta^S \mathbf{M}_S + \Delta t^2 \mathbf{K}_S)\mathbf{d}^{n+1} = \Delta t^2 \mathbf{b}_S + \mathbf{f}_{\text{old}} \text{ and } (\delta^F \mathbf{M}_F + \Delta t\, \mathbf{K}_F)\mathbf{u}^{n+1} = \Delta t\, \mathbf{b}_F + \mathbf{f}_{\text{old}} \tag{5}$$

that are coupled by

$$\delta^C d_{N_S}^{n+1} = \Delta t\, u_1^{n+1} + f_{\text{old}} \qquad \text{and} \qquad b_{S,N_S} = F_S = F_F = -b_{F,1},$$

where the factor δ^C can be used to vary the time integration in the first coupling condition of (2). Note that we require here that $\{\phi_i^{S/F}\}_i$ form nodal bases each.

2.3 Coarse spatial discretization and convergence of the Dirichlet-Neumann partitioning

The setting $N_S = N_F = 2$ leads to the coarsest spatial discretization that distinguishes between inner and interface variables, in the following denoted by subscripts I and Γ, respectively. From (5) one obtains for the structural displacements

$$(\delta^S m_I^S + \Delta t^2 k_I^S) d_I^{n+1} + (\delta^S m_{I\Gamma}^S + \Delta t^2 k_{I\Gamma}^S) d_\Gamma^{n+1} = f_{\text{old}} \tag{6a}$$

$$(\delta^S m_{\Gamma I}^S + \Delta t^2 k_{\Gamma I}^S) d_I^{n+1} + (\delta^S m_\Gamma^S + \Delta t^2 k_\Gamma^S) d_\Gamma^{n+1} = \Delta t^2 F_S + f_{\text{old}} \tag{6b}$$

and for the fluid velocities

$$(\delta^F m_\Gamma^F + \Delta t\, k_\Gamma^F) u_\Gamma^{n+1} + (\delta^F m_{\Gamma I}^F + \Delta t\, k_{\Gamma I}^F) u_I^{n+1} = -\Delta t\, F_F + f_{\text{old}} \tag{7a}$$

$$(\delta^F m_{I\Gamma}^F + \Delta t\, k_{I\Gamma}^F) u_\Gamma^{n+1} + (\delta^F m_I^F + \Delta t\, k_I^F) u_I^{n+1} = f_{\text{old}}. \tag{7b}$$

The discretized problem and its variables are displayed in Fig. 1.

If this problem is to be solved in a partitioned manner and a strong coupling is applied, one has to iterate over the fields. In wide use is the Dirichlet–Neumann algorithm (also called Gauss–Seidel algorithm, e.g., in [9]) that can be written for one time step in the following way:

1. Predict $d_\Gamma^{n+1,(0)} = d_\Gamma^{(0)}$ (in the following time indices are omitted for simplicity)
2. Iterate over the fields for $i = 0, 1, \ldots$ until convergence
   - Solve (7) for F_F (and u_I) using $u_\Gamma^{(i)} = \frac{\delta^C}{\Delta t} d_\Gamma^{(i)} + f_{\text{old}}$ (i.e., the FSI boundary is an additional Dirichlet boundary)
   - Solve (6) for $d_\Gamma^{(i+1)}$ (and d_I) using the other coupling equation $F_S = F_F$ (i.e., the FSI boundary is an additional Neumann boundary)
   - test for convergence $\left| d_\Gamma^{(i+1)} - d_\Gamma^{(i)} \right| / \left| d_\Gamma^{(i+1)} \right| \overset{?}{\leq} \text{Tol}$
3. go to next time step $n \to n+1$.

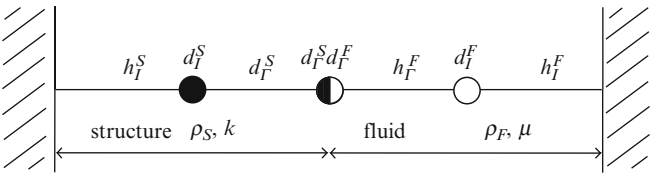


Fig. 1 One dimensional model problem with coarsest discretization

This algorithm leads to an iteration of the form

$$d_\Gamma^{(i+1)} = \lambda d_\Gamma^{(i)} + f_{\text{old}}$$

where the convergence factor λ can be written as

$$\lambda = -\delta^C \frac{\delta^F m_\Gamma^F + \Delta t\, k_\Gamma^F - \frac{(\delta^F m_{\Gamma I}^F + \Delta t\, k_{\Gamma I}^F)(\delta^F m_{I\Gamma}^F + \Delta t\, k_{I\Gamma}^F)}{\delta^F m_I^F + \Delta t\, k_I^F}}{\delta^S m_\Gamma^S + \Delta t^2 k_\Gamma^S - \frac{(\delta^S m_{\Gamma I}^S + \Delta t^2 k_{\Gamma I}^S)(\delta^S m_{I\Gamma}^S + \Delta t^2 k_{I\Gamma}^S)}{\delta^S m_I^S + \Delta t^2 k_I^S}} \tag{8}$$

$$= -\delta^C \frac{g_\Gamma^F - g_{\Gamma I}^F g_{I\Gamma}^F / g_I^F}{g_\Gamma^S - g_{\Gamma I}^S g_{I\Gamma}^S / g_I^S}$$

with $g^S = \delta^S m^S + \Delta t^2 k^S$ and $g^F = \delta^F m^F + \Delta t\, k^F$.

Especially for small Δt, the nominator and denominator of λ are governed by the first terms. i.e., the convergence depends on the ratio of the time discretization constants $\alpha_\delta := \delta^C \delta^F / \delta^S$ and the mass ratio at the interface $\alpha_m = m_\Gamma^F / m_\Gamma^S$. Since $m_\Gamma^{S/F} = \rho_{S/F} \cdot h_{S/F}^\Gamma$, latter shows the dependence on the spatial discretization at the interface $\alpha_h = h_F^\Gamma / h_S^\Gamma$ and the densities $\alpha_\rho = \rho_F / \rho_S$, which becomes crucial in hemodynamical applications [11]. Concluding we can write for small Δt

$$\lambda \approx -\alpha_\delta \cdot \alpha_m = -\alpha_\delta \cdot \alpha_h \cdot \alpha_\rho.$$

If $|\lambda| < 1$, the iteration converges linearly to the fixed point $d_\Gamma^\star = \frac{1}{1-\lambda} f_{\text{old}}$. Furthermore, it holds for the error $e_i = |d_\Gamma^{(i)} - d_\Gamma^\star|$

$$e_{i+1} = \lambda e_i.$$

Remark 1. There are only minor changes to these formulas if N_S and $N_F > 2$. The multiple inner unknowns can be written in a vector whereas there is still only one interface unknown per field. This leads to a modified equation for λ,

$$\lambda = -\delta^C \frac{g_\Gamma^F - (\mathbf{g}_{\Gamma I}^F)^T (\mathbf{G}_I^F)^{-1} \mathbf{g}_{I\Gamma}^F}{g_\Gamma^S - (\mathbf{g}_{\Gamma I}^S)^T (\mathbf{G}_I^S)^{-1} \mathbf{g}_{I\Gamma}^S} \qquad \text{where } \mathbf{G}^{S/F} = \begin{pmatrix} \mathbf{G}_I^{S/F} & \mathbf{g}_{I\Gamma}^{S/F} \\ (\mathbf{g}_{\Gamma I}^{S/F})^T & g_\Gamma^{S/F} \end{pmatrix}. \tag{9}$$

3 Interface Mass Shifting

There are different methods to increase the convergence radius of the iteration and to accelerate it, notably the widely used *Aitken* relaxation [8]. It reads

$$d_\Gamma^{(i+1)} = \omega_i \left(\lambda d_\Gamma^{(i)} + f_{\text{old}} \right) + (1 - \omega_i) d_\Gamma^{(i)},$$

and for this linear one-dimensional problem the optimal relexation factor $\omega^\star = \frac{1}{1-\lambda}$ is quickly determined, which leads to an immediate convergence.

Instead, our focus is placed on a technique that can be applied in addition to the Aitken relaxation. It is based on an observation from the last section, the dependence of the iteration on the interface masses.
Changing the spatial discretization and the time integration is not considered here although these also have influence on the coupling. The corresponding parameters $h^\Gamma_{S/F}$ and $\delta^{S/F/C}$ shall thus be fixed in the following.

3.1 Shifting Algorithm

A significant improvement of the iteration can be achieved by shifting m^F_Γ from the nominator to the denominator in (8). This is accomplished by changing the Dirichlet–Neumann algorithm

- in step 2 by partially incrementing the iteration index i, i.e., (7a) becomes

$$\delta^F m^F_\Gamma u^{(i+1)}_\Gamma + \Delta t\, k^F_\Gamma u^{(i)}_\Gamma + (\delta^F m^F_{\Gamma I} + \Delta t\, k^F_{\Gamma I}) u^{n+1}_I = -\Delta t\, F_F + f_{\text{old}}.$$

  Since $u^{(i+1)}_\Gamma$ is unknown up to that point, (7) is solved for $\hat{F}_F := F_F + \frac{\delta^F}{\Delta t} m^F_\Gamma u^{(i+1)}_\Gamma$.
- in step 3 by $F_S = F_F = \hat{F}_F - \delta^F m^F_\Gamma u^{(i+1)}_\Gamma = \hat{F}_F - \delta^F m^F_\Gamma \frac{\delta^C}{\Delta t} d^{(i+1)}_\Gamma + f_{\text{old}}$. This changes (6b) to

$$(\delta^S m^S_{\Gamma I} + \Delta t^2 k^S_{\Gamma I}) d_I + (\delta^S m^S_\Gamma + \delta^F \delta^C m^F_\Gamma + \Delta t^2 k^S_\Gamma) d^{(i+1)}_\Gamma = \Delta t^2 \hat{F}_F + f_{\text{old}}.$$

It holds for the resulting convergence factor

$$\hat{\lambda} = -\delta^C \frac{\Delta t\, k^F_\Gamma - (\mathbf{g}^F_{\Gamma I})^T (\mathbf{G}^F_I)^{-1} \mathbf{g}^F_{I\Gamma}}{\delta^S m^S_\Gamma + \delta^C \delta^F m^F_\Gamma + \Delta t^2 k^S_\Gamma - (\mathbf{g}^S_{\Gamma I})^T (\mathbf{G}^S_I)^{-1} \mathbf{g}^S_{I\Gamma}}.$$

This alteration of the algorithm can also be obtained by setting $m^S_\Gamma = m^S_\Gamma + \alpha_\delta m^F_\Gamma$ and $m^F_\Gamma = 0$ on the left-hand side of (6) and (7), respectively, and applying the old algorithm in a black box manner.

However, due to the alteration a new residual in the corresponding line of the fluid system shows up

$$r_\Gamma := m^F_\Gamma \left(u^{(i+1)}_\Gamma - u^{(i)}_\Gamma \right) = m^F_\Gamma \frac{\delta^C}{\Delta t} \left(d^{(i+1)}_\Gamma - d^{(i)}_\Gamma \right) \quad \text{with}$$
$$|r_\Gamma| \le \frac{m^F_\Gamma \delta^C}{\Delta t} \text{Tol} \left| d^{(i+1)}_\Gamma \right| \tag{10}$$

at the end of the iteration. Since all the other rows remain unchanged (note that $u_\Gamma^{(i)}$ is used there), there is no additional residual in the other components, i.e., $\mathbf{r}^F = (r_\Gamma, \mathbf{0})^T$.

3.2 Numerical Experiments

The default parameters for the simulations below are chosen as $\rho_S = 1.2\,\mathrm{g\,cm^{-1}}$, $\rho_F = 1.0\,\mathrm{g\,cm^{-1}}$, $k = 3 \cdot 10^6\,\mathrm{g\,cm\,s^{-2}}$, $\mu = 0.03\,\mathrm{g\,cm\,s^{-1}}$, $l_1 = 0.1\,\mathrm{cm}$ and $l_2 = 0.5\,\mathrm{cm}$, in analogy to the parameters in the 3d test simulations for hemodynamics, see, e.g., [3, 5, 11]. Furthermore, implicit Euler is employed for the time integration, i.e., $\delta_S = \delta_F = \delta_C = 1$. Using linear finite elements with $N_S = 10$, $N_F = 50$ equidistant nodes leads to $h_S = h_S^\Gamma = h_F^\Gamma = h_F = 0.01\,\mathrm{cm}$. In Fig. 2, the dependence of λ on Δt and the interface mass shifting is shown for several values of ρ_F. Without the mass shifting the same results as in [9] are obtained since the model problems are comparable. The different model parameters explain the shift in the time scale. The interface mass shifting decreases the convergence factor significantly and changes also its sign.

Figure 3 shows the time history of the interface and the number of iterations performed. The initial interface displacement is set to $d_\Gamma = 0.001\,\mathrm{cm}$ whereas all other displacements and velocities are 0. On the left the displacement of the interface node is plotted against the simulation time t. Using the relative Tolerance $\mathrm{Tol} = 10^{-6}$ and $\Delta t = 10^{-7}$ it can be observed that both runs give the same results, i.e., the additional residual of (10) is not interfering the results. Note that the small timestep is needed to resolve the fast oscillations in the solution as depicted. Figure 3 on the right shows the iterations that are needed by the different algorithms during the simulation. An $O(\Delta t)$ prediction $d_{\Gamma,n+1}^{(0)} = d_{\Gamma,n} + \Delta t u_{\Gamma,n}$ is applied but without

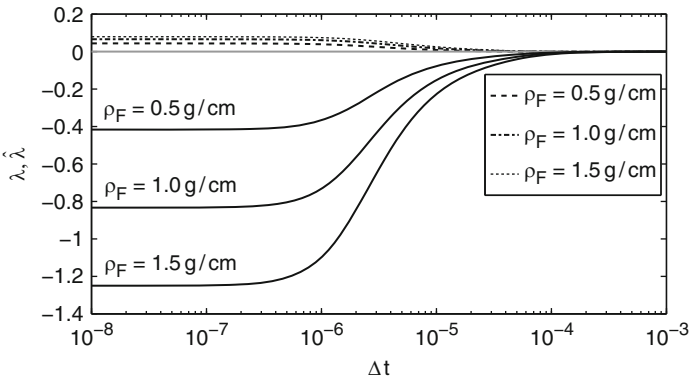


Fig. 2 Convergence factor $\lambda(\Delta t) < 0$ and mass shifted convergence factor $\hat{\lambda}(\Delta t) > 0$ with $\rho_S = 1.2\,\mathrm{g\,cm^{-1}}$ and different ρ_F

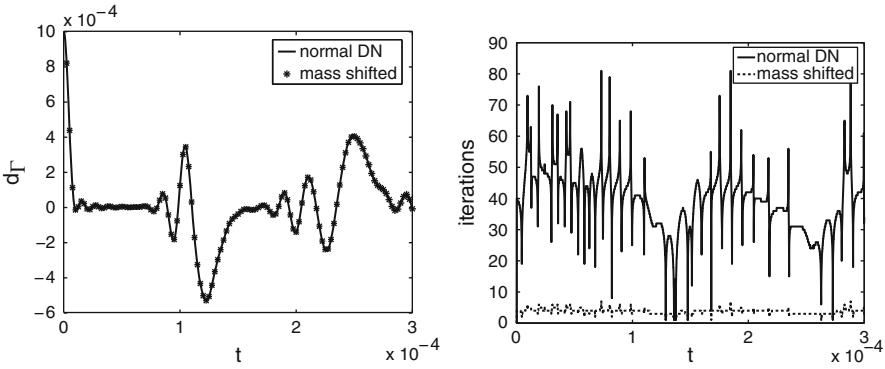


Fig. 3 Displacement of the interface node d_Γ and number of iterations per timestep

relaxation. In this plot the improvement by using the mass shifting becomes most apparent. It needs only a few, in the mean 3.79, iterations to converge while the standard method works hard on the mass ratio $\alpha_m = 0.87$ and requires 39.26 iterations per timestep on average.

4 Conclusions and Outlook

In this contribution a linear 1d model problem has been derived for analyzing the coupling of a fluid-structure interaction problem. The widely used Dirichlet–Neumann algorithm has been studied in this setting, and an acceleration method has been derived that shifts the fluid interface mass to the structural problem to gain a significant speed up both in terms number of iterations and computing time.
Currently, this approach is generalized to corresponding model problems in two and three dimensions. More specifically, using the Stokes problem on the fluid side renders the situation more complex due to the constraint that stems from the mass conservation. An upcoming article gives detailed information on the effects of the mass shifting in a 2d scenario depending on the FEM discretization and the handling of the constraint.

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. We are very grateful to W.A. Wall, M. Gee, and U. Küttler from the Institute for Computational Mechanics at TUM for fruitful discussions on the topic. We also want to thank the anonymous reviewer for the remarks that helped us to improve this contribution.

References

1. Badia, S., Quaini, A., Quarteroni, A.: Splitting methods based on algebraic factorization for fluid-structure interaction. SIAM J. Sci. Comput. 30(4), pp. 1778–1805, 2008
2. Bungartz, H.J., Schäfer, M.: Fluid-Structure Interaction, Modelling, Simulation, Optimisation. In: Lecture Notes in Computational Science and Engineering, Vol. 53, Springer, Berlin, 2006
3. Fernandez, M.A., Moubachir, M.: A Newton method using exact jacobians for solving fluid-structure coupling. Comput. Struct. 83, pp. 127–142, 2005
4. Förster, Ch., Wall, W.A., Ramm, E.: Artificial added mass instabilities in sequential staggered coupling of nonlinear structures and incompressible viscous flows. Comput. Methods Appl. Mech. Eng. 196, pp. 1278–1293, 2007
5. Gerbeau, J.F., Vidrascu, M.: A Quasi-Newton Algorithm based on a reduced model for fluid-structure interaction problems in blood flow. Math. Model. Numer. Anal. 37(4), pp. 631–647, 2003
6. Heil, M.: An efficient solver for the fully coupled solution of large-displacement fluid-structure interaction problems. Comput. Methods Appl. Mech. Eng. 193, pp. 1–23, 2004
7. Hughes, T.J.R.: The Finite Element Method. Dover, NY, 2003
8. Irons, B., Tuck, R.C.: A version of the Aitken accelerator for computer implementation. Int. J. Numer. Methods Eng. 1, pp. 275–277, 1969
9. Joosten, M.M., Dettmer, W.G., Peric, D.: Analysis of the block Gauss-Seidel solution procedure for a strongly coupled model problem with reference to fluid-structure interaction. Int. J. Numer. Methods Eng. 78, pp. 757–778, 2009
10. Küttler, U.: Effziente Lösungsverfahren für Fluid-Struktur-Interaktions-Probleme. PhD-Thesis In: Fakultät für Maschinenwesen, Technische Universität München, 2009
11. Küttler, U., Wall, W.: Fixed-point fluid-structure interaction solvers with dynamic relaxation. Comput. Mech. 43, pp. 61–72, 2008
12. Le Tallec, P., Mouro, J.: Fluid structure interaction with large structural displacements. Comput. Methods Appl. Mech. Eng. 190, pp. 3039–3067, 2001
13. Mok, D.P., Wall, W.A.: Partitioned analysis schemes for the transient interaction of incompressible flows and nonlinear flexible structures. In: Trends in Computational Structural Mechanics, W.A. Wall, K.-U. Bletzinger and K. Schweitzerhof (Eds.), pp. 689–698, 2001

Second Order Numerical Operator Splitting for 3D Advection–Diffusion-Reaction Models

Riccardo Fazio and Alessandra Jannelli

Abstract In this paper, we present a numerical operator splitting for time integration of 3D advection-diffusion-reaction problems. In this approach, three distinct methods of second order accuracy are proposed for solving, separately, each term involved in the model. Numerical results, obtained for advection – reported in [Fazio and Jannelli, IAENG Int. J. Appl. Math., **39**, 25–35, 2009] –, diffusion, and reaction terms, show the efficiency of proposed schemes.

1 Introduction

This paper concerns numerical methods for three dimensional advection–diffusion-reaction (ADR) models governed by the following system of equations

$$\frac{\partial \mathbf{c}}{\partial t} + \nabla \cdot (\mathbf{vc}) - \nabla \cdot (\mathrm{D}\nabla \mathbf{c}) = \mathbf{R}(\mathbf{c}) \,, \tag{1}$$

where $\mathbf{c} = \mathbf{c}(\mathbf{x}, t) \in \mathbb{R}^m$, $\mathbf{x} \in \Omega \subset \mathbb{R}^3$ are the space variables and t denotes the time. The diffusion coefficient matrix $\mathrm{D} = diag[d_{11}, d_{22}, \ldots, d_{mm}]$ and the velocity field $\mathbf{v}(\mathbf{x}) \in \mathbb{R}^3$ are, usually, supposed to be given. Several phenomena of relevant interest can be described by model (1). Among others, we can quote the applications to a chemotaxis model [8], the pollutant transport in atmosphere [11], mucilage dynamics [4], ash-fall from volcano [5], and groundwater and surface water [9]. The governing system takes into account physical and biological processes modelled by three distinct terms: transport of each component due to the velocity field $\mathbf{v}$, described by the advection terms; random motion of each component due to the turbulent nature of the flow field, modelled by the (turbulent) diffusion terms; interaction of the involved species described by reaction terms (e.g., chemical reactions,

R. Fazio (✉) and A. Jannelli
Department of Mathematics, University of Messina, Salita Sperone 31, 98166 Messina, Italy
e-mail: rfazio@dipmat.unime.it, jannelli@dipmat.unime.it

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_33, © Springer-Verlag Berlin Heidelberg 2010

growth of species, consumption of nutrients, etc.). From the numerical view-point, for the time integration of different terms of the model (1), we propose a fractional step approach. This method consists in separating in the discretized equations the part that accounts for hydrodynamics, described by advection term, usually linear, and the diffusion term on the left hand side, from the part accounting for biology, described by nonlinear reaction term on the right hand side. This splitting is reasonable when a loose coupling exists between the different phenomena and when they evolve with different characteristic times. The coupling between the components in each grid point, and not over the grid points, appears only in the solution of the reaction equations. In this contest such assumptions holds and the use of a fractional step seems promising.

2 The Operator Splitting Approach

In this section, we describe an efficient algorithm for solving ADR models (1) written in the following form

$$\frac{\partial \mathbf{c}}{\partial t} = A(\mathbf{c}) + D(\mathbf{c}) + R(\mathbf{c}). \tag{2}$$

We propose the use the Strang splitting [7] approach: if $\mathbf{c}^n$ is the approximate solution at time t^n, we obtain the solution $\mathbf{c}^{n+1}$ at next time $t^{n+1} = t^n + \Delta t$ by the following sequence of five steps:

$$\mathbf{c}^{n+1} = \mathscr{A}(\Delta t/2)\mathscr{D}(\Delta t/2)\mathscr{R}(\Delta t)\mathscr{D}(\Delta t/2)\mathscr{A}(\Delta t/2)\,\mathbf{c}^n,$$

where $\mathscr{A}(\cdot)$, $\mathscr{D}(\cdot)$ and $\mathscr{R}(\cdot)$ represent the discretized advection, diffusion and reaction operators, respectively. The advantage of the fractional step method is that, for each term, a different time integration method can be chosen. For the time integration of the advection part, explicit methods are usually more efficient than the implicit ones. On the other hand, the reaction part is sometimes very stiff and this requires the use of implicit methods, used also for the diffusion term. As far as accuracy is concerned, by using this splitting technique we get second order accuracy provided that each subproblem is solved by a second order accurate method.

2.1 Advection Solver

In this section, we consider the homogeneous hyperbolic equations

$$\frac{\partial \mathbf{c}}{\partial t} + \nabla \cdot (\mathbf{v}\mathbf{c}) = 0, \tag{3}$$

with given initial condition and appropriate boundary conditions (for instance: Dirichlet conditions at the inflow and no conditions at the outflow boundaries, or periodic boundary conditions, etc.). We set an uniform Cartesian grid $\Omega_J \in \mathbb{R}^3$. Let $\mathbf{c}^n_{ijk}$ be the average value of $\mathbf{c}$ over cell (x_i, y_j, z_k) at current time t^n, and $\mathbf{c}^{n+1}_{ijk}$ the average value of $\mathbf{c}$ at time $t^n + \Delta t$. For time integration, we use a high-resolution finite volume method written in the conservative form

$$\mathbf{c}^{n+1}_{ijk} = \mathbf{c}^n_{ijk} + \\ -\frac{\Delta t}{\Delta x}\left[\mathbf{F}^n_{i+\frac{1}{2}jk} - \mathbf{F}^n_{i-\frac{1}{2}jk}\right] - \frac{\Delta t}{\Delta y}\left[\mathbf{G}^n_{ij+\frac{1}{2}k} - \mathbf{G}^n_{ij-\frac{1}{2}k}\right] - \frac{\Delta t}{\Delta z}\left[\mathbf{H}^n_{ijk+\frac{1}{2}} - \mathbf{H}^n_{ijk-\frac{1}{2}}\right]$$

where $\mathbf{F}$, $\mathbf{G}$ and $\mathbf{H}$ are intercell numerical fluxes. A recent study on first and second order positive numerical methods for the advection equation is developed in [1] where several test problems are solved.

2.2 Diffusion Solver

The diffusion term is discretized implicitly to avoid using small time steps when are not dictated by accuracy reasons in detecting the correct dynamics of the concentration. We use the Crank–Nicolson scheme because it is second order accurate in space and time,

$$\mathbf{c}^{n+1}_{i,j,k} - \frac{\Delta t}{2\Delta x \Delta y \Delta z}\mathbf{w}^{n+1}_{i,j,k} = \mathbf{c}^n_{i,j,k} + \frac{\Delta t}{2\Delta x \Delta y \Delta z}\mathbf{w}^n_{i,j,k} \tag{4}$$

where

$$\mathbf{w}^n_{i,j,k} = -\Big\{\Delta y \Delta z\left(\widehat{\mathbf{F}}^n_{i+\frac{1}{2},j,k} - \widehat{\mathbf{F}}^n_{i-\frac{1}{2},j,k}\right) \\ + \Delta x \Delta z\left(\widehat{\mathbf{G}}^n_{i,j+\frac{1}{2},k} - \widehat{\mathbf{G}}^n_{i,j-\frac{1}{2},k}\right) + \Delta x \Delta y\left(\widehat{\mathbf{H}}^n_{i,j,k+\frac{1}{2}} - \widehat{\mathbf{H}}^n_{i,j,k-\frac{1}{2}}\right)\Big\} \tag{5}$$

with

$$\widehat{\mathbf{F}}_{i+\frac{1}{2},j,k} = -d_{i+\frac{1}{2},j,k}\frac{\mathbf{c}_{i+1,j,k} - \mathbf{c}_{i,j,k}}{\Delta x}, \\ \widehat{\mathbf{G}}_{i,j+\frac{1}{2},k} = -d_{i,j+\frac{1}{2},k}\frac{\mathbf{c}_{i,j+1,k} - \mathbf{c}_{i,j,k}}{\Delta y}, \\ \widehat{\mathbf{H}}_{i,j,k+\frac{1}{2}} = -d_{i,j,k+\frac{1}{2}}\frac{\mathbf{c}_{i,j,k+1} - \mathbf{c}_{i,j,k}}{\Delta z}. \tag{6}$$

As far as stability is concerned, the Crank–Nicolson scheme is an unconditionally stable one. We have no restriction on the time step but the extra labour involved is

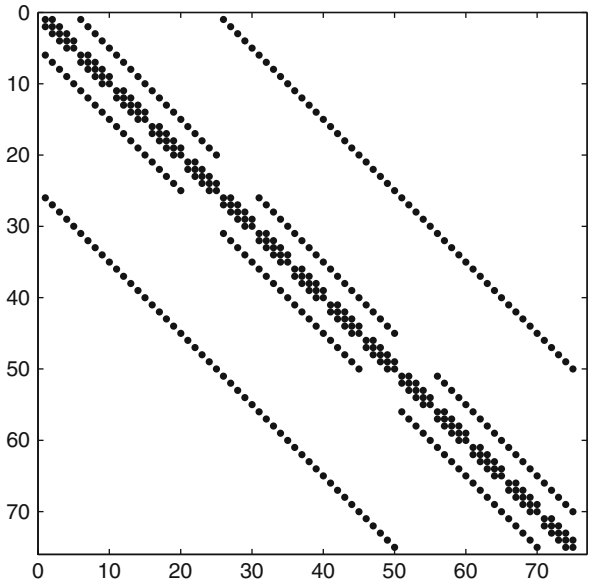


Fig. 1 Example of matrix of coefficients for the Crank–Nicolson method

very considerable. We have to solve a system of linear equations. These equations have a regular structure, each involving at most seven unknowns. The matrix of the system consists of zeroes, but it has not tridiagonal form. The linear system obtained is solved by the bi-conjugate gradient method of Van der Vorst [10] (for a simple description of the method see [2, pp. 362–379]). Figure 1 shows the matrix of coefficients on a sample domain of $5 \times 5 \times 3$ mesh-points. Note that there could be some instability in coupling with the reaction term. The presence of diffusion term in the system may cause some instabilities. When we individually test each step in the Strang splitting procedure, they are stable for reasonable time step intervals. When we test the coupled diffusion and reaction steps they could be unstable. When the full model is solved numerically, the time step interval necessary to prevent instability is very small when the diffusion term is discretized with Crank–Nicolson. A much longer time step is possible when diffusion step is discretized with the TR-BDF2, as Tyson et al. have done in [8], here TR stands for Trapezoidal Rule and BDF2 for the second order Backward Differentiation Formula.

2.2.1 Test Problem: Heat Equation

As an example, we consider the heat equation

$$\frac{\partial c}{\partial t} = \frac{\partial^2 c}{\partial x^2} + \frac{\partial^2 c}{\partial y^2} \tag{7}$$

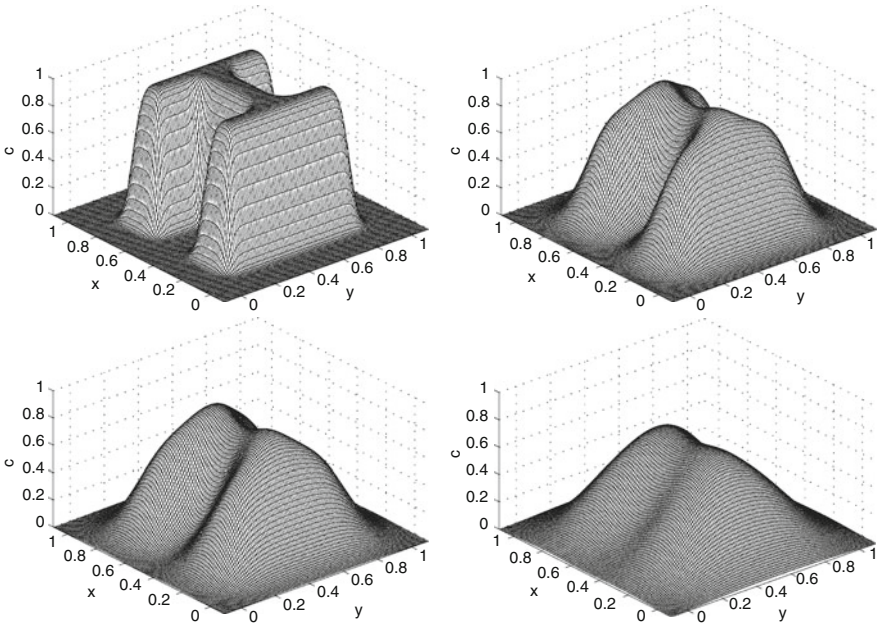


Fig. 2 The numerical solution and at time $t = 0.001$, $t = 0.003$, $t = 0.005$, and final time $t = 0.01$

on the unit square $0 < x < 1$, $0 < y < 1$, with homogeneous Dirichlet boundary conditions $c = 0$ on the boundary of the unit square. The initial condition is $c(x, y, 0) = f(x, y)$ with $f(x, y) = 1$ within the region shaped like the letter H, and $f(x, y) = 0$ in the rest of the square. In a narrow band surrounding the H, the function increases from 0 to 1, so that $f(x, y)$ is continuous; its derivatives are not continuous, being zero everywhere outside the narrow band and being greater than zero inside the band. The results of the implicit method are shown in Fig. 2. It shows the way in which the initial function diffuses throughout the square. This numerical results are obtained using $\Delta x = \Delta y = 0.01$ and $\Delta t = 0.001$ with $t_{max} = 0.01$.

2.3 Reaction Solver

The reaction step consists of solving a coupled system of ordinary differential equations in each grid cell. There are no spatial derivatives and hence no spatial coupling of different cells in this step. Moreover, the reaction equations are sometimes very stiff, requiring the use of implicit methods for stability reasons. In this contest, we propose the use of an adaptive procedure implemented with stiff solvers at low accuracy and complexity. In particular, we use the Milne device for the estimation of the local error, that is the error incurred in the integration from t^n to t^{n+1} under the

assumption that the approximate solution at time t^n is exact. In order to implement the Milne device, we use two different convergent multistep methods of same order of accuracy p in order to decide whether the numerical value is an acceptable approximation to exact solution evaluated at time t^{n+1}. Let us denote by $\mathbf{c}^{n+1}$ and $\tilde{\mathbf{c}}^{\,n+1}$ the two computed numerical approximations, and with C and $\widetilde{C}$ the corresponding local error constants. A naive approach is to require that the local error LE satisfies

$$\mathrm{LE} = \left| \frac{C}{\widetilde{C} - C} \right| ||\mathbf{c}^{n+1} - \tilde{\mathbf{c}}^{\,n+1}|| \leq tol, \tag{8}$$

with $||\cdot||$ a suitable norm.

2.3.1 Numerical Results: Robertson Problem

As sample numerical test, we consider the problem given by a stiff system of three non-linear differential equations with suitable initial conditions

$$\begin{cases} c_1' = -P_1 c_1 + P_3 c_2 c_3 \\ c_2' = P_1 c_1 - P_3 c_2 c_3 - P_2 c_2^2 \\ c_3' = P_2 c_2^2 \\ c_1(0) = 1, c_2(0) = 0, c_3(0) = 0, \end{cases} \tag{9}$$

where $P_1 = 0.04$, $P_2 = 3 \cdot 10^7$ and $P_3 = 10^4$. The model describes the kinetics of an auto-catalytic reaction described by Robertson [6]. The structure of reaction is reported in (10), where A, B and C represent the chemical species involved

$$\begin{aligned} A &\xrightarrow{P_1} B, \\ B + B &\xrightarrow{P_2} C + B, \\ B + C &\xrightarrow{P_3} A + C. \end{aligned} \tag{10}$$

This problem is sometimes used as a test problem for stiff solvers. The large difference among the reaction rate constants P_i, with $i = 1, 2, 3$, is the reason for the stiffness. As usual in problems arising in chemical kinetics, this system has a small very quick initial transient. This phase is following by a very smooth variation of the components where a large step-size would be appropriate for a numerical method. The problem (9) is integrated within the range $t \in [0, 10^6]$. Figure 3 shows the numerical solution of the species involved.

The numerical results are obtained in 267 steps (with 3 rejected steps) by Milne device implemented with the TR with $\widetilde{C} = -1/12$, and BDF2 with variable time steps, see [3], with

$$C = -\frac{(k^n + 1)^2}{6k^n(2k^n + 1)},$$

where $k^n = \Delta t^n / \Delta t^{n-1}$. Figure 4 shows the adaptive numerical results. In the top frame, we show the step-size selection Δt^n, in the bottom one the local error LE.

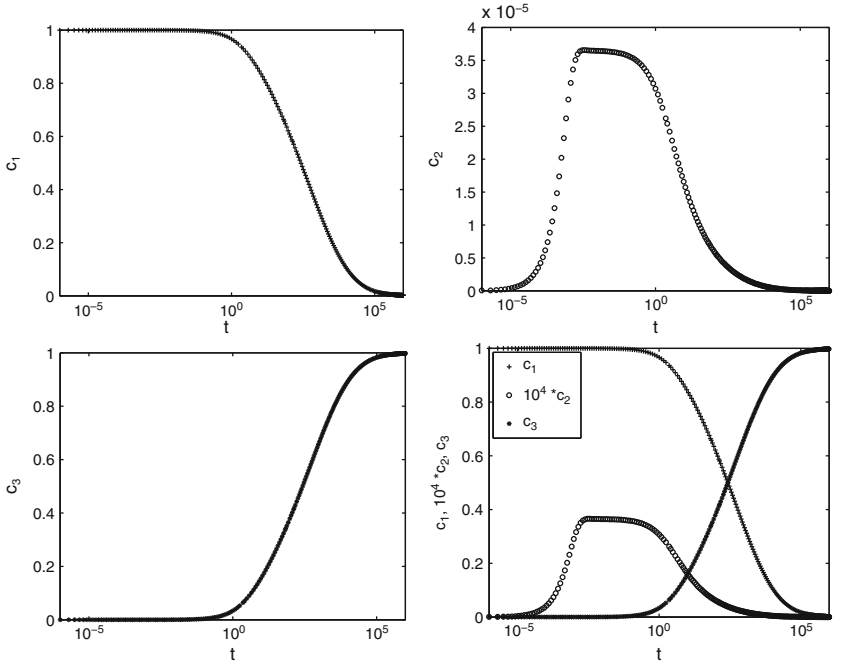


Fig. 3 Semi-log scale plot of numerical solution for the Robertson problem

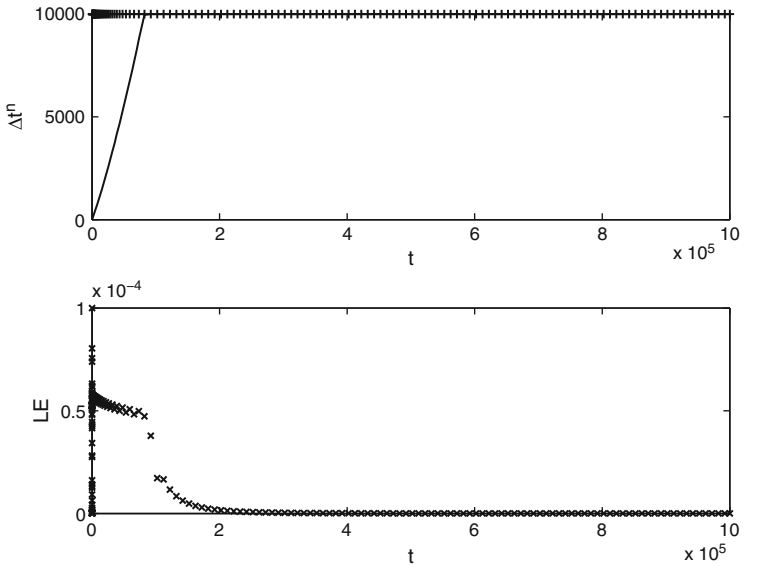


Fig. 4 Adaptive numerical results for the Robertson problem

It is easy to note, how, the adaptive procedure modifies the time step in relation to the value of the the local error for the solution second component. Initially, at the beginning of the process, the adaptive procedure sets a small Δt^n corresponding to fast transitory of the second component. Then, when this component becomes smooth, the procedure amplifies the step-size. A maximum value for step-size is set and this represents its upper bound.

For the adaptive procedure, we set: $\Delta t_{\min} \leq \Delta t^n \leq \Delta t_{\max}$ with $\Delta t_{\min} = 10^{-6}$ and $\Delta t_{\max} = 10^4$, $LE_{\min} \leq LE \leq LE_{\max}$ with $LE_{\min} = 10^{-5}$ and $LE_{\max} = 10\, LE_{\min}$. The time-step Δt^n is modified in the following cases: if $LE_{min} \leq LE \leq LE_{max}$, then $\Delta t^{n+1} = 0.9\, \Delta t^n\, (tol/LE)^{1/(p+1)}$, $p = 2$ in our case; if $LE < LE_{\min}$ then $\Delta t^{n+1} = 1.2\, \Delta t^n$; if $LE > LE_{\max}$ then the step is repeated with $\Delta t^n = 0.5\, \Delta t^n$.

References

1. R. Fazio, A. Jannelli: Second order positive schemes by means of flux limiters for the advection equation. IAENG Int. J. Appl. Math., **39**, 25–35, 2009
2. G. H. Golub, C. F. van Loan: *Matrix computations*, 2nd ed. Hopkins University Press, London, 1989
3. A. Jannelli, R. Fazio: Adaptive stiff solvers at low accuracy and complexity. J. Comp. Appl. Math., **191**, 246–258, 2006
4. A. Jannelli, R. Fazio, D. Ambrosi: A 3D mathematical model for the prediction of mucilage dynamics. Comput. Fluids, **32**, 47–57, 2003
5. R. McKibbin, L. L. Lim, T. A. Smith, W. L. Sweatman: A model for dispersal of eruption ejecta. In *Proceedings World Geothermal Congress*. 24–29 April 2005, Antalya, Turkey
6. H. H. Robertson: *The solution of a set of reaction rate equations*. Academic Press, London, 178–182, 1966
7. G. Strang: On the construction and comparison of difference schemes. SIAM J. Num. Anal., **5**, 506–517, 1968
8. R. Tyson, L. G. Stern, R. J. LeVeque: Fractional step methods applied to a chemotaxis model. J. Math. Biol., **41**, 455–475, 2000
9. M. Toro, L. C. van Rijn, K. Meijer: Three-dimensional modelling of sand and mud transport in currents and waves. Technical Report No. H461/Q407/Q791, Delft Hydraulics, Delft, The Netherlands, 1989
10. H. A. Van der Vorst: Bi-CGSTAB: A fast and smootly converging variant of Bi-CG for the solution of nonsymmetric linear systems. SIAM J. Sci. Stat. Comput., **13**, 631–644, 1992
11. J. G. Verwer, W. H. Hundsdorfer, J. G. Blom: Numerical time integration for air pollution models. Sur. Math. Ind., **2**, 107–174, 2002

Space-Time DG Method for Nonstationary Convection–Diffusion Problems

Miloslav Feistauer, Václav Kučera, Karel Najzar, and Jaroslava Prokopová

Abstract The paper is concerned with the theory of the discontinuous Galerkin finite element method for the space-time discretization of a nonlinear nonstationary convection–diffusion initial-boundary value problem. The discontinuous Galerkin method is applied separately in space and time using, in general, different space grids on different time levels and different polynomial degrees p and q in space and time dicretization. The analysis of error estimates is described.

1 Introduction

In a number of applications we meet the necessity to solve complicated partial differential equations. In some cases it is suitable to carry out the space discretization by the discontinuous Galerkin finite element method (DGFEM) using piecewise polynomial approximation of a sought solution without any requirement on the continuity between neighboring elements. See, e.g., [2, 3].

The numerical simulation of strongly nonstationary transient problems requires the application of numerical schemes of high order of accuracy in space as well as in time. From this point of view, it appears suitable to use the discontinuous Galerkin discretization with respect to both space and time. The discontinuous Galerkin time discretization was introduced and analyzed, e.g., in [10] for the solution of ordinary differential equations. In [1, 11–13] the solution of parabolic problems is carried out with the aid of conforming finite elements in space combined with the DG time discretization. In the present paper we are concerned with the space-time discontinuous Galerkin discretization applied separately in space and in time to the numerical solution of the following nonstationary nonlinear convection-diffusion problem.

M. Feistauer (✉), V. Kučera, K. Najzar, and J. Prokopová
Charles University Prague, Faculty of Mathematics and Physics, Sokolovská 83,
186 75 Praha 8, Czech Republic
e-mail: feist@karlin.mff.cuni.cz

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_34,

Let $\Omega \subset I\!R^d$ ($d = 2$ or 3) be a bounded polyhedral domain and $T > 0$. We want to find $u : Q_T = \Omega \times (0,T) \to I\!R$ such that

$$\frac{\partial u}{\partial t} + \sum_{s=1}^{d} \frac{\partial f_s(u)}{\partial x_s} - \varepsilon \Delta u = g \quad \text{in } Q_T = \Omega \times (0,T), \tag{1}$$

$$u\big|_{\partial\Omega\times(0,T)} = u_D, \tag{2}$$

$$u(x,0) = u^0(x), \quad x \in \Omega. \tag{3}$$

We assume that $\varepsilon > 0$ is a constant, g, u_D, u^0, f_s are fiven functions and $f_s \in C^1(I\!R), \quad |f_s'| \le C,\ s = 1,\dots,d$. This means that the functions f_s (called fluxes of the quantity u) are Lipschitz-continuous in $I\!R$.

2 Space-Time Discretization

In the time interval $[0,T]$ we shall construct a partition formed by time instants $0 = t_0 < \cdots < t_M = T$ and denote $I_m = (t_{m-1}, t_m),\ \tau_m = t_m - t_{m-1}$. For each I_m we consider a partition $\mathcal{T}_{h,m}$ of the closure $\overline{\Omega}$ of the domain Ω into a finite number of closed triangles for $d = 2$ and tetrahedra for $d = 3$ with mutually disjoint interiors. The partitions $\mathcal{T}_{h,m}$ are in general different for different m.

By $\mathcal{F}_{h,m}$ we denote the system of all faces of all elements $K \in \mathcal{T}_{h,m}$. Further, we define the set of all inner faces by $\mathcal{F}^I_{h,m} = \{\Gamma \in \mathcal{F}_{h,m};\ \Gamma \subset \Omega\}$ and by $\mathcal{F}^B_{h,m} = \{\Gamma \in \mathcal{F}_{h,m};\ \Gamma \subset \partial\Omega\}$ the set of all boundary faces. Each $\Gamma \in \mathcal{F}_{h,m}$ will be associated with a unit normal vector n_Γ. We assume that for $\Gamma \in \mathcal{F}^B_{h,m}$ the normal n_Γ has the same orientation as the outer normal to $\partial\Omega$. We set $h_K = \operatorname{diam}(K)$ for $K \in \mathcal{T}_{h,m}$, $h_m = \max_{K\in\mathcal{T}_{h,m}} h_K,\ h = \max_{m=1,\dots,M} h_m$. By ρ_K we denote the radius of the largest d-dimensional ball inscribed into K and by $|K|$ we denote the d-dimensional Lebesgue measure of K. Finally, we set $\tau = \max_{m=1,\dots,M} \tau_m$.

For a function φ defined in $\bigcup_{m=1}^M I_m$ we denote $\varphi_m^\pm = \varphi(t_m\pm) = \lim_{t\to t_m\pm} \varphi(t)$, $\{\varphi\}_m = \varphi(t_m+) - \varphi(t_m-)$.

Over a triangulation $\mathcal{T}_{h,m}$ we define the *broken Sobolev spaces* $H^k(\Omega, \mathcal{T}_{h,m}) = \{v; v|_K \in H^k(K)\ \forall K \in \mathcal{T}_{h,m}\}$ with seminorm $|v|_{H^k(\Omega,\mathcal{T}_{h,m})} = \big(\sum_{K\in\mathcal{T}_{h,m}} |v|^2_{H^k(K)}\big)^{1/2}$.

For each face $\Gamma \in \mathcal{F}^I_{h,m}$ there exist two neighbours $K_\Gamma^{(L)}, K_\Gamma^{(R)} \in \mathcal{T}_{h,m}$ such that $\Gamma \subset \partial K_\Gamma^{(L)} \cap \partial K_\Gamma^{(R)}$. We use the convention that n_Γ is the outer normal to $\partial K_\Gamma^{(L)}$ and the inner normal to $\partial K_\Gamma^{(R)}$. For $v \in H^1(\Omega, \mathcal{T}_{h,m})$ and $\Gamma \in \mathcal{F}^I_{h,m}$ we introduce the following notation: $v|_\Gamma^{(L)}$ = the trace of $v|_{K_\Gamma^{(L)}}$ on Γ, $v|_\Gamma^{(R)}$ = the trace of $v|_{K_\Gamma^{(R)}}$ on Γ, $\langle v\rangle_\Gamma = \frac{1}{2}(v|_\Gamma^{(L)} + v|_\Gamma^{(R)})$, $[v]_\Gamma = v|_\Gamma^{(L)} - v|_\Gamma^{(R)}$.

Let $C_W > 0$ be a fixed constant. We set

$$h(\Gamma) = \frac{h_{K_\Gamma^{(L)}} + h_{K_\Gamma^{(R)}}}{2C_W} \quad \text{for } \Gamma \in \mathscr{F}_{h,m}^I, \quad h(\Gamma) = \frac{h_{K_\Gamma^{(L)}}}{C_W} \quad \text{for } \Gamma \in \mathscr{F}_{h,m}^B. \tag{4}$$

By $(\cdot,\cdot)$ we denote the scalar product in $L^2(\Omega)$ and by $\|\cdot\|$ we denote the norm in $L^2(\Omega)$. If $\overline{u},\ \varphi \in H^2(\Omega, \mathscr{T}_{h,m})$, we define the forms

$$\begin{aligned} a_{h,m}(\overline{u},\varphi) &= \varepsilon \sum_{K\in\mathscr{T}_{h,m}} \int_K \nabla \overline{u}\cdot\nabla\varphi \,\mathrm{d}x - \varepsilon \sum_{\Gamma\in\mathscr{F}_{h,m}^I} \int_\Gamma (\langle\nabla\overline{u}\rangle\cdot n_\Gamma[\varphi] - \theta\langle\nabla\varphi\rangle\cdot n_\Gamma[\overline{u}])\,\mathrm{d}S \\ &\quad - \varepsilon \sum_{\Gamma\in\mathscr{F}_{h,m}^B} \int_\Gamma (\nabla\overline{u}\cdot n_\Gamma\,\varphi - \theta\,\nabla\varphi\cdot n_\Gamma\,\overline{u})\,\mathrm{d}S, \end{aligned} \tag{5}$$

$$J_{h,m}(\overline{u},\varphi) = \sum_{\Gamma\in\mathscr{F}_{h,m}^I} h(\Gamma)^{-1}\int_\Gamma [\overline{u}]\,[\varphi]\,\mathrm{d}S + \sum_{\Gamma\in\mathscr{F}_{h,m}^B} h(\Gamma)^{-1}\int_\Gamma \overline{u}\,\varphi\,\mathrm{d}S,$$

$$A_{h,m} = a_{h,m} + \varepsilon J_{h,m}, \tag{6}$$

$$\begin{aligned} b_{h,m}(\overline{u},\varphi) &= -\sum_{K\in\mathscr{T}_{h,m}} \int_K \sum_{s=1}^d f_s(\overline{u})\,\frac{\partial\varphi}{\partial x_s}\,\mathrm{d}x + \sum_{\Gamma\in\mathscr{F}_{h,m}^I}\int_\Gamma H\left(\overline{u}\big|_\Gamma^{(L)}, \overline{u}\big|_\Gamma^{(R)}, n_\Gamma\right)[\varphi]\big|_\Gamma\,\mathrm{d}S \\ &\quad + \sum_{\Gamma\in\mathscr{F}_{h,m}^B}\int_\Gamma H\left(\overline{u}\big|_\Gamma^{(L)}, \overline{u}\big|_\Gamma^{(L)}, n_\Gamma\right)\varphi\big|_\Gamma\,\mathrm{d}S. \end{aligned} \tag{7}$$

$$\ell_{h,m}(\varphi) = (g,\varphi) + \varepsilon \sum_{\Gamma\in\mathscr{F}_{h,m}^B}\left(h(\Gamma)^{-1}\int_\Gamma u_D\,\varphi\,\mathrm{d}S + \theta\int_\Gamma \nabla\varphi\cdot n_\Gamma u_D\,\mathrm{d}S\right) \tag{8}$$

In (7), H is a numerical flux with the following properties.

(H1) $H(u,v,n)$ is defined in $\mathbb{R}^2 \times B_1$, where $B_1 = \{n \in \mathbb{R}^d; |n| = 1\}$, and is *Lipschitz-continuous* with respect to $u,\ v$.

(H2) $H(u,v,n)$ is *consistent*: $H(u,u,n) = \sum_{s=1}^d f_s(u)\,n_s, u \in \mathbb{R}, n = (n_1,\ldots,n_d) \in B_1$.

(H3) $H(u,v,n)$ is *conservative*: $H(u,v,n) = -H(v,u,-n),\ u,\ v \in \mathbb{R},\ n \in B_1$.

In the above forms we take $\theta = -1,\ \theta = 0,\ \theta = 1$ and obtain the symmetric (SIPG), incomplete (IIPG) and nonsymmetric (NIPG) variants of the approximation of the diffusion terms, respectively.

In the space $H^1(\Omega, \mathscr{T}_{h,m})$, the following norm will be used:

$$\|\varphi\|_{\mathrm{DG},m} = \Big(\sum_{K\in\mathscr{T}_{h,m}} |\varphi|^2_{H^1(K)} + J_{h,m}(\varphi,\varphi)\Big)^{1/2}. \tag{9}$$

Let $p,\ q \geq 1$ be integers. For each $m = 1,\ldots,M$ we define the finite-dimensional space

$$S_{h,m}^p = \left\{\varphi \in L^2(\Omega); \varphi|_K \in P^p(K)\ \forall\, K \in \mathscr{T}_{h,m}\right\}. \tag{10}$$

We denote by Π_m the $L^2(\Omega)$-projection on $S^p_{h,m}$. The approximate solution will be sought in the space

$$S^{p,q}_{h,\tau} = \left\{\varphi \in L^2(Q_T); \varphi|_{I_m} = \sum_{i=0}^{q} t^i\, \varphi_i \quad \text{with } \varphi_i \in S^p_{h,m},\ m = 1, \ldots, M\right\}. \tag{11}$$

In what follows we shall use the notation $U' = \partial U/\partial t, u' = \partial u/\partial t, D^{q+1} = \partial^{q+1}/\partial t^{q+1}$.

Definition 1. We say that a function U is an approximate solution of problem (1) – (3), if $U \in S^{p,q}_{h,m}$ and

$$\int_{I_m} \left((U', \varphi) + A_{h,m}(U, \varphi) + b_{h,m}(U, \varphi)\right) \mathrm{d}t + \left(\{U\}_{m-1}, \varphi^+_{m-1}\right) \tag{12}$$
$$= \int_{I_m} \ell_{h,m}(\varphi)\, \mathrm{d}t, \quad \forall\, \varphi \in S^{p,q}_{h,\tau}, \quad m = 1, \ldots, M,\ U_0^- := \Pi_1 u^0.$$

The exact sufficiently regular solution u satisfies the identity

$$\int_{I_m} \left((u', \varphi) + A_{h,m}(u, \varphi) + b_{h,m}(u, \varphi)\right) \mathrm{d}t + \left(\{u\}_{m-1}, \varphi^+_{m-1}\right) \tag{13}$$
$$= \int_{I_m} \ell_{h,m}(\varphi)\, \mathrm{d}t \quad \forall\, \varphi \in S^{p,q}_{h,\tau}, \quad \text{with } u(0-) = u(0).$$

It is also possible to consider $q = 0$. In this case, scheme (12) represents a version of the backward Euler method. Since it can be analyzed in a similar way as, for example, in [6], we shall be concerned only with $q \geq 1$.

3 Error Analysis

In the derivation of the error we shall use the $S^{p,q}_{h,\tau}$-interpolation π of functions $v \in H^1(0, T; L^2(\Omega))$ defined by

$$\text{a)} \quad \pi\, v \in S^{p,q}_{h,\tau}, \quad \text{b)} \quad (\pi\, v)\,(t_m-) = \Pi_m\, v(t_m-), \tag{14}$$
$$\text{c)} \quad \int_{I_m} (\pi v - v,\ \varphi^*)\, \mathrm{d}t = 0 \quad \forall\, \varphi^* \in S^{p,q-1}_{h,\tau}, \quad \forall\, m = 1, \ldots, M.$$

It is possible to prove that πu is uniquely determined and $\pi v|_{I_m} = \pi(\Pi_m v)|_{I_m}$.

Our main goal will be the analysis of the estimation of the error $e = U - u$, which can be expressed in the form $e = \xi + \eta$, where $\xi = U - \pi u \in S^{p,q}_{h,\tau}$, $\eta = \pi u - u$. Then, in virtue of (12) and (13),

$$\int_{I_m} \big((\xi', \varphi) + A_{h,m}(\xi, \varphi)\big)\,\mathrm{d}t + \left(\{\xi_{m-1}\}, \varphi^+_{m-1}\right) = \int_{I_m} \big(b_{h,m}(u, \varphi) - b_{h,m}(U, \varphi)\big)\,\mathrm{d}t$$
$$- \int_{I_m} \big((\eta', \varphi) + A_{h,m}(\eta, \varphi)\big)\,\mathrm{d}t - \left(\{\eta\}_{m-1}, \varphi^+_{m-1}\right) \quad \forall \varphi \in S^{p,q}_{h,\tau}. \tag{15}$$

3.1 Derivation of an Abstract Error Estimate

In our further considerations, by C and c we shall denote positive generic constants, independent of $h, \tau, K, \varepsilon, u, U$, which can attain different values in different places. In the sequel, we shall consider a system of triangulations $\mathscr{T}_{h,m}$, $m = 1, \ldots, M$, $h \in (0, h_0)$, which is shape regular and locally quasiuniform: There exist positive constants C_R and C_Q, independent of K, Γ, m and h, such that for all $m = 1, \ldots, M$ and $h \in (0, h_0)$

$$\frac{h_K}{\rho_K} \le C_R, \quad \forall K \in \mathscr{T}_{h,m}, \tag{16}$$
$$h_{K^{(L)}_\Gamma} \le C_Q\, h_{K^{(R)}_\Gamma}, \quad h_{K^{(R)}_\Gamma} \le C_Q\, h_{K^{(L)}_\Gamma} \quad \forall \Gamma \in \mathscr{F}^I_{h,m}. \tag{17}$$

Important tools in the analysis of the DGFEM are the multiplicative trace inequality (see, e.g., [7]), the inverse inequality ([4]), the coercivity of the form $A_{h,m}$ ([8]) and the consistency of the form $b_{h,m}$ obtained in a similar way as in [5].

Let us substitute $\varphi := \xi$ in (15). Then a detailed and rather technical analysis yields the estimate

$$\left\|\xi^-_m\right\|^2 - \left\|\xi^-_{m-1}\right\|^2 + \varepsilon\left(1 - \frac{2}{k}\right)\int_{I_m} \|\xi\|^2_{\mathrm{DG},m}\,\mathrm{d}t \tag{18}$$
$$\le \frac{C}{\varepsilon}\int_{I_m} \|\xi\|^2\,\mathrm{d}t + 2\left\|\eta^-_{m-1}\right\|^2 + C\int_{I_m} R_m(\eta)\,\mathrm{d}t,$$

where $k > 2$ and

$$R_m(\eta) = \varepsilon\left(\|\eta\|^2_{\mathrm{DG},m} + \sum_{K\in\mathscr{T}_{h,m}} h^2_K |\eta|^2_{H^2(K)}\right) + \frac{1}{\varepsilon}\left(\sum_{K\in\mathscr{T}_{h,m}} \|\eta\|^2_{L^2(K)} + h^2_K |\eta|^2_{H^1(K)}\right). \tag{19}$$

Further, it is necessary to estimate $\int_{I_m} \|\xi\|^2\,\mathrm{d}t$. We apply here the approach from [1] based on the use of the so-called Gauss–Radau quadrature and interpolation. In the interval $(t_{m-1}, t_m]$ we shall consider the Gauss–Radau quadrature formula

$$\int_{I_m} \varphi(t)\,\mathrm{d}t \approx \tau_m \sum_{i=1}^{q+1} w_i \varphi\left(t^{m,i}\right), \tag{20}$$

with weights $w_i > 0$ and integration points $t^{m,i} = t_{m-1} + \tau_m \vartheta_i$, where $0 < \vartheta_1 < \cdots < \vartheta_{q+1} = 1$ are Gauss–Radau points. Formula (20) is exact for polynomials of degree $\le 2q$.

Now, in (15) we set $\varphi := \tilde{\xi}$, which is defined as the Lagrange interpolation of $\tau_m \xi(t)/(t - t_{m-1})$ at the points $t^{m,i}$, $i = 1, \ldots, q+1$. This means that for each $x \in \Omega$, the function $\tilde{\xi}(\cdot, x)$ is a polynomial (in t) of degree $\le q$. A rather technical analysis proves the existence of constants $C, C^* > 0$ such that

$$\int_{I_m} \|\xi\|^2 \,\mathrm{d}t \le C\,\tau_m \left(\left\|\xi_{m-1}^-\right\|^2 + \left\|\eta_{m-1}^-\right\|^2 + \int_{I_m} R_m(\eta)\,\mathrm{d}t \right), \tag{21}$$

provided

$$0 < \tau_m \le C^* \varepsilon. \tag{22}$$

From (18) with $k := 8$, (21), discrete Gronwall's lemma and the relation $e = \xi + \eta$ we get the *abstract error estimate.*

Theorem 1. *Let (22) hold. Then there exist constants $C, c > 0$ such that the error $e = U - u$ satisfies the estimate*

$$\begin{aligned} &\|e_m^-\|^2 + \frac{\varepsilon}{2} \sum_{j=1}^{m} \int_{I_m} \|e\|_{DG,j}^2 \,\mathrm{d}t \\ &\le C \exp(c t_m/\varepsilon) \left(\sum_{j=1}^{m} \|\eta_j^-\|^2 + \sum_{j=1}^{m} \int_{I_j} R_j(\eta)\,\mathrm{d}t \right) + 2\|\eta_m^-\|^2 + 2\varepsilon \sum_{j=1}^{m} \int_{I_j} \|\eta\|_{DG,j}^2 \,\mathrm{d}t, \\ &\quad m = 1, \ldots, M. \end{aligned} \tag{23}$$

3.2 Error Estimation in Terms of h and τ

The derivation of error estimates in dependence on h and τ is obtained from the abstract error estimate and estimation of terms containing η, under the assumption

$$u \in H^{q+1}\big(0, T; H^1(\Omega)\big) \cap C([0, T]; H^{p+1}(\Omega)), \tag{24}$$

and the assumption that the meshes satisfy conditions (16), (17), (22) and

$$\tau_m \ge C h_m^2, \quad m = 1, \ldots, M. \tag{25}$$

We use approximation properties of the operators Π_m and the estimate

$$\begin{aligned} &\|\pi\varphi(x,\cdot) - \varphi(x,\cdot)\|_{L^2(I_m)}^2 \le C\,\tau_m^{2q+2} \left\| D^{q+1}\varphi(x,\cdot)\right\|_{L^2(I_m)}^2, \\ &\quad x \in K,\ K \in \mathscr{T}_{h,m},\ m = 1, \ldots, M,\ h \in (0, h_0), \end{aligned} \tag{26}$$

obtained by scaling arguments for $\varphi \in H^{q+1}(I_m, S_{h,m}^p)$. Further, we show that for $j, m = 1, \ldots, M,\ K \in \mathcal{T}_{h,m}$ and $h \in (0, h_0)$ we have

$$\|\eta_j^-\|^2 \le C h^{p+1} |u(t_j)|_{H^{p+1}(\Omega)}, \tag{27}$$

$$\int_{I_m} \|\eta\|^2_{L^2(K)} \mathrm{d}t \le C \left(h_K^{2(p+1)} |u|^2_{L^2(I_m, H^{p+1}(K))} + \tau_m^{2(q+1)} |u|^2_{H^{q+1}(I_m, L^2(K))} \right), \tag{28}$$

$$\int_{I_m} |\eta|^2_{H^1(K)} \mathrm{d}t \le C \left(h_K^{2p} |u|^2_{L^2(I_m, H^{p+1}(K))} + \tau_m^{2(q+1)} |u|^2_{H^{q+1}(I_m, H^1(K))} \right), \tag{29}$$

$$h_K^2 \int_{I_m} |\eta|^2_{H^2(K)} \mathrm{d}t \le C \left(h_K^{2p} |u|^2_{L^2(I_m, H^{p+1}(K))} + \tau_m^{2(q+1)} |u|^2_{H^{q+1}(I_m, H^1(K))} \right). \tag{30}$$

The most delicate is the estimation of the expression $\int_{I_m} J_{h,m}(\eta, \eta)\, \mathrm{d}t$. In the same way as in [5] we get

$$\int_{I_m} J_{h,m}\big(\Pi_m u - u, \Pi_m u - u\big)\, \mathrm{d}t \le C\, h^{2p} \big|u\big|^2_{L^2(I_m, H^{p+1}(\Omega))}. \tag{31}$$

In the estimation of $\int_{I_m} J_{h,m}\big(\pi(\Pi_m u) - \Pi_m u,\ \pi(\Pi_m u) - \Pi_m u\big)\, \mathrm{d}t$ we distinguish two cases.

a. Let $\Gamma \in \mathcal{F}_{h,m}^I$ Using the relations $\big[\pi(\Pi_m)u - \Pi_m u\big] = \pi([\Pi_m u]) - [\Pi_m u]$, and $D^{q+1}[\Pi_m u(x,\cdot)] = [D^{q+1}\Pi_m u(x,\cdot)], \big[D^{q+1}u\big] = 0$, and $D^{q+1}(\Pi_m u - u) = \Pi_m(D^{q+1}u) - D^{q+1}u$, (26), Fubini's theorem, the multiplicative trace inequality and the approximation properties of Π_m we obtain

$$\int_{I_m} \Big(\sum_{\Gamma \in \mathcal{F}_{h,m}^I} h(\Gamma)^{-1} \int_\Gamma \big[\pi(\Pi_m u) - \Pi_m u\big]^2 \mathrm{d}S \Big)\, \mathrm{d}t \le C\, \tau_m^{2q+2} \sum_{K \in \mathcal{T}_{h,m}} |u|^2_{H^{q+1}(I_m, H^1(K))}.$$

b. If $\Gamma \in \mathcal{F}_{h,m}^B$, i.e., $\Gamma \subset \partial\Omega \cap \partial K$ for some $K \in \mathcal{T}_{h,m}$, it appears that in the case of general boundary data u_D depending on x and t we get a suboptimal estimate in τ_m of the expression

$$\int_{I_m} \Big(h(\Gamma)^{-1} \int_\Gamma |\pi(\Pi_m u) - \Pi_m u|^2 \, \mathrm{d}S \Big)\, \mathrm{d}t. \tag{32}$$

Therefore, we assume that

$$u_D(x,t) = \sum_{j=0}^{q} \psi_j(x)\, t^j, \tag{33}$$

where $\psi_j \in H^{p+1/2}(\partial\Omega)$ for $j = 0, \ldots, q$, and, thus, $D^{q+1}u|_{\partial\Omega} = D^{q+1}u_D = 0$. This and a similar process as in the case (a), when $\Gamma \in \mathcal{F}_{h,m}^I$, imply that

$$\int_{I_m} \Big(\sum_{\Gamma \in \mathcal{F}_{h,m}^B} h(\Gamma)^{-1} \int_\Gamma \big|\pi(\Pi_m u) - \Pi_m u\big|^2 \, \mathrm{d}S \Big)\, \mathrm{d}t \le C\, \tau_m^{2q+2} \sum_{K \in \mathcal{T}_{h,m}} |u|^2_{H^{q+1}(I_m, H^1(K))}. \tag{34}$$

From the above analysis we get the estimate

$$|J_{h,m}(\eta,\eta)| \le C \sum_{K\in\mathscr{T}_{h,m}} \left(h_K^{2p} |u|^2_{L^2(I_m,H^{p+1}(K))} + \tau_m^{2q+2} |u|^2_{H^{q+1}(I_m,H^1(K))} \right). \tag{35}$$

Finally, using the previous estimates, we obtain the *main result.*

Theorem 2. *Let u be the exact solution of problem (1)–(3) satisfying the regularity condition (24). Let U be the approximate solution to problem (1)–(3) obtained by scheme (12) in the case that the Dirichlet data u_D is defined by (33). Let conditions (16), (17), (22) and (25) be satisfied. Then there exist constants C, $c > 0$ independent of h, τ, m, ε, u, U such that*

$$\begin{aligned} &\|e_m^-\| + \frac{\varepsilon}{2} \sum_{j=1}^{m} \int_{I_m} \|e\|^2_{DG,j} \,\mathrm{d}t \qquad (36)\\ &\le C \exp(ct_m/\varepsilon) \Bigg(\left(h^{2p} |u|^2_{L^2(0,T;H^{p+1}(\Omega))} + \tau^{2q+2} |u|_{H^{q+1}(0,T;H^1(\Omega))} \right) \left(\varepsilon + \frac{1}{\varepsilon} \right) \\ &\qquad + h^{2p} |u|^2_{C([0,T];H^{p+1}(\Omega))} \Bigg), \quad m = 1,\dots,M. \end{aligned}$$

The detailed analysis will be a subject of a paper [9] in preparation. There are several topics for future work:

- Derivation of optimal error estimates in space and time in the case of the SIPG method,
- Numerical realization of the discrete problem and the demonstration of results by numerical experiments,
- Analysis of the effect of numerical integration in space and time integrals.

Acknowledgements This work is a part of the research project MSM 0021620839 financed by the Ministry of Education of the Czech Republic. It was also partly supported by the grant No. 201/08/0012 of the Grant Agency of the Czech Republic.

References

1. Akrivis, G., Makridakis, C.: Galerkin time-stepping methods for nonlinear parabolic equations. ESAIM: Math. Model. Numer. Anal., **38**, 261–289 (2004)
2. Arnold, D. N.: An interior penalty finite element method with discontinuous elements. SIAM J. Numer. Anal., **19**, 742–760 (1982)
3. Arnold, D. N., Brezzi, F., Cockburn, B., Marini, D.: Unified analysis of discontinuous Galerkin methods for elliptic problems. SIAM J. Numer. Anal., **39**, 1749–1779 (2001)
4. Ciarlet, P. G.: The Finite Element Method for Elliptic Problems. North-Holland, Amsterdam (1979)
5. Dolejší, V., Feistauer, M.: Error estimates of the discontinuous Galerkin method for nonlinear nonstationary convection-diffusion problems. Numer. Func. Anal. Optimiz. **26**(25–26), 2709–2733 (2005)

6. Dolejší, V., Feistauer, M., Hozman, J.: Analysis of semi-implicit DGFEM for nonlinear convection-diffusion problems on nonconforming meshes. Comput. Methods Appl. Mech. Eng. **196**, 2813–2827 (2007)
7. Dolejší, V., Feistauer, M., Schwab, C.: A finite volume discontinuous Galerkin scheme for nonlinear convection-diffusion problems. Calcolo, **39**, 1–40 (2002)
8. Feistauer, M.: A remark to the DGFEM for nonlinear convection-diffusion problems applied on nonconforming meshes. In: Numerical Mathematics and Advanced Applications, ENUMATH 2007, K. Kunisch, G. Of, O. Steinbach, Editors. Springer, Heidelberg, 323–330 (2008)
9. Feistauer, M., Kučera, V., Najzar, K., Prokopová, J.: Analysis of space-time discontinuous Galerkin method for nonlinear convection-diffusion problems
10. Eriksson, K., Estep, D., Hansbo, P., Johnson, C.: Computational Differential Equations. Cambridge University Press, Cambridge (1996)
11. Estep, D., Larsson, S.: The discontinuous Galerkin method for semilinear parabolic problems. Math. Model. Numer. Anal., **27**, 35–54 (1993)
12. Schötzau, D., Schwab, C.: An *hp* a priori error analysis of the Discontinuous Galerkin time-stepping method for initial value problems. Calcolo, **37**, 207–232 (2000)
13. Thomée, V.: Galerkin Finite Element Methods for Parabolic Problems. Springer, Berlin (2006)

High Order Finite Volume Schemes for Numerical Solution of Unsteady Flows

Petr Furmánek, Jiří Fürst, and Karel Kozel

Abstract The aim of this contribution is to present two modern high-order finite volume (FVM) schemes for numerical solution of unsteady transonic flows. The first one is derived from the total variation diminishing (TVD) version of the classical MacCormack scheme proposed by Causon. In our case it is used with slight modifications and hence refered to as Modified Causon's scheme. It is no more TVD, but with no loss of accuracy to the TVD version and with a significantly lower demands on computational power and memory (cca 30% less). The second one, based on a similar approach as the WENO family schemes, is the implicit Weighted Least-Square Reconstruction scheme (WLSQR) used in combination with the AUSMPW+ numerical flux. For the turbulence modelling the Kok's TNT turbulence model is employed. Unsteady effects (forced oscillatory motion) are simulated by Arbitrary Lagrangian–Eulerian method (ALE). As the transonic test cases the inviscid and turbulent flow around the NACA 0012 profile and inviscid flow over the ONERA M6 wing were chosen. Comparison of numerical and experimental results for inviscid flow is very good, which is unfortunately not the case of turbulent flow.

1 Introduction

The unsteady effects appear in many physical processes (blood flow, atmospheric boundary layer, turbo-machinery, aeronautics) and have a huge impact on the flow field (sometimes even with fatal consequences, e.g. flutter). Investigation of unsteady flows may be done generally in two ways. Either by experimental measurements or by numerical simulations. One of possible approaches is the Arbitrary

P. Furmánek (✉)
VZLÚ a.s.,
e-mail: petr.furmanek@fs.cvut.cz

J. Fürst and K. Kozel
CTU in Prague,
e-mail: jiri.furst@fs.cvut.cz, karel.kozel@fs.cvut.cz

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_35,

Lagrangian–Eulerian method [3], which combines the Lagrangian and Eulerian way of moving fluid investigation, i.e. both the fluid and its reference frame move. The motion is in our case presented by prescribed oscillations of profile/wing around the reference point/axis. The chosen schemes were tested with a very good results for a number of steady test cases before used for numerical solution of unsteady flow.

2 Mathematical Model

Viscous compressible flow in general 2D case is described by the following set of Navier–Stokes equations (conservative vector form):

$$W_t + F_x + G_y = 0, \tag{1}$$

where

$$F = F^c - \frac{1}{Re}F^v,\ G = G^c - \frac{1}{Re}G^v,\ \ W = (\rho, \rho u, \rho v, e)^T, \tag{2}$$
$$F^c = (\rho u, \rho u^2 + p, \rho uv, (e+p)u)^T,\ \ F^v = (0, \tau_{xx}, \tau_{xy}, u\tau_{xx} + v\tau_{xy} + \frac{\kappa}{Pr}\lambda u_x)^T,$$
$$G^c = (\rho v, \rho uv, \rho v^2 + p, (e+p)v)^T,\ \ G^v = (0, \tau_{xy}, \tau_{yy}, u\tau_{xy} + v\tau_{yy} + \frac{\kappa}{Pr}\lambda v_y)^T,$$
$$p = (\kappa - 1)\big[e - \frac{1}{2}\rho(u^2 + v^2)\big],\ \kappa = \frac{c_p}{c_v} \qquad (\textit{equation of state})$$

with W being the vector of conservative variables, F^c, G^c – convective fluxes, F^v, G^v – viscous fluxes, ρ – density; (u, v) – velocity vector; p – pressure; e – total energy per unit volume, τ – tensor of viscous stresses, Re – Reynolds number, Pr – Prandtl number, λ – heat transfer coefficient. Subscripts $_t$, $_x$, $_y$ signify time and spatial partial derivatives. In the case of inviscid flow the viscous fluxes are neglected and the system of the Euler equations is obtained.

3 Numerical Methods

Unsteady flows were numerically simulated with the use of the Arbitrary Lagrangian-Eulerian method. System (1) is solved by the Finite Volume Method. The time-dependent computational domain $\Omega(t)$ is divided into a set of non-overlapping computational cells $D_i(t)$. Using the *geometric conservation law* [8] the following identity is obtained

$$\frac{\mathrm{d}}{\mathrm{d}t}\iint_{D_i(t)} W(t)\mathrm{d}x\,\mathrm{d}y = \iint_{D_i(t)} W(t)\mathrm{d}x\,\mathrm{d}y + \oint_{\partial D_i(t)} W(t)\cdot\dot{\mathbf{x}}\cdot\mathbf{n_0}\mathrm{d}S \tag{3}$$

where $\dot{\mathbf{x}} = (w_1, w_2)$ is velocity of a point on the boundary $\partial D_i(t)$.Then in each time-dependent cell $D_i(t)$ the following relation should be fulfilled (with the use of Gauss–Ostrogradsky theorem and the mean value theorem):

$$\begin{aligned}
&\iint_{D_i(t)} W_t(t)\mathrm{d}x\,\mathrm{d}y + \iint_{D_i(t)} (F_x(W(t)) + G_y(W(t)))\mathrm{d}x\,\mathrm{d}y = \\
&\frac{\mathrm{d}}{\mathrm{d}t}(|D_i(t)|W_i(t)) + \oint_{\partial D_i(t)} \left[(F(W(t)), G(W(t))) - W(t)\cdot\dot{\mathbf{x}}\right]\cdot\mathbf{n_0}\mathrm{d}S = \\
&\frac{\mathrm{d}}{\mathrm{d}t}(|D_i(t)|W_i(t)) + \oint_{\partial D_i(t)} \left[F^*(W(t)), G^*(W(t))\right]\cdot\mathbf{n_0}\mathrm{d}S = 0
\end{aligned} \tag{4}$$

where

$$\begin{aligned}
F^*(W(t)) &= F^c(W(t)) - w_1 W(t) - \frac{1}{Re}F^v(W(t)) \\
G^*(W(t)) &= G^c(W(t)) - w_2 W(t) - \frac{1}{Re}G^v(W(t))
\end{aligned} \tag{5}$$

$\mathbf{n_0}$ is outer normal vector to the interface of cell $D_i(t)$ and $|D_i(t)|$ is its volume. $(F, G)\cdot\mathbf{n_0}$ denotes product of a matrix $(F, G) \in \mathbb{R}^{4,2}$ and vector $\mathbf{n_0} \in \mathbb{R}^{1,2}$.

4 Numerical Schemes

4.1 Modified Causon's Scheme

Numerical solution of (4) was obtained by two different FVM schemes. The first was the so called *Modified Causon's scheme* [4, 6]. It is based on classical explicit MacCormack predictor-corrector scheme in TVD form, which is able to deliver very good results. However, it also entails disadvantageous demands for both computational memory and power. Therefore a simplification saving approximately 30% of computational time was proposed by Causon [1] by introducing a special type of artificial dissipation (AD). This new scheme was still TVD, but the influence of AD turned out to be too strong. The authors on the other hand proposed another modification based on Causon's scheme (refered to as the Modified Causon's scheme), which is no more TVD, but keeps the advantages of the Causon's scheme while clearing out its drawbacks in the same time.

4.2 Weighted Least-Square Reconstruction Scheme (WLSQR)

When solving (4) with the WLSQR scheme [4,5], the real inviscid fluxes in the surface integrals are approximated by numerical ones (in our case by the AUSMPW+ flux [7]). The high order accuracy in time is obtained in a standard way by using

the interpolated values at the cell faces. The interpolation is obtained by using the weighted least-square approach, which usually gives better convergence to steady state than the methods with Barth's limiter. Advancing in time is realised by the non-linear implicit dual-time backward Euler method. Resulting sparse system of linear equations is solved by GMRES with ILU(0) preconditioning. Dimension of the Krylov subspace is chosen between 10–40 and maximum number of iteration is set to 10–50. If the steady solution is not found in prescribed number of iterations the computation proceeds in the next time step.

4.3 Modification of the Computational Mesh

Because the ALE method uses moving meshes, also the algorithm for mesh modification has to be prescribed. The actual position of mesh vertices during the unsteady computation was given by the following prescription for each mesh vertex $\mathbf{x}(t)$

$$\mathbf{x}(t) = \mathbb{Q}\big[\phi(t, ||\mathbf{x}(0) - \mathbf{x}_{ref}||)\big](\mathbf{x}(0) - \mathbf{x}_{ref}) + \mathbf{x}_{ref} \tag{6}$$

$$\mathbb{Q}(\phi) = \begin{pmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{pmatrix}, \qquad \phi(\mathrm{t}, r) = \begin{cases} -\alpha_1(\mathrm{t}) & \text{for } r < r_1, \\ -\alpha_1(\mathrm{t})\, f_D(r) & \text{for } r_1 \le r < r_2, \\ 0 & \text{for } r_2 < r. \end{cases}$$

$$f_D(r) = \left[2\left(\frac{r - r_1}{r_2 - r_1}\right)^3 - 3\left(\frac{r - r_1}{r_2 - r_1}\right)^2 + 1 \right]$$

It means that the circle with center in $\mathbf{x}_{ref}$ and radius r_1 is rotating according to the prescribed change of pitching angle as a solid body. The outer area of the second circle with the radius $r_2 > r_1$ is motion-less and in the annulus between the two circles the motion of the mesh is damped with the use of damping function $f_D()$.

5 Numerical Results

5.1 2D Unsteady Transonic Flow

Considered test case is transonic flow over an oscillating NACA 0012 profile for which the experimental data are available in [2]. It is characterised by the inlet Mach number $M_\infty = 0.755$. The oscillatory motion of the profile around the reference point $x_{ref} = [0.25, 0.00]$ is given by the pitching angle $\alpha_1(t) = 0.016° + 2.51° \sin(\omega t)$. The angular velocity is defined as $\omega = \frac{2kU_\infty}{c}$, where U_∞ is the free-stream velocity (since the non-dimensional form of (1) is considered and

angle of attack $\alpha = 0°$ then $U_\infty = M_\infty$), $c = 1$ is the chord length and the reduced frequency $k = 0.0814$. The unsteady state development was observed on the behaviour of the lift coefficient (c_l) given as $c_l = \frac{\oint_{\Gamma_{prof}} p \, \mathrm{d}x}{\frac{1}{2} U_\infty^2 \rho_\infty}$, where $\rho_\infty = 1$ and Γ_{prof} is the curve defining the profile. The used computational schemes and meshes were

- Modified Causon's scheme – structured C-mesh with 15,096 elements (124 cells around profile),
- WLSQR scheme with AUSMPW+ flux – unstructured mesh with 6,720 quadrilateral cells (120 cells around profile). For the turbulent flow simulation the Kok's TNT turbulence model was used.

As can be seen from Figs. 1 and 2 the numerical results obtained by both schemes in the case of inviscid flow are very good. For the c_l comparison the results correspond qualitatively, but experimental data show a bit higher c_l values (Fig. 1). Considering symmetry of the problem, also the behaviour of the c_l should be symmetric with the center of symmetry in the point $[0, 0]$. The experimental data however do not have this characteristic and therefore the suspicion of their systematical error comes in mind. Important characteristics, e.g. the position and intensity

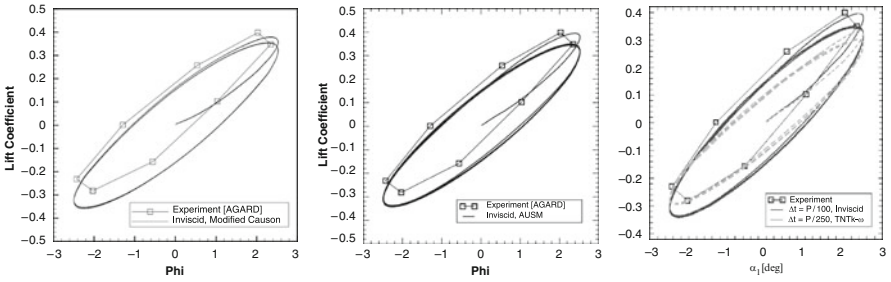


Fig. 1 NACA 0012, lift coefficient behaviour, comparison of numerical (*squares*) and experimental (*lines*) results. To the *left* is Modified Causon's scheme, inviscid computation, the *centre* is WLSQR scheme, AUSMPW+, inviscid computation, and to the *right* is WLSQR scheme, AUSMPW+, comparison of inviscid (*line*) and turbulent(*dashed*) computation

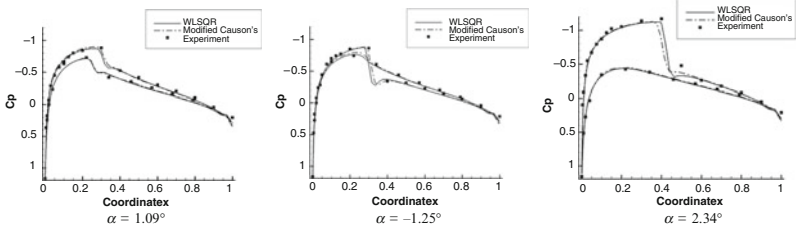


Fig. 2 c_p coefficient during the 5th period of forced oscillatory motion, inviscid flow, comparison of experimental (*dots*) and numerical (*lines*) results (Modified Causon's scheme, WLSQR scheme with AUSMPW+flux)

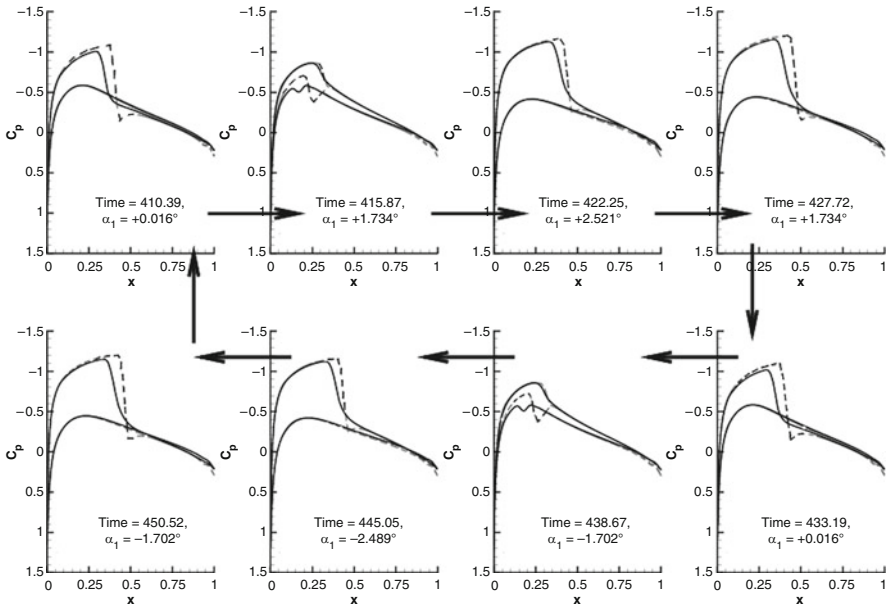


Fig. 3 c_p coefficient behaviour during the 5th period of forced oscillatory motion, WLSQR scheme, comparison of inviscid (*dashed*) and turbulent (*line*) model (Kok's TNT)

of the shock wave (minimal and maximal reached value of c_p), are however in a very good correspondence, which is unfortunately not the case of the turbulent computation, where both the c_p and c_l coefficient differ significantly (Fig. 3). It is therefore necessary to use another turbulence model (EARSM) or large eddy simulation (LES).

5.2 3D Unsteady Inviscid Transonic Flow

The initial conditions for 3D unsteady inviscid transonic flow were taken from the standard test case mentioned in [9]. The forced oscillatory motion of the wing around the elastic axis parallel with the axis z and going through the reference point $x_{ref} = [\frac{1}{3}; 0.00; 0.00]$ was given by the same relation for pitching angle as in 2D. The inlet Mach number was considered $M_\infty = 0.8395$, initial deviation $\alpha_0 = 3.06°$, amplitude $\alpha_1 = 1.5°$ and frequency $f = 10\,\text{Hz}$. The structured computational mesh had 467,313 elements and its deformation during the ALE computation was given by the 3D extension of (6). Computation was carried out using the Modified Causon's scheme.

The Modified Causon's scheme has proved itself well – the results (Fig. 4) show that the fully periodic state has been achieved at least during the 5th period of the oscillatory motion. Pressure coefficient decreases with increasing angle of attack

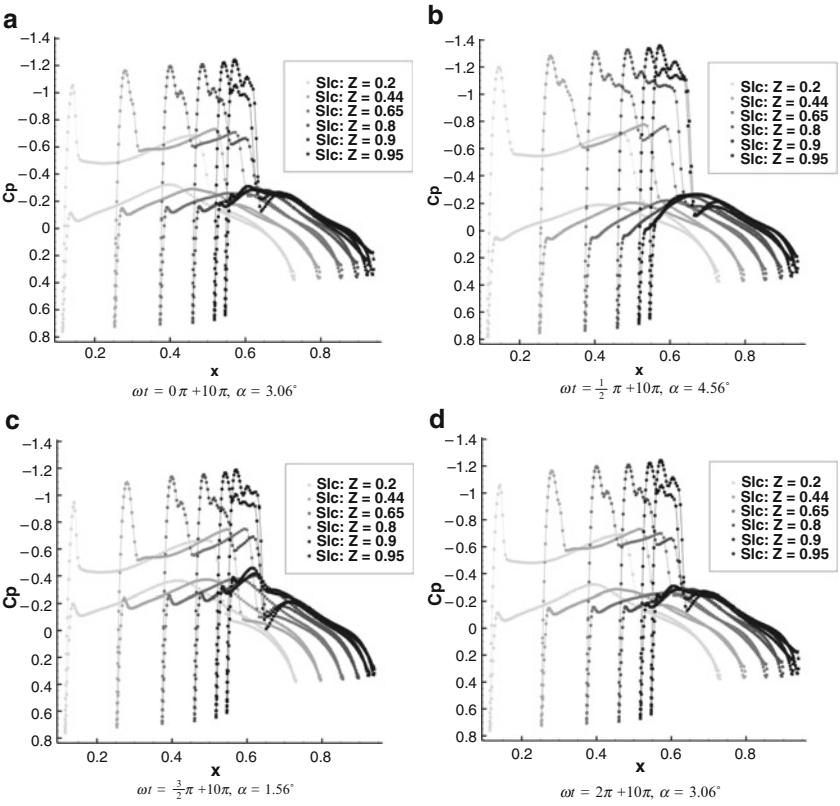


Fig. 4 c_p coefficient behaviour during the 5th period of forced oscillatory motion

(and vice versa) and the scheme does not produce spurious oscillations. Comparison with the experimental data is unfortunately not yet available, but work is in progress at the present time on implementation of another wing geometry used in the experiments with oscillating wing at the Aeronautical Research and Test Institute in Prague (VZLÚ a. s.).

6 Conclusion

Proposed FVM schemes for numerical solution of unsteady 2D and 3D transonic inviscid flows show very good accuracy. They were able to capture important flow characteristics as the position and intensity of shockwaves and have proved themselves as a reliable numerical simulation of investigated cases. Chosen turbulent model however does not suit the simulated unsteady flow regime and hence another one has to be employed (e.g. the EARSM model). Both schemes would need some further improvements (implicit form in the case of Modified Causon's scheme, matrix-free GMRES in the case of WLSQR scheme).

Acknowledgements This work was partially supported by the Research Plans VZ MSM 6840770 010, VZ MSM 0001066902 and grants GACR 201/08/0012, GACR 101/07/1508.

References

1. D. M. Causon: High resolution finite volume schemes and computational aerodynamics. In: J. Ballmann and R. Jeltsch, editors, Nonlinear Hyperbolic Equations – Theory, Computation Methods and Applications, volume 24 of Notes on Numerical Fluid Mechanics, pages 63–74, Braunschweig, Vieweg, March 1989
2. Compendium of unsteady aerodynamic measurements. AGARD Advisory Report No. 702, 1982
3. J. Donea: An arbitrary Lagrangian–Eulerian finite element method for transient fluid- structur interactions. Comput. Methods Appl. Mech. Eng., (1982), 33:689–723
4. J. Fürst: Numerical Solution of Transonic Flow Using Modern Schemes of Finite volume Method and Finite Differences, Dissertation thesis (in Czech), ČVUT, Praha, 2001
5. J. Fürst: A weighted least square scheme for compressible flows. Flow, Turbulence and Combustion (2005)
6. J. Fürst, K. Kozel: Application of second order TVD and ENO schemes in internal aerodynamics. J. Sci. Comput., (2002), 17(1–4): 263–272. ISSN 0885-7474
7. Kyu Hong Kim, Chongam Kim, Oh-Hyun Rho: Methods for the accurate computations of hypersonic fows I. AUSMPW+ scheme. J. Comput. Phys., (2001), 174:38–80
8. M. Lesoinne, C. Farhat: Geometric conservation laws for flow problems with moving boundaries and deformable meshes, and their impact on aeroelastic computations. Comp. Methods Appl. Mech. Eng., (1996), 134:71–90
9. V. Schmitt, F. Charpin: Pressure Distributions on the ONERA-M6-Wing at Transonic Mach Numbers. Experimental Data Base for Computer Program Assessment. Report of the Fluid Dynamics Panel Working Group 04, AGARD AR 138, May 1979

Multigrid Finite Element Method on Semi-Structured Grids for the Poroelasticity Problem

F.J. Gaspar, F.J. Lisbona, and C. Rodrigo

Abstract An efficient finite element multigrid method on semi-structured triangular grids, based on box-relaxation, is proposed for the poroelasticity problem. A stabilized finite element scheme for these equations, based on the perturbation of the flow equation is considered. Numerical results confirm the good performance of Vanka smoothers for this saddle point type problem.

1 Introduction

Multigrid methods [4, 8, 11] are one of the most powerful techniques for solving the corresponding large systems of equations arising from the discretization of partial differential equations. Geometric multigrid methods are characterized by employing a hierarchy of grids. For an irregular domain, it is very common to consider an unstructured mesh as coarsest grid in order to fit well its geometry, and to apply regular refinement to its elements. We are interested in the use of semi-structured triangular grids, where a nested hierarchy of grids is obtained by dividing each triangle into four congruent ones, connecting the midpoints of their edges. These grids provide a suitable framework for the implementation of a geometric multigrid algorithm, permitting the use of stencil-based data structures, see [2, 6].

The choice of a suitable smoother is an important feature for the design of an efficient geometric multigrid method, and even it requires special attention when one works with systems of PDEs because the smoother should smooth the error for all unknowns. Moreover, for saddle point problems (they have a zero or almost zero block in the matrix for one of the unknowns) numerical experiments show that smoothing factors of standard collective point-wise relaxations are not satisfactory. The poroelasticity model is an example of such problems, and its resolution by multigrid on semi-structured grids is the aim here. An overview of multigrid

C. Rodrigo (✉), F.J. Lisbona, and F.J. Gaspar
University of Zaragoza, Pedro Cerbuna 12, 50009, Zaragoza, Spain
e-mail: fjgaspar@unizar.es, lisbona@unizar.es, carmenr@unizar.es

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_36, © Springer-Verlag Berlin Heidelberg 2010

methods for discretizations on rectangular grids of saddle point problems is presented in [10], where coupled or box-relaxation appears as one of the most suitable smoothers for this kind of problems. It consists of decomposing the mesh into small subdomains and treating them separately, that is, all (or a part of) the equations corresponding to the points in each subdomain are solved simultaneously as a system. This class of smoothers was introduced by Vanka [13] to solve the finite difference discretization on rectangular grids of the Navier–Stokes equations. Since then, much literature can be found about the application of this type of smoothers, mainly in the field of Computational Fluid Dynamics (CFD) [9, 12]. There are less papers concerning to the performance of this relaxation in the context of Computational Solid Mechanics (CSM), see for example [14]. However, for discretizations of the poroelasticity problem on rectangular grids, it has been proved to obtain very good results with these smoothers. For instance, in [5] a box-relaxation is performed for a discretization of the problem on staggered grids. Hence, it seems a good idea to extend box-relaxation to triangular grids.

In Sect. 2, the formulation of the poroelasticity problem, as well as its stabilized finite element discretization, will be introduced. Section 3 is devoted to present the proposed multigrid algorithm, based on Vanka-type smoothers on triangular grids. Finally, in Sect. 4 some numerical experiments will be presented in order to illustrate the obtained results, in which some troubles with regard to the coarse-grid correction of the stabilized problem are shown.

2 Poroelasticity Problem

Poroelasticity theory addresses the time dependent coupling between the deformation of a porous material and the fluid flow inside. The general statement of this problem was given by Biot in [3]. We assume the porous medium to be linearly elastic, homogeneous and isotropic, and the porous matrix is supposed to be saturated by an incompressible fluid. The state of this continuous medium is characterized by the knowledge of elastic displacements $\mathbf{u}$, and fluid pressure p at each point, and in terms of these unknowns the governing equations of the consolidation problem are given by

$$-\mu\Delta\mathbf{u} - (\lambda+\mu)\nabla(\nabla\cdot\mathbf{u}) + \nabla p = \mathbf{g}(\mathbf{x},t), \tag{1}$$

$$\frac{\partial}{\partial t}(\nabla\cdot\mathbf{u}) - \frac{\kappa}{\eta}\Delta p = f(\mathbf{x},t), \quad \mathbf{x}\in\Omega, \ \ 0<t\leq T, \tag{2}$$

where Ω is an open bounded region of $\mathbb{R}^n$, $n \leq 3$, λ and μ are the Lamé coefficients, κ is the permeability of the porous medium, and η is the viscosity of the fluid. The source terms $\mathbf{g}(\mathbf{x},t)$ and $f(\mathbf{x},t)$ represent a density of applied body forces and a forced fluid extraction or injection process, respectively. We consider the following boundary and initial conditions

$$\begin{aligned} & p = 0, \quad \boldsymbol{\sigma}' \mathbf{n} = \mathbf{t}, \quad \text{on } \Gamma_t, \\ & \mathbf{u} = \mathbf{0}, \quad \frac{\kappa}{\eta}(\nabla p) \cdot \mathbf{n} = 0, \quad \text{on } \Gamma_c, \\ & \nabla \cdot \mathbf{u}\,(\mathbf{x}, 0) = 0, \quad \mathbf{x} \in \Omega, \end{aligned} \tag{3}$$

where $\boldsymbol{\sigma}'$ is the effective stress tensor for the porous medium, $\mathbf{n}$ is the unit outward normal to the boundary and $\Gamma_t \cup \Gamma_c = \Gamma$, with Γ_t and Γ_c disjoint subsets of Γ with non null measure.

Considering the following function spaces $\mathscr{Q} = \{q \in H^1(\Omega) \,|\, q = 0 \text{ on } \Gamma_t\}$, and $\mathscr{U} = \{\mathbf{u} \in (H^1(\Omega))^n \,|\, \mathbf{u} = \mathbf{0} \text{ on } \Gamma_c\}$, and introducing the corresponding bilinear forms

$$a(\mathbf{u}, \mathbf{v}) = 2\mu \sum_{i,j=1}^{n} (\epsilon_{ij}(\mathbf{u}), \epsilon_{ij}(\mathbf{v})) + \lambda(\nabla \cdot \mathbf{u}, \nabla \cdot \mathbf{v}), \quad b(p, q) = \frac{\kappa}{\eta} \sum_{i=1}^{n} \left(\frac{\partial p}{\partial x_i}, \frac{\partial q}{\partial x_i} \right),$$

the variational formulation of problem (1) and (2) with boundary and initial conditions (3) reads:

For each $t \in (0, T]$, find $(\mathbf{u}(t), p(t)) \in \mathscr{U} \times \mathscr{Q}$ such that

$$\begin{aligned} & a(\mathbf{u}(t), \mathbf{v}) + (\nabla p(t), \mathbf{v}) = (\mathbf{g}, \mathbf{v}) + h(\mathbf{v}), \quad \forall\, \mathbf{v} \in \mathscr{U}, \\ & \left(\frac{\partial}{\partial t}(\nabla \cdot \mathbf{u}(t)), q\right) + b(p(t), q) = (f, q), \quad \forall q \in \mathscr{Q}, \end{aligned}$$

with the initial condition $(\nabla \cdot \mathbf{u}(0), q) = 0, \forall\, q \in L^2(\Omega)$, and where $h(\mathbf{v}) = \int_{\Gamma_t} \mathbf{t} \cdot \mathbf{v}\, \mathrm{d}\Gamma$. Let $\mathscr{T}_h$ be a triangulation of Ω satisfying the usual admissibility assumption. Let $S_h^k \subset H^1(\Omega)$ be the spaces of C^0 piecewise polynomial finite element interpolations of degree k. In the two-dimensional case, we can define finite element approximations for $\mathscr{U}$ and $\mathscr{Q}$ as $\mathscr{U}_h^k = \mathscr{U} \bigcap (S_h^k \times S_h^k)$ and $\mathscr{Q}_h^{k'} = \mathscr{Q} \bigcap S_h^{k'}$, respectively. Using an implicit time discretization, the following discrete formulation of the considered problem is obtained:

For $m \geq 1$, find $(\mathbf{u}_h^m, p_h^m) \in \mathscr{U}_h^k \times \mathscr{Q}_h^{k'}$ such that

$$a(\mathbf{u}_h^m, \mathbf{v}_h) + (\nabla p_h^m, \mathbf{v}_h) = (\mathbf{g}^m, \mathbf{v}_h) + h(\mathbf{v}_h), \quad \forall \mathbf{v}_h \in \mathscr{U}_h^k, \tag{4}$$

$$(\nabla \cdot \mathbf{u}_h^m, q_h) + \tau b(p_h^m, q_h) = (\nabla \cdot \mathbf{u}_h^{m-1}, q_h) + \tau(f^m, q_h) \quad \forall q_h \in \mathscr{Q}_h^{k'}, \tag{5}$$

where τ is the time discretization parameter.

Here, we use linear finite elements to approximate the problem, however it is well-known that choosing the same polynomial space for approximation of displacements and pressure, i.e. $k = k'$ in (4) and (5), strong nonphysical oscillations can appear in the approximation for the pressure field, when the space discretization parameter h is

not sufficiently small. To overcome this oscillating behavior, in [1] a stabilized finite element scheme, based on the perturbation of the flow equation, was proposed. This technique allows to use continuous piecewise linear approximation spaces for both displacements and pressure, providing solutions without oscillations independently of the chosen discretization parameters. This scheme is based on the perturbation of the flow equation with a term which arises from the discretization of the time derivative of the Laplacian of the pressure multiplied by a coefficient $\beta = h^2/4(\lambda + 2\mu)$. Thus, the corresponding discrete variational problem reads:
For $m \geq 1$, find $(\mathbf{u}_h^m, p_h^m) \in \mathscr{U}_h^k \times \mathscr{Q}_h^k$ such that

$$a(\mathbf{u}_h^m, \mathbf{v}_h) + (\nabla p_h^m, \mathbf{v}_h) = (\mathbf{g}^m, \mathbf{v}_h) + h(\mathbf{v}_h), \quad \forall\, \mathbf{v}_h \in \mathscr{U}_h^k,$$

$$(\nabla \cdot \mathbf{u}_h^m, q_h) + (\tau + \beta')b(p_h^m, q_h) = \tau(f^m, q_h) + (\nabla \cdot \mathbf{u}_h^{m-1}, q_h) + \beta' b(p_h^{m-1}, q_h), \quad \forall q_h \in \mathscr{Q}_h^k,$$

where $\beta' = \beta \dfrac{\eta}{\kappa}$.

3 Multigrid Based on Vanka-type smoothers

Geometric multigrid methods are strongly dependent on the choice of adequate components to the considered problem. These components have to be chosen so that they efficiently interplay with each other in order to obtain a good connection between the relaxation and the coarse-grid correction. In this paper, linear interpolation is chosen as the prolongation, and its adjoint as the restriction operator. The discrete operator on each mesh in the hierarchy results from the direct discretization of the partial differential equation on the corresponding grid. As it has been previously commented, box-relaxation is a suitable smoother to deal with poroelasticity problem. There are many variants of box-type smoothers, they can differ in the choice of the subdomains which are solved simultaneously, and in the way in which the local systems to be solved are built. Firstly, we consider a point-wise box Gauss–Seidel iterative algorithm, which consists of simultaneously updating all unknowns appearing in the discrete divergence operator in the second equation of the system. This means that 12 unknowns corresponding to displacements and one pressure unknown (see left Fig. 1) are relaxed simultaneously and therefore, a 13×13 system has to be solved for each point. Another version of box-relaxation is the line-wise variant. In triangular grids three different line box smoothers can be defined, each one associated with each vertex of the triangle, and they consist of looping over each line parallel to the opposite edge to the corresponding vertex and relaxing simultaneously all the unknowns appearing in the divergence operator associated with each pressure point of the line, that is, all these unknowns marked in right Fig. 1. These smoothers are much more expensive than their point-wise counterpart, but in some situations in which point-wise relaxation does not give satisfactory results, they are

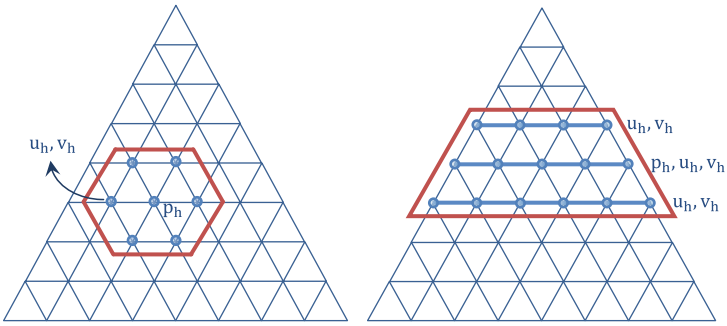


Fig. 1 Unknowns simultaneously updated in point-wise and line-wise box Gauss–Seidel

a very good option. For example, point-wise box-smoothers can be less robust on anisotropic meshes, being the line-wise box-smoothers preferred in this situation.

4 Numerical Experiments

Following the concept developed in [7] for the efficient design of geometric multigrid algorithms on semi-structured triangular grids, we are interested in using a block-wise multigrid algorithm, where each of the triangles of the coarsest grid is treated as a different block with regard to the smoothing process. That is, different smoothers will be chosen for triangles with different geometries. In particular, point-wise box smoothers will be used for equilateral or almost equilateral triangles, and line-wise box smoothers will be considered when the grid becomes anisotropic, that is, for triangles with some small angle.

Depending on the value of the space discretization parameter h, the artificial stabilization term has more or less influence. If this parameter is very small the artificial term is negligible, whereas if h is big enough this term becomes more dominating and is well-known that the multigrid convergence for this problem slows down, what is due mainly to poor coarse-grid correction to certain error components. In order to see this behavior, results for both cases will be presented.

We begin considering the case in which h is sufficiently small to neglect the effect of the artificial term. Some results for two triangular domains, an equilateral and an isosceles with common angle 85°, which are representative for the application of point-wise and line-wise box relaxation, respectively, are presented. In Table 1, for the considered triangular domains and for a wide range of values of parameter $k = \tau\kappa/\eta$, the asymptotic convergence factors, experimentally computed by taking a random initial guess and a zero right-hand side, are shown for different numbers of pre- and post-smoothing steps. It is observed that both smoothers provide a convergence independent on k, even for small values of this parameter, and the obtained

Table 1 Asymptotic computed convergence factors for an equilateral triangle with a point-wise box smoother and for an isosceles triangle with common angle 85° with a line-wise box smoother

	Equilateral			Isosceles (85°)		
k	$(1,0)$	$(1,1)$	$(2,1)$	$(1,0)$	$(1,1)$	$(2,1)$
10^{-4}	0.325	0.121	0.069	0.332	0.133	0.087
10^{-6}	0.325	0.121	0.069	0.332	0.133	0.087
10^{-8}	0.325	0.121	0.069	0.332	0.133	0.087
10^{-10}	0.325	0.121	0.069	0.332	0.133	0.087
10^{-12}	0.325	0.121	0.069	0.332	0.133	0.087
10^{-14}	0.325	0.121	0.069	0.332	0.133	0.087

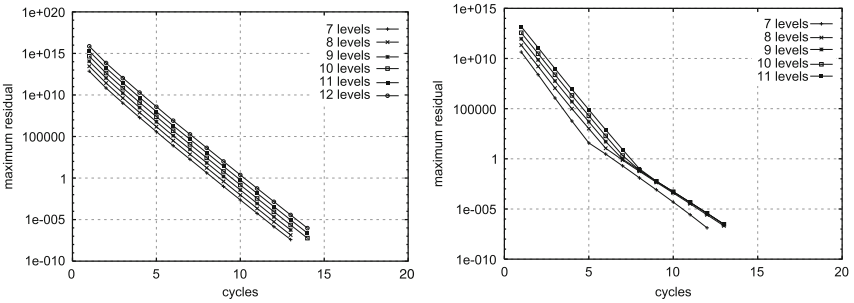


Fig. 2 Robustness of point-wise and line-wise box-smoothers, respectively

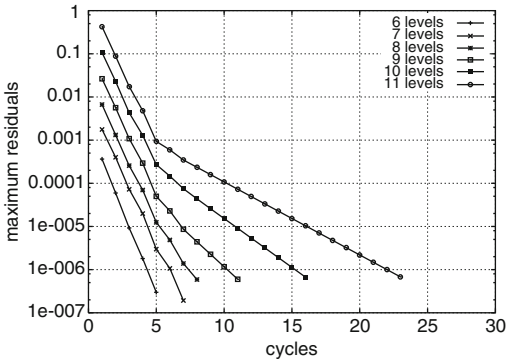
factors are very satisfactory. In order to see the robustness of the method with respect to the space discretization parameter, in Fig. 2 the history of the convergence of the method, with an F(2,1), $k = 10^{-8}$, and with stopping criterion as the maximum residual over all unknowns to be less than 10^{-6}, is depicted for both smoothers, and the independency on the number of refinement levels is shown.

Next, we deal with the case of h sufficiently large. The dominance of the artificial stabilization term causes some convergence problems of the proposed multigrid algorithm. In Fig. 3a, although the robustness with regard to the parameter k is observed, a significative deterioration of the experimentally computed convergence factors obtained with the point-wise box smoother for an equilateral triangle is seen. Besides, in Fig. 3b, the history of the convergence for this algorithm, with an F(2,1) and with stopping criterion as the maximum residual over all unknowns to be less than 10^{-6}, shows how the number of iterations, necessary to reach the desired value for the residual, grows up as the number of refinement levels increases. The same behavior has been observed for the line version of the smoother. As commented before, this poor performance is due to the coarse-grid correction, and some techniques to overcome these troubles will be investigated in a forthcoming publication.

Finally, to see how efficiently the block-wise multigrid works, we solve a poroelasticity problem on the rectangular domain depicted in Fig. 4, in which the considered space discretization parameter h is small enough to discard problems

	Equilateral		
k	$(1,0)$	$(1,1)$	$(2,1)$
10^{-4}	0.838	0.734	0.677
10^{-6}	0.838	0.734	0.677
10^{-8}	0.838	0.734	0.677
10^{-10}	0.838	0.734	0.677
10^{-12}	0.838	0.734	0.677
10^{-14}	0.838	0.734	0.677

a)



b)

Fig. 3 (**a**) Asymptotic experimentally computed convergence factors with point-wise box smoother. (**b**) History of the convergence for different refinement levels with $k = 10^{-8}$

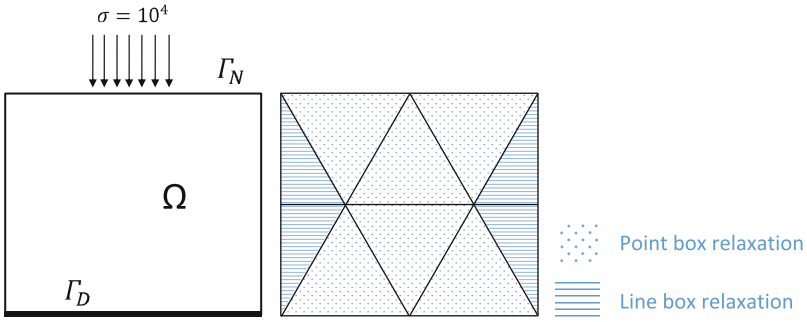


Fig. 4 Rectangular computational domain, and smoothers applied on each coarse triangle

with the stabilization term on the coarse-grid correction. The body is assumed rigid at the bottom and in the central part of the upper edge a uniform load of density 10^4 is applied.

We apply the proposed multigrid method, considering different smoothers for different triangles of the coarsest mesh, composed of 10 triangles. In particular, in Fig. 4 the triangles in which point-wise or line-wise box smoothing is considered are distinguished. The obtained results for the convergence of the method, considering an F(2,1) and for a value of parameter $k = 10^{-8}$, are shown in Table 2, where the number of cycles necessary to reduce the initial residual with a factor of 10^{-8} are shown for different refinement levels, together with the number of elements and the number of unknowns of the problem. An independent convergence of the stepsize of the mesh is observed, and the efficiency of the proposed algorithm is also shown, since the residual is reduced in only 10 iterations.

Table 2 Number of cycles necessary to reduce the initial residual with a factor of 10^{-8}

No. of levels	No. of elements	No. of unknowns	No. of cycles
6	10,240	15,747	10
7	40,960	62,211	10
8	163,840	247,299	10
9	655,360	986,115	10
10	2,621,440	3,938,307	10
11	10,485,760	15,740,931	10

Acknowledgements This research has been partially supported by FEDER/MCYT Projects MTM2007-63204 and the DGA (Grupo consolidado PDIE).

References

1. Aguilar, G., Gaspar, F., Lisbona, F., Rodrigo, C.: Numerical stabilization of Biot's consolidation model by a perturbation on the flow equation. Int. J. Numer. Methods Eng. **75**, 1282–1300 (2008)
2. Bergen, B., Gradl, T., Hülsemann, F., Rüde, U.: A massively parallel multigrid method for finite elements. Comput. Sci. Eng. **8**, 56–62 (2006)
3. Biot, M.: General theory of three dimensional consolidation. J. Appl. Phys. **12**, 155–169 (1941)
4. Brandt, A.: Multi-level adaptive solutions to boundary-value problems. Math. Comput. **31**, 333–390 (1977)
5. Gaspar, F.J., Lisbona, F.J., Oosterlee, C.W., Wienands, R.: A systematic comparison of coupled and distributive smoothing in multigrid for the poroelasticity system. Numer. Linear Algebra Appl. **11**, 93–113 (2004)
6. Gaspar, F.J., Gracia, J.L., Lisbona, F.J., Rodrigo, C.: Efficient geometric multigrid implementation for triangular grids. J. Comput. Appl. Math. **234**, 1027–1035 (2010)
7. Gaspar, F.J., Gracia, J.L., Lisbona, F.J., Rodrigo, C.: Multigrid finite element methods on semi-structured triangular grids for planar elasticity. NLAA, **17**, 473–493 (2010)
8. Hackbusch, W.: Multi-grid methods and applications. Springer, Berlin, 1985
9. John, V., Tobiska, L.: Numerical performance of smoothers in coupled multigrid methods for the parallel solution of the incompressible Navier–Stokes equations. Int. J. Numer. Methods Fluids **33**, 453–473 (2000)
10. Oosterlee, C.W., Gaspar F.J.: Multigrid relaxation methods for systems of saddle point type. Appl. Numer. Math. **58**, 1933–1950 (2008)
11. Trottenberg, U., Oosterlee, C.W., Schüller, A.: Multigrid. Academic Press, New York, 2001
12. Turek, S.: Efficient solvers for incompressible flow problems: an algorithmic and computational approach. Springer, Berlin, 1999
13. Vanka, S.P.: Block-implicit multigrid solution of Navier-Stokes equations in primitive variables. J. Comput. Phys. **65**, 138–158 (1986)
14. Wobker, H., Turek, S.: Numerical studies of Vanka-type smoothers in computational solid mechanics. Adv. Appl. Math. Mech. **1**, 29–55 (2009)

A Posteriori Error Bounds for Discontinuous Galerkin Methods for Quasilinear Parabolic Problems

Emmanuil H. Georgoulis and Omar Lakkis

Abstract We derive a posteriori error bounds for a quasilinear parabolic problem, which is approximated by the hp-version interior penalty discontinuous Galerkin method (IPDG). The error is measured in the energy norm. The theory is developed for the semidiscrete case for simplicity, allowing to focus on the challenges of a posteriori error control of IPDG space-discretizations of strictly monotone quasilinear parabolic problems. The a posteriori bounds are derived using the elliptic reconstruction framework, utilizing available a posteriori error bounds for the corresponding steady-state elliptic problem.

1 Introduction

Discontinuous Galerkin (DG) methods [1, 2, 17], have enjoyed substantial development in recent years. For parabolic problems DG methods are interesting due to their good local conservation properties as well as due to their block-diagonal mass matrices.

This work is concerned with the derivation of a posteriori error bounds for the space-discrete interior penalty discontinuous Galerkin method (IPDG) for quasilinear parabolic problems with strictly monotone non-linearities of Lipschitz growth.

A posteriori error bounds for h-version DG methods are derived in [3, 10, 12, 13] and for DG-in-space parabolic problems in [5–8, 16, 18]. The contribution of this work is twofold:

- The derivation of a posteriori energy-norm error bounds for IPDG methods for quasilinear parabolic problems, and

E.H. Georgoulis (✉)
Department of Mathematics, University of Leicester, University Road, Leicester LE1 7RH, UK
e-mail: Emmanuil.Georgoulis@mcs.le.ac.uk

O. Lakkis
Department of Mathematics, University of Sussex, Falmer, East Sussex BN1 9RF, UK
e-mail: O.Lakkis@sussex.ac.uk

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_37,

- The resulting a posteriori bounds are are explicit with respect to the local elemental polynomial degree.

A key tool in our a posteriori error analysis is the *elliptic reconstruction technique* [8, 14, 15]. Roughly speaking, in the elliptic reconstruction framework the error is split into a *parabolic* and an *elliptic* part, respectively. In the interest of being explicit with respect to the dependence of the a posteriori error bounds in the elemental polynomial degree p, we restrict the presentation to quadrilateral elements of tensor-product type (cf. Remark 2).

2 Model Problem and the IPDG Method

Let Ω be a bounded open (curvilinear) polygonal domain with Lipschitz boundary $\partial\Omega$ in $\mathbb{R}^d$, $d = 2, 3$. For $\omega \subset \Omega$, we consider the standard spaces $L^2(\omega)$ (whose norm is denoted by $\|\cdot\|_\omega$ for brevity), $H^1(\omega)$ and $H^1_0(\omega)$, whose norm will be denoted by $\|\cdot\|_1$, along with its dual $H^{-1}(\Omega)$, with norm $\|\cdot\|_{-1}$. For brevity, the standard inner product on $L^2(\Omega)$ will be denoted by $\langle\cdot,\cdot\rangle$ and the corresponding norm by $\|\cdot\|$. We also define the spaces $L^2(0,T,X)$, $X \in \{L^2(\omega), H^{\pm1}(\omega)\}$ and $L^\infty(0,T,L^2(\Omega))$, consisting of all measurable functions $v : [0,T] \to X$, for which $\|v\|_{L^2(0,T;X)} := \left(\int_0^T \|v(t)\|_X^2\right)^{1/2} < +\infty$ and $\|v\|_{L^\infty(0,T;L^2(\Omega))} := \operatorname{ess\,sup}_{t\in[0,T]}\|v(t)\|$. (The differentials in the integrals with respect to t are suppressed for brevity throughout this work.)

We identify function $v \in [0,T]\times\Omega \to \mathbb{R}$ with $v : t \to X$ and we denote $v(t)$, $t \in [0,T]$, for $v \in [0,T]\times\Omega \to \mathbb{R}$.

For $t \in (0,T]$, we consider the problem of finding a function u satisfying

$$u_t(t,x) - \nabla\cdot(a(t,x,|\nabla u(t,x)|)\nabla u(t,x)) = f(t,x) \quad \text{in } (0,T]\times\Omega, \tag{1}$$

where $f \in L^\infty(0,T;L^2(\Omega))$ and a scalar uniformly continuous function, subject to initial condition $u(0,x) = u_0(x)$ on $\{0\}\times\Omega$, for $u_0 \in L^2(\Omega)$, and homogeneous Dirichlet boundary conditions on $[0,T]\times\partial\Omega$.

We assume that the non-linearity a in equation (1) is of strongly monotone type with Lipschitz growth so that there exist positive constants $\underline{a}$ and $\overline{a}$ such that the following inequalities hold:

$$|a(t,x,|y|)y - a(t,x,|z|)z| \le \overline{a}|y-z| \tag{2}$$

$$\big(a(t,x,|y|)y - a(t,x,|z|)z\big)\cdot(y-z) \ge \underline{a}|y-z|^2, \tag{3}$$

for all vectors $y, z \in \mathbb{R}^d$, and all $(t,x) \in [0,T]\times\bar\Omega$.

Let $\mathscr{T}$ be a shape-regular subdivision of Ω into disjoint closed quadrilateral elements $\kappa \in \mathscr{T}$. We assume that $\kappa \in \mathscr{T}$ are constructed via C^∞-diffeomorphisms with non-singular Jacobian $F_\kappa : (-1,1)^d \to \kappa$, so as to ensure $\bar\Omega = \cup_{\kappa\in\mathscr{T}}\bar\kappa$.

For $p \in \mathbb{N}$, $\mathscr{Q}_p(\hat{\kappa})$ is the set of all tensor-product polynomials on $(-1,1)^d$ of degree p in each variable and let

$$S^p := \{v \in L^2(\Omega) : v|_{F_\kappa} \in \mathscr{Q}_p(\hat{\kappa}),\ \kappa \in \mathscr{T}\}, \tag{4}$$

be the (discontinuous) *finite element space*. Let Γ be the union of all $(d-1)$-dimensional element faces e associated with the subdivision $\mathscr{T}$ (including the boundary). Let also $\Gamma_{\text{int}} := \Gamma \backslash \partial\Omega$, so that $\Gamma = \partial\Omega \cup \Gamma_{\text{int}}$.

Let κ^+, κ^- be two (generic) elements sharing a face $e := \kappa^+ \cap \kappa^- \subset \Gamma_{\text{int}}$ with respective outward normal unit vectors $\mathbf{n}^+$ and $\mathbf{n}^-$ on e. For $q : \Omega \to \mathbb{R}$ and $\boldsymbol{\phi} : \Omega \to \mathbb{R}^d$, let $q^\pm := q|_{e\cap\partial\kappa^\pm}$ and $\boldsymbol{\phi}^\pm := \boldsymbol{\phi}|_{e\cap\partial\kappa^\pm}$, and set

$$\begin{aligned} \{q\}|_e &:= \frac{1}{2}(q^+ + q^-), & \{\boldsymbol{\phi}\}|_e &:= \frac{1}{2}(\phi^+ + \phi^-), \\ [\![q]\!]|_e &:= q^+\mathbf{n}^+ + q^-\mathbf{n}^-, & [\![\boldsymbol{\phi}]\!]|_e &:= \phi^+ \cdot \mathbf{n}^+ + \boldsymbol{\phi}^- \cdot \mathbf{n}^-; \end{aligned}$$

if $e \subset \partial\kappa \cap \partial\Omega$, we set $\{\boldsymbol{\phi}\}|_e := \boldsymbol{\phi}^+$ and $[\![q]\!]|_e := q^+\mathbf{n}^+$. Finally, we introduce the *meshsize* $h : \Omega \to \mathbb{R}$, defined by $h(x) = \text{diam}\kappa$, if $x \in \kappa\backslash\partial\kappa$ and $h(x) = \{h\}$, if $x \in \Gamma$.

Consider the IPDG semi-linear form $B(\cdot,\cdot) : S^p \times S^p \to \mathbb{R}$, introduced in [9] for the solution of the corresponding steady-state problem, defined by

$$\begin{aligned} B(w,v) := \sum_{\kappa\in\mathscr{T}} \int_\kappa \alpha(w)\cdot\nabla v\,\mathrm{d}x + \int_\Gamma \big(\theta\{a(t,x,h^{-1}|[\![w]\!]|)\nabla v\}\cdot[\![w]\!] \\ - \{\alpha(w)\}\cdot[\![v]\!] + \sigma[\![w]\!]\cdot[\![v]\!]\big)\,\mathrm{d}s, \end{aligned} \tag{5}$$

where $\alpha(w) := a(t,\cdot,|\nabla w|)\nabla w$, $w \in H^1(\Omega) + S^p$, for $\theta \in \{-1,0,1\}$, with the function $\sigma : \Gamma \to \mathbb{R}_+$ defined piecewise by $\sigma|_e := C_\sigma p^2/(h|_e)$, for some sufficient large constant $C_\sigma > 0$. The corresponding energy norm $|||\cdot|||$ is defined $|||w||| := \big(\sum_{\kappa\in\mathscr{T}} \|\nabla w\|_\kappa^2 + \int_\Gamma \sigma[\![w]\!]^2\mathrm{d}s\big)^{1/2}$, for $w \in H^1(\Omega) + S^p$. The (spatially semidiscrete) *interior penalty discontinuous Galerkin method* (IPDG) for the initial/boundary value model problem reads:

$$\text{find } U : (0,T] \to S^p \text{ such that } \langle U_t, V\rangle + B(U,V) = \langle f, V\rangle\ \forall t \in (0,T], V \in S^p. \tag{6}$$

3 A Posteriori Error Bounds

For $w \in H^1(\Omega) + S^p$, and $T > 0$, we define the norm $|||w|||_{L^2(0,T;H^1(\Omega))} := \Big(\int_0^T |||w|||^2\Big)^{1/2}$, $t > 0$. We shall derive a posteriori bounds for the error $|||u - U|||_{L^2(0,T;H^1(\Omega))}$.

Definition 1 (elliptic reconstruction). Let U be the (semi-discrete) solution to the problem (6) and fix $t \in [0,T]$. We define the *elliptic reconstruction* $w \equiv w(t) \in H_0^1(\Omega)$ of U to be the solution to the elliptic problem

$$\langle \alpha(w), \nabla v\rangle = \langle g, v\rangle \quad \forall v \in H_0^1(\Omega), \tag{7}$$

where $g \equiv g(t)$ is given by $g := -AU + f - \Pi f$, with $\Pi : L^2(\Omega) \to S^p$ is the orthogonal L^2-projection operator onto S^p and $A \equiv A(t) : S^p \to S^p$ is the discrete operator defined by

$$\text{for } Z \in S^p, \quad \langle -AZ, V\rangle = B(Z,V) \quad \forall V \in S^p. \tag{8}$$

The construction of w and that of AZ are both well defined in view of the elliptic problem's unique solvability and the Riesz representation, respectively.

Remark 1. The key property of the construction in Definition 1 is that U is the IPDG solution of an elliptic problem with analytical solution w. Namely, for each fixed $t \in [0,T]$ it satisfies

$$\text{find } U \in S^p \text{ such that } B(U,V) = \langle g, V\rangle \quad \forall V \in S^p. \tag{9}$$

We can now decompose the error as follows:

$$U - u = \rho - \epsilon, \text{ with } \rho := w - u, \text{ and } \epsilon := w - U, \tag{10}$$

where $w \equiv w(t)$ denotes the elliptic reconstruction of $U \equiv U(t)$, $t \in [0,T]$.

Lemma 1 (differential error relation). *Let u, w, U, e, ρ, ϵ as above. Then, for all $v \in H_0^1(\Omega)$, we have*

$$\langle e_t, v\rangle + \langle \alpha(w) - \alpha(u), \nabla v\rangle = 0. \tag{11}$$

Proof. We have

$$\begin{aligned} &\langle e_t, v\rangle + \langle \alpha(w) - \alpha(u), \nabla v\rangle = \langle U_t, v\rangle + \langle \alpha(w), \nabla v\rangle - \langle f, v\rangle \\ &= \langle U_t, v\rangle + \langle g, v\rangle - \langle f, v\rangle = \langle U_t, v\rangle + \langle -AU, v\rangle - \langle \Pi f, v\rangle \\ &= \langle U_t, \Pi v\rangle + \langle -AU, \Pi v\rangle - \langle f, \Pi v\rangle = 0, \end{aligned} \tag{12}$$

using (1), (7) and the properties of the L^2-projection, respectively. □

We consider further the decomposition of U into *conforming* and *non-conforming* (discontinuous) parts $U = U^c + U^d$, where $U^c \in S^p \cap H_0^1(\Omega)$ and $U^d := U - U^c \in S^p$. Note that there are many ways of performing this decomposition (e.g., by projecting U onto the conforming space) whereof the specific nature remains at our disposal until further.

We also use the shorthand notation $e^c := U^c - u$ and $\epsilon^c := w - U^c$; note that $e^c = \rho - \epsilon^c$, $e = e^c + U^d$ and that $e^c \in H_0^1(\Omega)$.

Theorem 1 (abstract a posteriori energy-error estimate). *With u, U, U^d, e, and ϵ as defined above, the following error estimate is satisfied:*

$$\begin{aligned}|||e|||_{L^2(0,T,H^1(\Omega))} \leq & C_1|||\epsilon|||_{L^2(0,T,H^1(\Omega))} + \underline{a}^{-\frac{1}{2}}\big(\|u_0 - U(0)\| + \|U^d(0)\|\big) \\ & + C_1|||U^d|||_{L^2(0,T,H^1(\Omega))} + C_2\|U_t^d\|_{L^2(0,T,H^{-1}(\Omega))},\end{aligned} \tag{13}$$

with $C_1 := 1 + \sqrt{2}\overline{a}\underline{a}^{-1}$ and $C_2 := \sqrt{2}\underline{a}^{-1}$.

Proof. Set $v = e^c$ in (11), to deduce

$$\langle e_t^c, e^c\rangle + \langle \alpha(U^c) - \alpha(u), \nabla e^c\rangle = -\langle U_t^d, e^c\rangle + \langle \alpha(U^c) - \alpha(w), \nabla e^c\rangle. \tag{14}$$

Conditions (3) and (2) imply, respectively,

$$\langle \alpha(U^c) - \alpha(u), \nabla e^c\rangle \geq \underline{a}\|\nabla e^c\|^2, \text{ and } \langle \alpha(U^c) - \alpha(w), \nabla e^c\rangle \leq \overline{a}\|\nabla \epsilon^c\|\,\|\nabla e^c\|,$$

and the duality pairing (H^{-1}, H_0^1) gives $|\langle U_t^d, e^c\rangle| \leq \|U_t^d\|_{-1}\|\nabla e^c\|$. Using the last 3 relations on (14), we deduce

$$\langle e_t^c, e^c\rangle + \underline{a}\|\nabla e^c\|^2 \leq \Big(\|U_t^d\|_{-1} + \overline{a}\|\nabla \epsilon^c\|\Big)\|\nabla e^c\|, \tag{15}$$

which, in turn, implies

$$\langle e_t^c, e^c\rangle + \frac{\underline{a}}{2}\|\nabla e^c\|^2 \leq \frac{1}{2\underline{a}}\big(\|U_t^d\|_{-1} + \overline{a}\|\nabla \epsilon^c\|\big)^2. \tag{16}$$

Integrating (16) with respect to t between 0 and T, yields

$$\|e^c(t)\|^2 + \underline{a}\int_0^T \|\nabla e^c\|^2 \leq \|e^c(0)\|^2 + \frac{1}{\underline{a}}\int_0^T \Big(\|U_t^d\|_{-1} + \overline{a}\|\nabla \epsilon^c\|\Big)^2,$$

or

$$\begin{aligned}\Big(\int_0^T \|\nabla e^c\|^2\Big)^{\frac{1}{2}} \leq & \underline{a}^{-\frac{1}{2}}\|e^c(0)\| + \underline{a}^{-1}\Big(\int_0^T \Big(\|U_t^d\|_{-1} + \overline{a}\|\nabla \epsilon^c\|\Big)^2\Big)^{\frac{1}{2}} \\ \leq & \underline{a}^{-\frac{1}{2}}\|e^c(0)\| + C_2\|U_t^d\|_{L^2(0,T,H^{-1}(\Omega))} \\ & + (C_1 - 1)|||\epsilon^c|||_{L^2(0,T,H^1(\Omega))}\end{aligned} \tag{17}$$

noting that $|||\epsilon^c||| = \|\nabla \epsilon^c\|$. Using the bounds $|||\epsilon^c||| \leq |||\epsilon||| + |||U^d|||$, $\|e^c(0)\| \leq \|e(0)\| + \|U^d(0)\|$ on (17) and the resulting bound on the triangle inequality

$$|||e|||_{L^2(0,T,H^1(\Omega))} \leq |||\epsilon^c|||_{L^2(0,T,H^1(\Omega))} + |||U^d|||_{L^2(0,T,H^1(\Omega))},$$

yields the result. □

For the above result to yield a formally a posteriori bound, we need to estimate $|||\epsilon|||_{L^2(0,T,H^1(\Omega))}$ further. In particular, in view of Remark 1, we require an a posteriori error bound for the IPDG method for the corresponding elliptic quasilinear problem (9). Such a result is available in [11], an instance of which and is presented next.

Theorem 2 ([11]). *Let $w \in H^1_0(\Omega)$ be the elliptic reconstruction defined in (7) and let $W \in S^p$ be the solution of (9). Then, for $C_\sigma > 1$ sufficiently large the bound*

$$|||w - W|||^2 \le \mathscr{E}(W, g, S^p) := C_{\text{est}} \sum_{\kappa\in\mathscr{T}} \Big(\eta_\kappa^2 + \mathscr{O}(g, W)\Big), \tag{18}$$

holds, with

$$\eta_\kappa^2 = \frac{h_\kappa^2}{p^2}\|\tilde{\Pi}\big(g + \nabla\cdot\alpha(W)\big)\|_\kappa^2 + \frac{h_\kappa}{p}\|\tilde{\Pi}_\Gamma[\![\alpha(W)]\!]\|^2_{\partial\kappa\setminus\partial\Omega} + C_\sigma^2\frac{p^3}{h_\kappa}\|[\![W]\!]\|^2_{\partial\kappa},$$

and

$$\mathscr{O}(g, w_{\text{DG}}) = \sum_{\kappa\in\mathscr{T}}\Big(\frac{h_\kappa^2}{p^2}\|\big(\mathbb{I} - \tilde{\Pi}\big)\big(g + \alpha(W)\big)\|_\kappa^2 + \frac{h_\kappa}{p}\|\big(\mathbb{I} - \tilde{\Pi}_\Gamma)[\![\alpha(W)]\!]\|^2_{\partial\kappa\setminus\partial\Omega}\Big),$$

where $\mathbb{I}$ denotes a generic identity operator, $\tilde{\Pi}$ denotes the L^2-projection operator onto S^{p-1}, $\tilde{\Pi}_\Gamma$ is defined piecewise by $\tilde{\Pi}_\Gamma v|_e := \pi_e^{p-1}v$, for all elemental faces $e \subset \Gamma$, $v \in L^2(\Omega)$, where $\pi_e^{p-1} : L^2(\Omega) \to \mathscr{P}_{p-1}(e)$ denotes the L^2-projection operator of the trace on the face e of a function in S^{p-1} (with $\mathscr{P}_{p-1}(e)$, for $e \subset \bar{\kappa}$ the space of mapped univariate polynomials of degree at most $p-1$ on e), and $C_{\text{est}} > 0$ is independent of C_σ, θ, h and p. □

Also, it is possible to further estimate the terms involving U^d, to avoid computing U^d explicitly. This is done (with, crucially, *explicit* dependence on p) using the following result based on [4, Lemma 3.2].

Lemma 2. *Suppose $\mathscr{T}$ does not contain any hanging nodes. Then, for any $v \in S^p$ and any multi-index γ, with $|\gamma| = 0, 1$, there exists a function $v^c \in S^p \cap H^1_0(\Omega)$ such that*

$$\sum_{\kappa\in\mathscr{T}}\|D^\gamma(v - v^c)\|_\kappa^2 \le C_3\|\Big(\frac{h}{p^2}\Big)^{\frac{1}{2}-|\gamma|}[\![v]\!]\|_\Gamma^2, \tag{19}$$

with $C_3 > 0$ depending on the maximal angle of $\mathscr{T}$ only.

Proof. [4, Lemma 3.2] implies that for every $\kappa \in \mathscr{T}$ there exists an Oswald-type operator $I_{\text{Os}} : S^p \to S^p \cap H^1_0(\Omega)$, such that

$$\|v - I_{\text{Os}}v\|_\kappa^2 \le C\sum_{e\subset\mathscr{F}(\kappa)}\frac{h_\kappa}{p^2}\|[\![v]\!]\|_e^2, \tag{20}$$

for all $v \in S^p$, with $\mathscr{F}(\kappa) := \{e \in \Gamma : e \cap \bar{\kappa} \neq \emptyset\}$. Summing over all the elements $\kappa \in \mathscr{T}$, and observing that the maximal angle and the lack of hanging nodes gives an upper bound on the cardinality of $\mathscr{F}(\kappa)$ for all $\kappa \in \mathscr{T}$, we deduce that

$$\sum_{\kappa\in\mathscr{T}} \|v - I_{\mathrm{Os}}v\|_\kappa^2 \leq C \sum_{e\subset\Gamma} \frac{h_\kappa}{p^2} \|[\![v]\!]\|_e^2, \tag{21}$$

which shows (19) for $|\gamma| = 0$. To show (19) for $|\gamma| = 1$, we observe that $(v - I_{\mathrm{Os}}v) \in S^p$; thus, the standard inverse estimate yields:

$$\sum_{\kappa\in\mathscr{T}} \|\nabla(v - I_{\mathrm{Os}}v)\|_\kappa^2 \leq C \sum_{\kappa\in\mathscr{T}} \frac{p^4}{h_\kappa^2} \|v - I_{\mathrm{Os}}v\|_\kappa^2 \leq C \sum_{e\subset\Gamma} \frac{p^2}{h_\kappa} \|[\![v]\!]\|_e^2, \tag{22}$$

using the shape regularity of $\mathscr{T}$. Setting $v^c = I_{\mathrm{Os}}v$, the result follows. □

Remark 2. The assumptions of Lemma 2 pose the following restrictions on the finite element space S^p: the use of quadrilateral elements (as the tensor-product nature of the local elemental bases is of crucial importance here), the exclusion of hanging nodes and the uniformity of the polynomial degree. If explicit knowledge of the polynomial degree p in the a posteriori bounds presented in this work is not required, then these restrictions are not needed in view of [12, Lemma 4.1], i.e., triangular elements containing hanging nodes can be employed.

Combining the results of Theorems 1 and 2, together with the approximation properties described in Lemma 2, we obtain an a posteriori error bound in the energy norm for the semi-discrete problem (6).

Theorem 3 (energy-norm a posteriori bound). *With the notation of Theorem 1 and the assumptions of Lemma 2, the following error bound holds:*

$$\begin{aligned} |||e|||_{L^2(0,T,H^1(\Omega))} \leq & C_1 \int_0^T \mathscr{E}^2(U, g, S^p) + \underline{a}^{-\frac{1}{2}} \|u_0 - U(0)\| \\ & + \underline{a}^{-\frac{1}{2}} C_3 \|\big(\frac{h}{p^2}\big)^{\frac{1}{2}} [\![U(0)]\!]\|_\Gamma^2 + C_4 \|\sqrt{\sigma}[\![U]\!]\|_{L^2(0,T;L^2(\Gamma))} \\ & + C_5 \|\big(\frac{h}{p^2}\big)^{\frac{1}{2}} [\![U_t]\!]\|_{L^2(0,T;L^2(\Gamma))}, \end{aligned} \tag{23}$$

with $C_4 := C_1\sqrt{C_3/C_\sigma}$, $C_5 := C_2 C_{\mathrm{PF}}$ *and* $C_{\mathrm{PF}} > 0$ *(the Poincaré–Friedrichs constant), such that* $\|v\|_{-1} \leq C_{\mathrm{PF}}\|v\|$, *for all* $v \in L^2(\Omega)$.

Proof. Combining the results from Theorems 1 and 2, together with the approximation properties described in Lemma 2, the result follows. □

Acknowledgements Omar Lakkis acknowledges the support of Royal Society UK International Travel Grants.

References

1. D. N. Arnold, *An interior penalty finite element method with discontinuous elements*, SIAM J. Numer. Anal., 19 (1982), pp. 742–760
2. G. A. Baker, *Finite element methods for elliptic equations using nonconforming elements*, Math. Comp., 31 (1977), pp. 45–59
3. R. Becker, P. Hansbo, and M. G. Larson, *Energy norm a posteriori error estimation for discontinuous Galerkin methods*, Comput. Methods Appl. Mech. Engrg., 192 (2003), pp. 723–733
4. E. Burman and A. Ern, *Continuous interior penalty hp-finite element methods for advection and advection-diffusion equations*, Math. Comp., 76 (2007), pp. 1119–1140 (electronic)
5. Y. Chen and J. Yang, *A posteriori error estimation for a fully discrete discontinuous Galerkin approximation to a kind of singularly perturbed problems*, Finite Elem. Anal. Des., 43 (2007), pp. 757–770
6. A. Ern and J. Proft, *A posteriori discontinuous Galerkin error estimates for transient convection-diffusion equations*, Appl. Math. Lett., 18 (2005), pp. 833–841
7. A. Ern and M. Vohralík, *A posteriori error estimation based on potential and flux reconstruction for the heat equation* (Submitted)
8. E. H. Georgoulis and O. Lakkis, *A posteriori error control for discontinuous Galerkin methods for parabolic problems* (Submitted)
9. P. Houston, J. Robson, and E. Süli, *Discontinuous Galerkin finite element approximation of quasilinear elliptic boundary value problems I: The scalar case*, IMA J. Numer. Anal., 25 (2005), pp. 726–749
10. P. Houston, D. Schötzau, and T. P. Wihler, *Energy norm a posteriori error estimation of hp-adaptive discontinuous Galerkin methods for elliptic problems*, Math. Models Methods Appl. Sci., 17 (2007), pp. 33–62
11. P. Houston, E. Süli, and T. P. Wihler, *A posteriori error analysis of hp-version discontinuous Galerkin finite element methods for second-order quasilinear elliptic problems*, IMA J. Numer. Anal. (to appear)
12. O. A. Karakashian and F. Pascal, *A posteriori error estimates for a discontinuous Galerkin approximation of second-order elliptic problems*, SIAM J. Numer. Anal., 41 (2003), pp. 2374–2399 (electronic)
13. O. A. Karakashian and F. Pascal, *Convergence of adaptive discontinuous Galerkin approximations of second-order elliptic problems*, SIAM J. Numer. Anal., 45 (2007), pp. 641–665 (electronic)
14. O. Lakkis and C. Makridakis, *Elliptic reconstruction and a posteriori error estimates for fully discrete linear parabolic problems*, Math. Comp., 75 (2006), pp. 1627–1658 (electronic)
15. C. Makridakis and R. H. Nochetto, *Elliptic reconstruction and a posteriori error estimates for parabolic problems*, SIAM J. Numer. Anal., 41 (2003), pp. 1585–1594 (electronic)
16. S. Sun and M. F. Wheeler, *$L^2(H^1)$ norm a posteriori error estimation for discontinuous Galerkin approximations of reactive transport problems*, J. Sci. Comput., 22/23 (2005), pp. 501–530
17. M. F. Wheeler, *An elliptic collocation-finite element method with interior penalties*, SIAM J. Numer. Anal., 15 (1978), pp. 152–161
18. J.-M. Yang and Y.-P. Chen, *A unified a posteriori error analysis for discontinuous Galerkin approximations of reactive transport equations*, J. Comput. Math., 24 (2006), pp. 425–434

An A Posteriori Analysis of Multiscale Operator Decomposition

Victor Ginting

Abstract We analyze an operator decomposition time integration method for systems of ordinary differential equations that present significantly different scales within the components of the model. Using adjoint formulation of the problems, we derive an a posteriori error analysis for the average error over certain time interval as quantity of interest.

1 Problem Setting

In this paper, we analyze an iterative operator decomposition technique for solving a system of ordinary differential equations that presents significantly different scales for the rate of change of individual components of the model. The technique employs discretizations on significantly different time scales for different components of the problem. We consider a model that can be decomposed into two vector-valued components: find $y = (y_1, y_2)^\top \in \mathbb{R}^n$ that satisfies

$$\begin{cases} \dot{y}_1 = F_1(y_1, y_2), & t \in (0, T], \\ \dot{y}_2 = F_2(y_1, y_2), & t \in (0, T], \\ y_1(0) = g_1, y_2(0) = g_2, \end{cases} \tag{1}$$

for a given initial condition $g = (g_1, g_2)^\top$. Here, $y_i \in \mathbb{R}^{n_i}$, $i = 1, 2$, $n = n_1 + n_2$, and $F = (F_1, F_2)^\top \in \mathbb{R}^n$, with $F_i(y) = F_i(y_1, y_2) \in \mathbb{R}^{n_i}$, $i = 1, 2$. If F_1 and F_2 induce significantly different rates of change in the respective solution components, then an heuristic consideration of accuracy suggests that it is most efficient to solve (1) using small time steps for the fast component and large time steps for the slow

V. Ginting (✉)
Department of Mathematics, University of Wyoming, Dept. 3036 1000 E. University Ave. Laramie, WY, USA 82071
e-mail: vginting@uwyo.edu

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_38,

component. The natural issues to consider are accuracy, stability, and estimation of the error of particular numerical solutions. Our goal in this paper is to derive accurate error estimates for a quantity of interest computed from the decomposition technique. This decomposition can also be viewed as a multirate numerical method. There is a large literature on multirate numerical methods, see for example [1–3, 7–11]. By and large, these references are focused on application and standard a priori analysis issues, e.g., stability, accuracy, and convergence properties. The main goal of this paper is to derive a computable a posteriori error representation that accurately estimates the error in a specified quantity of interest computed from a operator decomposition of (1). In this case, we choose our quantity interest to be the average of the solution over certain time interval.

We discretize $[0, T]$ into $0 = t_0 < t_1 < t_2 < \cdots < t_N = T$ with time steps $\{\Delta t_n =\}_{n=1}^{N}$, with $\Delta t_n = t_n - t_{n-1}$ and $\Delta t = \max_{1\le n\le N}\{\Delta t_n\}$ and time intervals $I_n = (t_{n-1}, t_n)$. For the purpose of operator decomposition we think of these nodes as synchronization times. To each I_n, we assign a positive integer M_n which is the number of iterations to be used when synchronizing the fast and slow components. Let $\tilde{y} = (\tilde{y}_1, \tilde{y}_2)^\top$ denote the analytic solution of (1) using the iterative procedure defined in Alg. 1.

Algorithm 1 Analytic Iterative Operator Decomposition

for $n = 1$ to N **do**
  Set $\tilde{y}_2^{(0)} = \tilde{y}_2^{(M_{n-1})}(t_{n-1})$
  for $m = 1$ to M_n **do**

$$\text{Compute } \tilde{y}_1^{(m)}(t) \text{ for } t \in I_n \text{ satisfying } \begin{cases} \dot{\tilde{y}}_1^{(m)} = F_1(\tilde{y}_1^{(m)}, \tilde{y}_2^{(m-1)}) \\ \tilde{y}_1^{(m)}(t_{n-1}) = \tilde{y}_1^{(M_{n-1})}(t_{n-1}) \end{cases} \tag{2}$$

$$\text{Compute } \tilde{y}_2^{(m)}(t) \text{ for } t \in I_n \text{ satisfying } \begin{cases} \dot{\tilde{y}}_2^{(m)} = F_2(\tilde{y}_1^{(m)}, \tilde{y}_2^{(m)}) \\ \tilde{y}_2^{(m)}(t_{n-1}) = \tilde{y}_2^{(M_{n-1})}(t_{n-1}). \end{cases} \tag{3}$$

  end for
end for

Let $L_{i,n}$, $i = 1, 2$ be two positive integers, where $L_{1,n}$ denotes the number of time steps used to solve the fast subsystem and $L_{2,n}$ the number of steps used for the slow subsystem. Without loss of generality, we assume $L_{1,n} = d_n L_{2,n}$ for some positive integer d_n, i.e., $L_{1,n}$ is divisible by $L_{2,n}$. We denote time steps for each component in the Galerkin formulation by $\Delta s_{i,n} = \Delta t_n / L_{i,n}$, with $\Delta s_i = \max_{1\le n\le N}\{\Delta s_{i,n}\}$. We use the continuous Galerkin method, which is more appropriate for problems with conserved quantities [4]. The finite element approximate solutions are sought in piecewise continuous polynomial spaces,

$$\mathscr{V}^{(q_i)}(I_n) = \left\{U : U|_{I_{l,n}} \in \mathscr{P}^{(q_i)}(I_{l,n}), 1 \le l \le L_{i,n}\right\}, \quad i = 1, 2.$$

The finite element solution of the multiscale iterative operator decomposition is to find $\widetilde{Y} = (\widetilde{Y}_1, \widetilde{Y}_2)^\top$ with $\widetilde{Y}_1|_{I_n} \in \mathscr{V}^{(q_1)}(I_n)$ and $\widetilde{Y}_2|_{I_n} \in \mathscr{V}^{(q_2)}(I_n)$ determined by Alg. 2.

Algorithm 2 Finite Element Multiscale Iterative Operator Decomposition

for $n = 1$ to N **do**
  Set $\widetilde{Y}_2^{(0)} = \widetilde{Y}_2^{(M_{n-1})}(t_{n-1})$
  for $m = 1$ to M_n **do**
    Set $\widetilde{Y}^{(m)}(t_{n-1}) = \widetilde{Y}^{(M_{n-1})}(t_{n-1})$.
    for $j = 1$ to $L_{1,n}$ **do**
      Compute $\widetilde{Y}_1^{(m)}(t)$ for $t \in I_{j,n}$ satisfying

$$\int_{I_{j,n}} \left(\dot{\widetilde{Y}}_1^{(m)} - F_1(\widetilde{Y}_1^{(m)}, \widetilde{Y}_2^{(m-1)}), V \right) dt = 0, \quad \forall\, V \in \mathscr{P}^{(q_1-1)}(I_{j,n}).$$

    end for
    for $k = 1$ to $L_{2,n}$ **do**
      Compute $\widetilde{Y}_2^{(m)}(t)$ satisfying

$$\int_{I_{k,n}} \left(\dot{\widetilde{Y}}_2^{(m)} - F_2(\widetilde{Y}_1^{(m)}, \widetilde{Y}_2^{(m)}), W \right) dt = 0, \quad \forall\, W \in \mathscr{P}^{(q_2-1)}(I_{k,n}).$$

    end for
  end for
end for

We note that many standard finite difference schemes can be obtained by applying appropriate quadrature formulas to the integrals defining the finite element approximation.

2 A Posteriori Analysis

A key feature of the analysis is the realization that the multiscale operator decomposition problem is naturally associated with a different adjoint operator than the original problem. Our approach [5] to overcome this issue is to use a different linearization than commonly used for nonlinear problems. We assume that the operators for the original problem and the analytic operator decomposition version share a common solution, and use that as a linearization point. The simplest example of such a solution is a steady state solution, which can be guaranteed to exist by assuming homogeneity in the right-hand side, i.e., $F(0) = 0$. This is generally not restrictive in practice, but this assumption can be generalized (see [5]). We let

$$\overline{F'_{ij}(y)} = \int_0^1 \frac{\partial F_i}{\partial y_j}(sy)\, ds, \quad i, j = 1, 2, \tag{4}$$

and $\overline{F'}$ denotes the square matrix whose entries are (4). Then $F(y) = \overline{F'(y)}y$ and so $\dot{y} = \overline{F'(y)}y$. Associated with this linearized form, we denote by φ the generalized Green's function satisfying the following adjoint problem:

$$\begin{cases} -\dot{\varphi} = \overline{F'(y)}^{\top}\varphi + \psi, & t \in (T, 0], \\ \varphi(T) = 0. \end{cases} \tag{5}$$

We can obtain a solution representation using the Green's functions, by multiplying y with the adjoint equation (5) and integrating by parts over I_n

$$\int_{I_n} (y, \psi)\, dt + (y_n, \varphi_n) = (y_{n-1}, \varphi_{n-1}). \tag{6}$$

To simplify presentation, we express the analytic iterative operator decomposition in Alg. 1 in a more compact format. In particular, for any iteration index m, we write (2) and (3) as

$$\dot{\tilde{y}}^{(m)} = F(\tilde{y}^{(m)}) + \delta_{\tilde{y}}^{(m)}, \tag{7}$$

where

$$\delta_{\tilde{y}}^{(m)} = -\left[F_1(\tilde{y}_1^{(m)}, \tilde{y}_2^{(m)}) - F_1(\tilde{y}_1^{(m)}, \tilde{y}_2^{(m-1)}), \quad 0\right]^{\top}. \tag{8}$$

The vector $\delta_{\tilde{y}}^{(m)}$ can be interpreted as a residual at the iteration level m.

To define an adjoint for the operator decomposition problem in Alg. 1, we let $\tilde{\varphi}_i$ denote the generalized Green's function that satisfies an adjoint problem on time interval I_n as given in Alg. 3.

Algorithm 3 Adjoint for the Analytic Iterative Operator Decomposition

Set $\tilde{\varphi}_1^{(0)} = \tilde{\varphi}_1^{(K_n+1)}$
for $k = 1$ to K_n **do**
Compute $\tilde{\varphi}_2^{(k)}$ satisfying

$$-\dot{\tilde{\varphi}}_2^{(k)} = \overline{F'_{22}(\tilde{y}^{(m)})}^{\top}\tilde{\varphi}_2^{(k)} + \overline{F'_{12}(\tilde{y}^{(m)})}^{\top}\tilde{\varphi}_1^{(k-1)} + \psi_2, \quad t \in (t_n, t_{n-1}]$$

Compute $\tilde{\varphi}_1^{(k)}$ satisfying

$$-\dot{\tilde{\varphi}}_1^{(k)} = \overline{F'_{11}(\tilde{y}^{(m)})}^{\top}\tilde{\varphi}_1^{(k)} + \overline{F'_{21}(\tilde{y}^{(m)})}^{\top}\tilde{\varphi}_2^{(k)} + \psi_1, \quad t \in (t_n, t_{n-1}]$$

end for

In this algorithm, it is understood that we set $\tilde{\varphi}(T) = 0$. Notice that the adjoint problems are solved backward in time and in the reverse order to that of the forward problem, starting with $\tilde{\varphi}_2$ followed by $\tilde{\varphi}_1$. These generalized Green's functions are an iterative approximation of (5). As in the forward problem, we can also rewrite this last algorithm into a compact form

$$-\dot{\tilde{\varphi}}^{(k)} = \overline{F'(\tilde{y}^{(m)})}^{\top} \tilde{\varphi}^{(k)} + \psi + \xi^{(k)}, \tag{9}$$

for adjoint iteration level k. Here

$$\xi^{(k)} = -\left[0, \quad \overline{F'_{12}(\tilde{y}^{(m)})}^{\top} (\tilde{\varphi}_1^{(k)} - \tilde{\varphi}_1^{(k-1)})\right]^{\top},$$

is the residual of the adjoint at iteration level k. To get a representation of the iterative operator decomposition solution, we follow a similar derivation as for the fully coupled problem. Multiplying $\tilde{y}^{(m)}$ with (9), integrating over I_n, and applying integration by parts along with (7), we obtain

$$\begin{aligned}\int_{I_n} (\tilde{y}^{(m)}, \psi)\, dt + (\tilde{y}_n^{(m)}, \tilde{\varphi}_n^{(k)}) = (\tilde{y}_{n-1}^{(m)}, \tilde{\varphi}_{n-1}^{(k)}) + \int_{I_n} (\delta_{\tilde{y}}^{(m)}, \tilde{\varphi}^{(k)})\, dt \\ - \int_{I_n} (\tilde{y}^{(m)}, \xi^{(k)})\, dt.\end{aligned} \tag{10}$$

We note that this representation is not in the standard format as in (6), in which the solution at the current time level solely depends on the previous time level values. It contains artifacts arising from the iterative procedure used to compute both forward and backward problems. The second term on the right hand side of (10) can be interpreted as the weighted average of the forward problem residual over a time step. The third term, on the other hand, is the weighted average of the backward problem residual over a time step. Thus, the iterative nature of the operator decomposition is reflected in this representation. Once convergence is reached both on forward and backward problems, then the standard convention of solution representation using the adjoint technique is recovered. We are now able to express the error representation of the iterative operator decomposition method. By subtracting (10) from (6) and setting $\tilde{y}_{n-1}^{(m)} = y_{n-1}$ and $\varphi_n = \tilde{\varphi}_n^{(k)}$, we get an error equation over one time step:

$$\begin{aligned}\int_{I_n} (e^{(m)}, \psi)\, dt + (e_n^{(m)}, \tilde{\varphi}_n^{(k)}) = (y_{n-1}, \Delta\tilde{\varphi}_{n-1}^{(k)}) - \int_{I_n} (\delta_{\tilde{y}}^{(m)}, \tilde{\varphi}^{(k)})\, dt \\ + \int_{I_n} (\tilde{y}^{(m)}, \xi^{(k)})\, dt,\end{aligned}$$

where $e^{(m)} = y - \tilde{y}^{(m)}$ and $\Delta\tilde{\varphi}_{n-1}^{(k)} = \varphi_{n-1} - \tilde{\varphi}_{n-1}^{(k)}$. Note that there are terms that are not computable in this expression. The term $\Delta\tilde{\varphi}_{n-1}^{(k)}$ is definitely not computable,

though when convergence in the adjoint computation is reached, this term vanishes. Nevertheless, in the context of finite number of iterations, we desire to numerically quantify this term.

Furthermore, the adjoint used to derive an error equation for the finite element solution of the operator decomposition is similar to the one described in Alg. 3, except that the Jacobian is linearized around $z^{(m)} = s\tilde{y}^{(m)} + (1-s)\widetilde{Y}^{(m)}$, with $s \in [0,1]$, such that $F(\tilde{y}^{(m)}) - F(\widetilde{Y}^{(m)}) = \overline{F'(z^m)}\,(\tilde{y}^{(m)} - \widetilde{Y}^{(m)})$. We denote by $\widetilde{\vartheta}$ the generalized Green's function for the finite element solution, which is written in compact form

$$-\dot{\widetilde{\vartheta}}^{(k)} = \overline{F'(z^{(m)})}^{\top}\widetilde{\vartheta}^{(k)} + \psi + \eta^{(k)}, \tag{11}$$

where

$$\eta^{(k)} = -\left[0, \quad \overline{F'_{12}(z^{(m)})}^{\top}(\widetilde{\vartheta}_1^{(k)} - \widetilde{\vartheta}_1^{(k-1)})\right]^{\top},$$

is the residual of the adjoint at iteration level k. To derive an error equation for the finite element solution of the operator decomposition, we let $\tilde{e}^{(m)} = \tilde{y}^{(m)} - \widetilde{Y}^{(m)}$. We multiply $\tilde{e}^{(m)}$ to (11), integrate on time interval $I_{l,n}$, $l = 1, 2, \cdots, L_{1,n}$, and use the identity

$$-\dot{\tilde{e}}^{(m)} + \overline{F'(z^{(m)})}\tilde{e}^{(m)} = -\delta_{\tilde{y}}^{(m)} + \dot{\widetilde{Y}}^{(m)} - F(\widetilde{Y}^{(m)}) = -\delta_{\tilde{y}}^{(m)} + R^{(m)},$$

to get

$$\begin{aligned}\int_{I_{l,n}} \left(\tilde{e}^{(m)}, \psi\right) dt = \left(\tilde{e}_{l-1,n}^{(m)}, \widetilde{\vartheta}_{l-1,n}^{(k)}\right) - \left(\tilde{e}_{l,n}^{(m)}, \widetilde{\vartheta}_{l,n}^{(k)}\right) - \int_{I_{l,n}} \left(R^{(m)}, \widetilde{\vartheta}^{(k)}\right) dt \\ + \int_{I_{l,n}} (\delta_{\tilde{y}}^{(m)}, \widetilde{\vartheta}^{(k)})\, dt - \int_{I_{l,n}} (\tilde{e}^{(m)}, \eta^{(k)})\, dt.\end{aligned} \tag{12}$$

This is the basis for the equation for the average error at time level I_n. The equation reflects the error arising from the consistent finite element numerical discretization of the analytical iterative operator decomposition. Notice that the last term is not computable since it contains the error $\tilde{e}^{(m)}$ weighted by the iteration residual in the adjoint computation. Again provided that an a priori estimate on $\tilde{e}^{(m)}$ is available, we consider this term higher order due the fact that the residual can be made as small as needed when the adjoint computation is driven to convergence (see [6] for details). Now we may write a computable error estimator for the finite element multiscale iterative operator decomposition method.

Theorem 1. *The computable average error of finite element multiscale iterative operator decomposition is*

$$\int_0^T (y - \widetilde{Y}^{(M_N)}, \psi) \approx \sum_{n=1}^{N} \sum_{l=1}^{L_{1,n}} \left(Q_{1,l,n}^{(M_N)} + Q_{2,l,n}^{(M_N)} + Q_{3,l,n}^{(M_N)} + Q_{4,l,n}^{(M_N)} \right) \tag{13}$$
$$= Q_1 + Q_2 + Q_3 + Q_4.$$

where

$$Q_{1,n}^{(m)} = -\int_{I_{l,n}} \left(R_1^{(m)}, \widetilde{\vartheta}_1^{(K_n)} \right) dt \text{ and } Q_{2,n}^{(m)} = -\int_{I_{l,n}} \left(R_2^{(m)}, \widetilde{\vartheta}_2^{(K_n)} \right) dt$$
$$Q_{3,n}^{(m)} = \int_{I_{l,n}} \left(\delta_{\widetilde{Y}}^{(M_n)}, \widetilde{\vartheta}^{(K_n)} \right) dt \text{ and } Q_{4,n}^{(m)} = \int_{I_{l,n}} \left(\delta_{\tilde{y}}^{(M_n)}, \widetilde{\vartheta}^{(K_n)} - \tilde{\varphi}^{(K_n)} \right) dt,$$

and $R_1^{(m)} = \dot{\widetilde{Y}}_1^{(m)} - F_1(\widetilde{Y}_1^{(m)}, \widetilde{Y}_2^{(m-1)})$, *and* $R_2^{(m)} = \dot{\widetilde{Y}}_2^{(m)} - F_2(\widetilde{Y}_1^{(m)}, \widetilde{Y}_2^{(m)})$.

Theorem 1 has decomposed the average error over $(0, T)$ into several components. The term $Q_{1,n}$ represents the finite element residual associated with the fast time scale subsystem, while $Q_{2,n}$ represents the finite element residual associated with the slow time scale. The term $Q_{3,n}$ represents the iteration error quantified by the iteration residual $\delta_{\widetilde{Y}}^{(M_n)}$. Recall that the adjoints $\widetilde{\vartheta}$ and $\tilde{\varphi}$ differ in the functions which are used for linearization. Thus, the term $Q_{4,n}$ also vanishes when $\widetilde{\vartheta} = \tilde{\varphi}$, which may be true if, for example, the system (1) is linearly coupled, i.e., if $F_i(y_1, y_2) = A_{i1} y_1 + A_{i2} y_2$ for some matrix A_{i1} and A_{i2}.

3 A Numerical Experiment

We illustrate the robustness of the proposed error estimator by solving a 3×3 system

$$\begin{cases} \dot{x} &= 100y + z, \quad x(0) = \dfrac{9001}{10001} \\ \dot{y} &= -100x, \quad y(0) = -\dfrac{10^5}{10001} \\ \dot{z} &= -z + y, \quad z(0) = 1000. \end{cases} \tag{14}$$

There are two distinct time scales, fast $\mathcal{O}(2\pi/100)$ and slow $\mathcal{O}(1)$. We set $y_1 = [x\ y]^\top$ (associated with the fast time scale) and $y_2 = z$ (associated with the slow time scale). A typical solution is depicted in Fig. 1.The forward problem is solved using the second order, piecewise linear and continuous cG method, which is equivalent to Crank–Nicholson scheme. We consider a quantity of interest

$$\int_0^2 z(t)\, dt = 858.7488805 \tag{15}$$

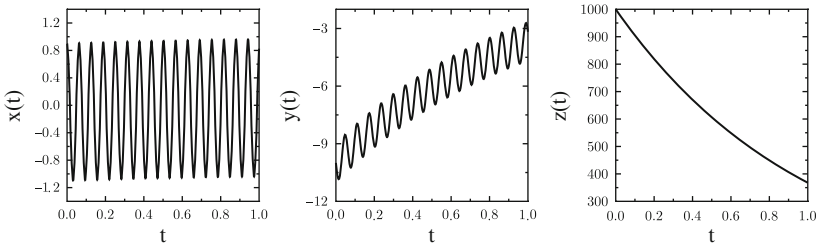


Fig. 1 Typical solution of (14)

Table 1 Error estimate for quantity of interest (15)

Δs_2	Exact error	Q_1	Q_2	Eff. index
2.00	−136.2729668	-0.0001734	−145.2920051	1.066
1.00	−24.2340044	-0.0001769	−24.6426795	1.017
0.50	−5.7329837	-0.0001759	−5.7570522	1.004
0.25	−1.4129595	-0.0001756	−1.4142761	1.001

We solve (14) until $T = 2$ using $\Delta t = 2$, $\Delta s_1 = 5 \times 10^{-3}$ and varying Δs_2. The result is shown in Table 1. The problem is solved until the residual is very small so that $Q_3 = 0$. Notice also since the problem is linearly coupled we have $Q_4 = 0$. This means that the only error contributions come from Q_1 and Q_2. Eventhough the error is still considerably large, the prediction of the proposed error estimate is very good as can be seen from the effectivity index. Notice also that the dominant contribution is from the slow scale component Q_2. This example shows the potential for using the above estimate to adaptively determine the parameters controlling accuracy. Since, the error estimate is written as a sum of contributing components, namely the finite element residual associated with the fast scale variables (Q_1), the finite element residual associated with the slow scale variables (Q_2), and the iteration residuals (Q_3 and Q_4), we can determine the largest source of error and adjust the corresponding parameter.

References

1. Biesiadecki, J.J., Skeel, R.: Dangers of multiple time step methods. J. Comp. Phys. **109**, 318–328 (1993)
2. Chen, J., Crow, M., A variable partitioning strategy for the multirate method in power systems. IEEE Trans. Power Syst. **23**, 259–266 (2008)
3. Constantinescu, E., Sandu, A.: Multirate timestepping methods for hyperbolic conservation laws. Report TR-06-15 (2006)
4. Estep, D., Larson, M.G., Williams, R.D.: Estimating the error of numerical solutions of systems of reaction-diffusion equations. Mem. Amer. Math. Soc. **146**, viii+109 (2000)
5. Estep, D., Ginting, V., Ropp, D., Shadid, J., Tavener, S.: An a posteriori-a priori analysis of multiscale operator splitting. SIAM J. Num. Anal. **46**, 1116–1146 (2008)

6. Estep, D., Ginting, V., Tavener, S.: A posteriori analysis of a multirate numerical method for ordinary differential equations. submitted to SIAM J. Num. Anal. (2008)
7. Gear, C.W.: Multirate methods for ordinary differential equations. Tech. Rep. UIUCDCS-74-880, Dept. of Computer Science, University of Illinois, Urbana-Champaign (1974)
8. Logg, A.: Multi-adaptive galerkin methods for odes I. SIAM J. Sci. Comput. **24**, 1879–1902 (2003)
9. Logg, A.: Multi-adaptive Galerkin methods for ODEs. II. Implementation and applications. SIAM J. Sci. Comput. **25**, 1119–1141 (2003)
10. Logg, A.: Multiadaptive Galerkin methods for ODEs. III. A priori error estimates. SIAM J. Numer. Anal. **43**, 2624–2646 (2006)
11. Verhoeven, A., Tasic, B., Beelen, T.G.J., ter Maten, E.J.W., Mattheij, R.M.M.: Automatic partitioning for multirate methods, In: Scientific Computing in Electrical Engineering, 11, Springer, Berlin (2007)

Goal-Oriented Error Estimation for the Discontinuous Galerkin Method Applied to the Biharmonic Equation

João L. Gonçalves, Philippe R.B. Devloo, and Sônia M. Gomes

Abstract A posteriori goal-oriented error estimation for approximation of discontinuous Galerkin finite element method is considered for the biharmonic equation. The methodology is based on the dual problem associated to the target functional. Using our estimation, we design two error indicators in order to ensure an efficient error control of the prescribed functional. We present numerical experiments to illustrate the performance of the error indicators.

1 Introduction

Error control of approximations is a crucial aspect in numerical resolution of partial differential equations. The most commonly used error control algorithms use error estimators related to the norm or seminorm associated with the differential equation. In many cases one is interested in controlling the error related to functionals applied to the solution; this approach is called goal-oriented error control. Pioneering work in this area has been developed by Babuška and collaborators [1, 2] .

Our group has developed research in the area of discontinuos Galerkin finite element methods (DGFEM). In the present we develop an a posteriori goal-oriented error estimator and two error indicators to approximations obtained by symmetric DGFEM applied to the biharmonic equation.

In [4], Harrimann et al. presents an a posteriori goal-oriented error estimator of second order equations. Our purpose is to extend this kind of estimator to the biharmonic equation. The resulting indicator is globally efficient, but not locally efficient. Modifying this indicator by balancing the interior edge error contributions, we

J.L. Gonçalves (✉) and S.M. Gomes
Unicamp – IMECC, 651 Sérgio Buarque de Holanda Street, Campinas, SP Brazil
e-mail: jluis@ime.unicamp.br, soniag@ime.unicamp.br

P.R.B. Devloo
Unicamp – FEC, 951 Albert Einstein Avenue, Campinas, SP Brazil
e-mail: phil@fec.unicamp.br

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_39,

obtain another error indicator that is globally and locally efficient. The performance of these indicators is illustrated in the numerical results.

2 Notation and Finite Element Spaces

Let Ω be an open bounded polygonal domain in $\mathbb{R}^2$ whith Lipschitz boundary; by $\partial\Omega$ we denote the union of the open edges of Ω. Let $\{\mathscr{T}_h\}_{h>0}$ be a family of partitions of Ω into pairwise disjoint open convex elements k, such that

$$\overline{\Omega} = \bigcup_{k\in\mathscr{T}_h} \overline{k}.$$

For a fixed master-element $\widehat{k} \subset \mathbb{R}^2$, we assume that each $k \in \mathscr{T}_h$ is the image by an affine function of $\widehat{k}$, i.e., $k = F_k(\widehat{k})$. Let $h_{\mathscr{T}_h}(x) = h_k = \text{diam}(k),\ x \in k$ be a piecewise constant function, being h the maximum of h_k, $k \in \mathscr{T}_h$. The set of all open edges e of all elements $k \in \mathscr{T}_h$ is denoted by ε. Define also the piecewise constant function $h_\varepsilon(x) = h_e = \text{diam}(e)$, $x \in e$. We assume that $\{\mathscr{T}_h\}_{h>0}$ is a shape-regular family in the sense that there is a positive constant c, independent of h, such that $ch_k \le h_e \le h_k$ for all $k \in \cup_{h>0}\mathscr{T}_h$ and $e \in \partial k$.

For positive integers m we denote by $Q_m(\widehat{k})$ the linear space of tensor-product polynomials of degree $\le m$ in each co-ordinate direction restricted to $\widehat{k}$. The maximum degree in each element k is denoted by p_k. The local Sobolev indices are denoted by s_k. Grouping p_k, s_k and F_k for $k \in \mathscr{T}_h$ into the vectors $\mathbf{p} = (p_k)$, $\mathbf{s} = (s_k)$ and $\mathbf{F} = (F_k)$ respectively, we introduce the finite element space

$$S^{\mathbf{p}}(\Omega, \mathscr{T}_h, \mathbf{F}) := \left\{u \in L^2(\Omega) : u|_k \circ F_k \in Q_{p_k}(\widehat{k})\right\}, \tag{1}$$

and the broken Sobolev space of composite index $\mathbf{s}$,

$$H^{\mathbf{s}}(\Omega, \mathscr{T}_h) := \left\{u \in L^2(\Omega) : u|_k \in H^{s_k}(k)\ \forall k \in \mathscr{T}_h\right\}. \tag{2}$$

The set ε_{int} denote all open interior edges, i.e., $\varepsilon_{int} = \{e \in \varepsilon : e \subset \Omega\}$, and the set ε_∂ denote all open boundary edges, i.e., $\varepsilon_\partial = \{e \in \varepsilon : e \subset \partial\Omega\}$. Moreover, we define $\Gamma^0 = \{x \in \Omega : x \in e \text{ for } e \in \varepsilon_{int}\}$ and $\Gamma = \Gamma^0 \cup \partial\Omega$. For $u, v \in L^2(\Gamma)$ the inner product is denoted by $\langle u, v\rangle_{L^2(\Gamma)}$ with corresponding norm $||u||_{L^2(\Gamma)}$. For each edge $e \in \varepsilon_{int}$ there are exactly two elements k_i and k_j, with $i > j$, such that $\overline{k_i} \cap \overline{k_j} = \overline{e}$.

The jump $[u]_e = u|_{\partial k_i \cap e} - u|_{\partial k_j \cap e}$ and the mean-value $\{u\}_e = 0.5(u|_{\partial k_i \cap e} + u|_{\partial k_j \cap e})$ over an edge $e \in \varepsilon_{int}$ between the elements k_i and k_j are defined for functions $u \in H^{\mathbf{s}}(\Omega, \mathscr{T}_h)$, such that $s_k > 1/2$ for all k. These definitions are extended to $e \in \partial\Omega$ by $[u]_e = u|_e$ and $\{u\}_e = u|_e$. To each edge $e \in \varepsilon_{int}$ we associate the unit vector $\mathbf{n}$, normal to e from k_i to k_j and for each $e \in \varepsilon_\partial \cap \varepsilon$, $\mathbf{n}$

denotes the outward unit normal vector to $\partial\Omega$. For each $k \in \mathscr{T}_h$, we also consider $\mathbf{n}_k$ the outward unit vector to ∂k.

3 Model Problem and the DGFEM

The boundary value problem for the biharmonic equation subject to Dirichlet boundary conditions is: find $u \in H^4(\Omega)$ such that

$$\Delta^2 u = f \text{ in } \Omega, \tag{3}$$
$$u = g_0 \text{ in } \partial\Omega, \tag{4}$$
$$\nabla u \cdot \mathbf{n} = g_1 \text{ in } \partial\Omega, \tag{5}$$

where $\Delta^2 u = \Delta(\Delta(u))$, $f \in L^2(\Omega)$, g_0 and $g_1 \in L^2(\partial\Omega)$.

Consider the following broken weak formulation of the boundary value problem (3–5): find $u \in H^4(\Omega, \mathscr{T}_h)$ such that,

$$B(u, v) = l(v) \ \ \forall\, v \in H^4(\Omega, \mathscr{K}), \tag{6}$$

where

$$B(u, v) = B_{\mathscr{T}_h}(u, v) + J_1(u, v) + J_1(v, u) - J_2(u, v) - J_2(v, u) + B_s(u, v), \tag{7}$$

is the bilinear form composed by the following terms

$$B_{\mathscr{T}_h}(u, v) = \sum_{K \in \mathscr{K}} \langle \Delta u, \Delta v \rangle_{L^2(K)}, \tag{8}$$
$$J_1(u, v) = \langle \{\mathbf{n} \cdot \nabla(\Delta u)\}, [v] \rangle_{L^2(\Gamma)}, \tag{9}$$
$$J_2(u, v) = \langle \{\Delta u\}, [\mathbf{n} \cdot \nabla v] \rangle_{L^2(\Gamma)}, \tag{10}$$
$$B_s(u, v) = \langle \alpha[u], [v] \rangle_{L^2(\Gamma)} + \langle \beta[\mathbf{n} \cdot \nabla u], [\mathbf{n} \cdot \nabla v] \rangle_{L^2(\Gamma)}. \tag{11}$$

The terms $B_{\mathscr{T}_h}(u, v)$, $J_1(u, v)$ and $J_2(u, v)$ come from the integration by parts of $\int_k (\Delta^2 u) v dx$ on each element, techniques for the decomposition of numerical fluxes and the definitions of jump and mean-value. The terms $J_1(v, u)$ and $J_2(v, u)$ are introduced in order to obtain symmetry. In the estabilization term $B_s(\cdot, \cdot)$, α and β are called discontinuos penalty parameters and are described in [5].

The linear functional

$$l(v) = l_{\mathscr{T}_h}(v) + l_s(v),$$

is composed by

$$l_{\mathscr{T}_h}(v) = (f, v)_{L^2(\Omega)} + \langle g_0, \mathbf{n} \cdot \nabla(\Delta v) \rangle_{L^2(\partial\Omega)} - \langle g_1, \Delta v \rangle_{L^2(\partial\Omega)},$$
$$l_s(v) = \langle \alpha\ g_0, v \rangle_{L^2(\partial\Omega)} + \langle \beta\ g_1, \mathbf{n} \cdot \nabla v \rangle_{L^2(\partial\Omega)}.$$

The DGFEM for the formulation (6) is: find $u_h \in S^{\mathbf{p}}(\Omega, \mathscr{T}_h, \mathbf{F})$ such that,

$$B(u_h, v) = l(v) \ \ \forall\ v \in S^{\mathbf{p}}(\Omega, \mathscr{T}_h, \mathbf{F}). \tag{12}$$

For a stability analysis and a priori estimates for (12) we refer to [5], where one of the results is the Galerkin orthogonality

$$B(u - u_h, v) = 0 \ \ \forall\ u_h \text{ e } \ v \in S^{\mathbf{p}}(\Omega, \mathscr{T}_h, \mathbf{F}). \tag{13}$$

4 Goal-Oriented Error Estimation

Let $J(u)$ be a arbitrary linear functional of the solution of problem (6). Our goal is to estimate the aproximation error $J(u) - J(u_h)$ in this quantity of interest, where u_h is the approximation solution of (12). Based on the theory presented in [4] and [6], we define the following dual problem: find $w \in H^4(\Omega, \tau_h)$ such that

$$B(v, w) = J(v)\ , \ \ \forall\ v \in H^4(\Omega, \mathscr{T}_h). \tag{14}$$

We assume that (14) has a unique solution. For functionals of the form $J(u) = \int_\Omega \Psi u dx$, the symmetric DGFEM (12) is adjoint-consistent, see [5].

We define the interior and boundary residues:

$$R_{int}(u) = f - \Delta^2 u, \quad R_{D1}(u) = g_0 - u \quad \text{and} \quad R_{D2}(u) = g_1 - \mathbf{n} \cdot \nabla u.$$

Taking $v = u - u_h$ in (14), and using the linearity of J, we get

$$J(u) - J(u_h) = J(u - u_h) = B(u - u_h, w). \tag{15}$$

This relation between the functional J and the bilinear form B is the starting point for the error estimation. Considering $w \in H^4(\Omega, \mathscr{T}_h)$, using the linearity of B, and integrating by parts twice, we obtain the expansion

$$B(u - u_h, w) = \sum_k \eta_k(u_h, w), \tag{16}$$

where the element indicators $\eta_k(u_h, w)$ have the expression

$$\begin{aligned}
\eta_k(u_h, w) = {} & \langle R_{int}(u_h), w\rangle_{L^2(k)} + \langle R_{D1}(u_h), \nabla \Delta(w) \cdot \mathbf{n}\rangle_{L^2(\partial\Omega \cap \partial k)} \\
& + \langle \alpha R_{D1}(u_h), w\rangle_{L^2(\partial\Omega \cap \partial k)} - \langle R_{D2}(u_h), \Delta(w) \cdot \mathbf{n}\rangle_{L^2(\partial\Omega \cap \partial k)} \\
& + \langle \beta R_{D2}(u_h), \nabla(w) \cdot \mathbf{n}\rangle_{L^2(\partial\Omega \cap \partial k)} - \langle [\Delta u_h], 0.5(\nabla(w) \cdot \mathbf{n}_k)\rangle_{L^2(\Gamma^0 \cap \partial k)} \\
& + \langle 0.5(w), [\nabla \Delta u \cdot \mathbf{n}_k]\rangle_{L^2(\Gamma^0 \cap \partial k)} - \langle 0.5(\nabla \Delta w \cdot \mathbf{n}_k), [u_h]\rangle_{L^2(\Gamma^0 \cap \partial k)} \\
& + \langle 0.5(\Delta(w)), [\nabla u_h \cdot \mathbf{n}_k]\rangle_{L^2(\Gamma^0 \cap \partial k)}.
\end{aligned} \tag{17}$$

A similar procedure, without using the linearity of B, gives another expansion

$$B(u-u_h,w)=\sum_k \mu_k(u_h,w,u) \tag{18}$$

in terms of new element indicators

$$\begin{aligned}\mu_k(u_h,w,u) &= \langle R_{int}(u_h),w\rangle_{L^2(k)} - \langle \nabla\Delta(u-u_h)\cdot\mathbf{n}_k,w\rangle_{L^2(\partial k\cap\Gamma^0)}\\ &+ \langle \Delta(u-u_h),\nabla w\cdot\mathbf{n}_k\rangle_{L^2(\partial k\cap\Gamma^0)} + \langle \nabla\Delta w\cdot\mathbf{n}_k,u-u_h\rangle_{L^2(\partial k\cap\Gamma^0)}\\ &+ \langle \nabla\Delta w\cdot\mathbf{n},R_{D1}(u_h)\rangle_{L^2(\partial k\cap\partial\Omega)} - \langle \Delta w,\nabla(u-u_h)\cdot\mathbf{n}_k\rangle_{L^2(\partial k\cap\Gamma^0)}\\ &- \langle \Delta w,R_{D2}(u_h)\rangle_{L^2(\partial k\cap\partial\Omega)}.\end{aligned} \tag{19}$$

As a consequence of the equalities (15), (16) and (18), the following a posteriori goal-oriented error estimates can be proved similarly to those showed in [4], and will be presented in the PhD thesis [3], in preparation.

Theorem 1. *Let u be the solution of (6), u_h the solution of (12) and w the solution of (14). Then*

$$J(u)-J(u_h)=\sum_{k\in\tau_h}\eta_k(u_h,w) \tag{20}$$

Theorem 2. *Let u be the solution of (6), u_h the solution of (12) and w the solution of (14). Then*

$$J(u)-J(u_h)=\sum_{k\in\tau_h}\mu_k(u_h,w,u) \tag{21}$$

Corollary 1. *Under the assumptions of Theorem 1, the following a posteriori error bound holds:*

$$|J(u)-J(u_h)|\le\sum_{k\in\tau_h}|\eta_k(u_h,w)| \tag{22}$$

5 Numerical Results

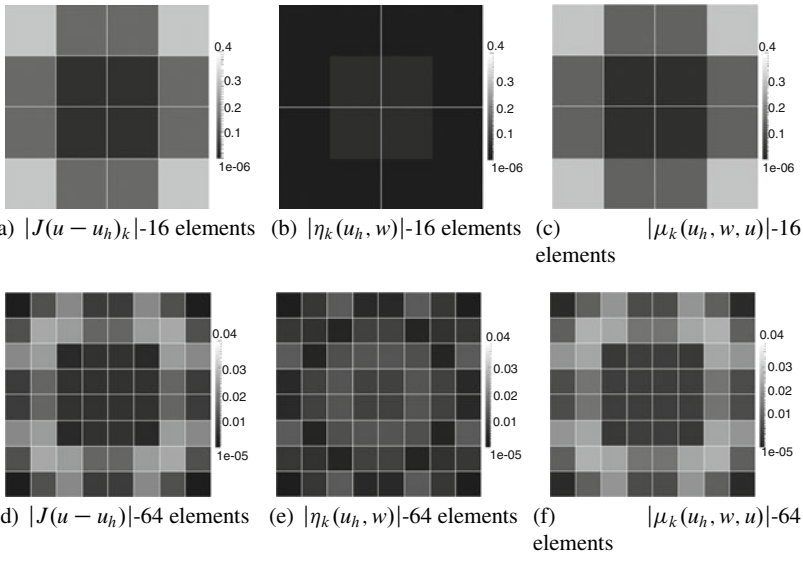
Let us consider an example where we know the primal and dual solutions, so that we can study the local and global efficiency of the proposed error indicators. On the domain $\Omega=(0,1)\times(0,1)$ we impose f and boundary conditions (4) and (5) so that the exact solution of (6) is $u(x,y)=\sin(\pi x)^2\sin(\pi y)^2$. Choosing the function $w(x,y)=\sin(2\pi x)^2\sin(2\pi y)^2$, it can be verifyied that it is the solution of the dual problem (14) associated to the target functional $J(u)=\int_\Omega \Delta^2(w)\,u\,dxdy$. The next results are for simulations using $p_k=3$ for all $k\in\mathcal{T}_h$.

The results presented in Table 1 show that both error indicators are globally efficient, confirming the equalities (20) and (21).

If one of these error indicators are supposed to be used to adapt the approximation space, we have to be sure that it is also efficient by element. Unlike $\eta_k(u_h,w)$, Fig. 1

Table 1 Globalcomparison of the error indicators with the true error

	16 elements	64 elements	256 elements
$\lvert J(u-u_h)\rvert$	0.147431014467342	0.009339820568385	0.000340218616789
$\lvert\sum \eta_k(u_h,w)\rvert$	0.147431014469882	0.009339820571323	0.000340218616006
$\lvert\sum \mu_k(u_h,w,u)\rvert$	0.147431014469875	0.009339820394009	0.000340219072805



(a) $|J(u-u_h)_k|$-16 elements (b) $|\eta_k(u_h,w)|$-16 elements (c) $|\mu_k(u_h,w,u)|$-16 elements

(d) $|J(u-u_h)|$-64 elements (e) $|\eta_k(u_h,w)|$-64 elements (f) $|\mu_k(u_h,w,u)|$-64 elements

Fig. 1 Comparison between $|J(u-u_h)_k|$ and the indicators $|\eta_k(u_h,w)|$ and $|\mu_k(u_h,w,u)|$

shows that the error indicator $\mu_k(u_h,w,u)$ is efficient by element, what makes it interesting for adaptation of the approximation space.

In practice, since the primal and dual solutions u and w are not know, we can not use the error indicators as stated. Instead, we propose to use $\mu_k(u_h,w_+,u_+)$ where w_+ and u_+ are approximations to w and u, respectively. Because of the Galerkin orthogonality (13), w_+ can not be in $S^{\mathbf{p}}(\Omega,\mathscr{T}_h,\mathbf{F})$. One option for w_+ and u_+ could be DGFEM approximations with increased polinomial degree, i.e., instead of p_k we use p_k+inc. To avoid the discontinuity of u_+ and w_+ on interior edges, we use the mean-values $\{u_+\}$ and $\{w_+\}$ instead.

Table 2 corresponds to the procedure of replacing w by $\{w_+\}$. It can be observed that for less enriched dual approximations $inc \leq 3$ there is a loss of global efficiency, but this effect disappears with increasing inc. The use of $u_+ = \{u_{h,p+inc}\}$ instead of u does not affect the equality (21), as can be seen in Table 3. The effect of replacing u by u_+ only contributes to a new arrangement of the error indicators $\mu_k(u_h,w,u)$ on the interior edges, without modification of the global error.

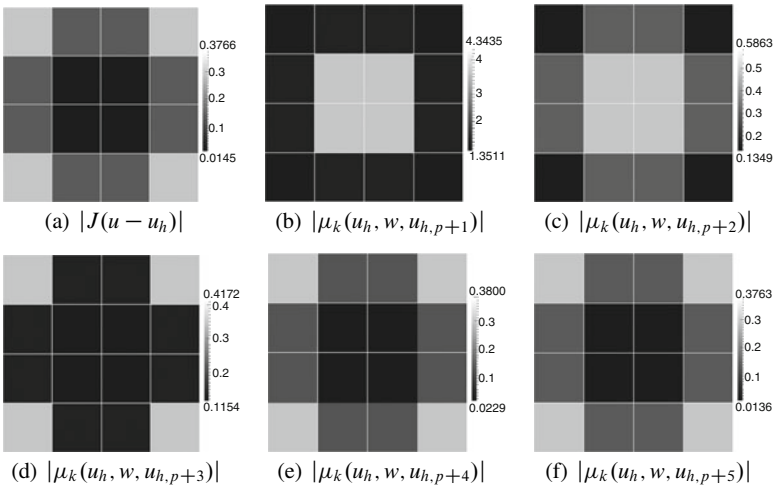
On the other hand, the efficiency by element of the error indicator is affected by this procedure, as we can observe in Fig. 2. Using 16 elements, and $inc \leq 3$ there are significant differences between the true error and the estimated error by element, but for higher order approximations of u_+, the agreement is perfect.

Table 2 Comparison of $|\sum \mu_k(u_h, \{w_{h,p+inc}\}, u)|$ with the true error

	16 elements	64 elements	256 elements
$J(u-u_h)$	0.147431014467	0.009339820568	0.000340218616
$inc = 1$	0.026080411968	0.007909482959	0.000327866783
$inc = 2$	0.162005604253	0.009774521521	0.000344525157
$inc = 3$	0.162635881593	0.009386956498	0.000340327529
$inc = 4$	0.146921660651	0.009334112842	0.000340210216
$inc = 5$	0.146917322016	0.009339421506	0.000340224326

Table 3 Comparison of $|\sum \mu_k(u_h, w, \{u_{h,p+inc}\})|$ with the true error

	16 elements	64 elements	256 elements
$J(u-u_{hp})$	0.147431014467	0.009339820568	0.000340218616
$inc = 1$	0.147431014469	0.009339821134	0.000340220187
$inc = 2$	0.147431014469	0.009339820131	0.000340220517
$inc = 3$	0.147431014469	0.009339821214	0.000340217549
$inc = 4$	0.147431014469	0.009339820029	0.000340216715
$inc = 5$	0.147431014469	0.009339820099	0.000340219336



(a) $|J(u-u_h)|$ (b) $|\mu_k(u_h, w, u_{h,p+1})|$ (c) $|\mu_k(u_h, w, u_{h,p+2})|$

(d) $|\mu_k(u_h, w, u_{h,p+3})|$ (e) $|\mu_k(u_h, w, u_{h,p+4})|$ (f) $|\mu_k(u_h, w, u_{h,p+5})|$

Fig. 2 Local comparison of $|\mu_k(u_h, w, u_+)|$ with the true error

6 Conclusions

We present two goal-oriented error indicators $\eta_k(u_h, w)$ and $\mu_k(u_h, w, u)$ of DGFEM approximations u_h of the biharmonic equation. The first one is a standart goal-oriented error indicator, which shows global but not local efficiency. The second indicator μ_k, obtained after a balance of the error distribution on the interior edges, results to be both global and local efficient. Therefore, the indicator μ_k

is more appropriate for adaptating the approximation spaces. In practice, the exact solutions u and w have to be replaced by enriched approximations u_+ and w_+. The present results, using enriched spaces with degree increments greater than 3 show perfect agreement with the true error.

Acknowledgements J. L. Gonçalves is grateful to FAPESP – Research Agency of the state of São Paulo, for the PhD grant (project 2007/00072-0). P. Devloo and S. Gomes thankfully acknowledges financial support from CNPq – the Brazilian Research Council.

References

1. Babuška, I., Miller, A.: The post-processing approach in the finite element method – part 1: Calculation of displacements, stress and other higher derivatives of the displacements. Int. J. Num. Meth. Eng. 20(6), 1085–1109 (1984)
2. Babuška, I., Miller, A.: The post-processing approach in the finite element method – part 2: Calculation of stress intesity factors. Int. J. Num. Meth. Eng. 20(6), 1111–1129 (1984)
3. Gonçalves, J. L.: *Indicadores de Erro a Posteriori para o Método de Galerkin Descontínuo para Aproximação de Funcionais de Soluções de Equações Elípticas de Quarta Ordem*. PhD thesis, in preparation, State University of Campinas, 2010
4. Harriman, K., Houston, P., Senior, B., Süli, E.: hp-Version Discontinuous Galerkin Methods with Interior Penalty for Partial Differential Equations with Nonnegative Characteristic Form. In C.-W. Shu, T. Tang, S.-Y. Cheng (Ed.). *Recent Advances in Scientific Computing and Partial Differential Equations*. Contemporary Mathematics, Vol. 330, pp. 89–119. AMS (2003)
5. Mozolevski, I., Süli, E., Bösing, P. R.: hp-version a priori error analysis of interior penalty discontinuos Galerkin finite element approximations to the biharmonic equation. J. Sci. Comput. 30(3), 465–491 (2007)
6. Oden, J. T., Prudhomme, S.: Goal-oriented error estimation and adaptivity for the finite element method. Comp. Math. Appl. 41, 735–756 (2001)

Solving Stochastic Collocation Systems with Algebraic Multigrid

Andrew D. Gordon and Catherine E. Powell

Abstract Stochastic collocation methods facilitate the numerical solution of PDEs with random data and give rise to large sequences of linear systems. For elliptic PDEs, algebraic multigrid (AMG) is a robust solver and considered individually, the systems are trivial to solve. The challenge lies in exploiting the systems' similarities to minimize the cost of solving the entire sequence. We propose an efficient solver that is more robust than other solution strategies in the literature. In particular, we show that it is feasible to use a finely-tuned AMG preconditioner for each system if key set-up information is reused. The method is robust with respect to variations in discretization and statistical parameters for stochastically linear and nonlinear data.

1 Introduction

Our starting point is the stochastic steady-state diffusion problem

$$\begin{aligned} -\nabla\cdot(a(\mathbf{x},\omega)\nabla p(\mathbf{x},\omega)) &= f(\mathbf{x}) && \text{in } D\times\Omega, \quad D\subset\mathbb{R}^2 \\ p(\mathbf{x},\omega) &= 0 && \text{on } \partial D\times\Omega, \end{aligned} \tag{1}$$

which arises when only limited information about a is available. Here Ω is a sample space from a probability space and the input a and solution p are second-order random fields. We assume the mean $\mu_a(\mathbf{x}) = \mathbb{E}[a(\mathbf{x},\omega)]$ and covariance function

$$C_a(\mathbf{x},\mathbf{y}) \;=\; \mathbb{E}\big[\big(a(\mathbf{x},\omega)-\mu_a(\mathbf{x})\big)\big(a(\mathbf{y},\omega)-\mu_a(\mathbf{y})\big)\big] = \sigma^2 V(\mathbf{x},\mathbf{y}) \tag{2}$$

are known and follow the well-established procedure ([2, 7]) of assuming that $a(\mathbf{x},\omega)$ is, or can be well approximated by, a function of M independent random variables $\xi_k(\omega)$. A common choice is a (truncated) Karhunen–Loève expansion [9]

A.D. Gordon and C.E. Powell
School of Mathematics, University of Manchester, Oxford Road, Manchester, M13 9PL, UK
e-mail: gordona@cs.man.ac.uk, c.powell@manchester.ac.uk

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_40,

$$a_M(\mathbf{x}, \_) = \mu_a(\mathbf{x}) + \sigma \sum_{k=1}^{M} \sqrt{\lambda_k} c_k(\mathbf{x}) \xi_k(\omega) \tag{3}$$

in terms of M uncorrelated random variables ξ_k, or the exponential thereof

$$a_M(\mathbf{x}, \boldsymbol{\xi}) = \exp\Big(\mu_a(\mathbf{x}) + \sigma \sum_{k=1}^{M} \sqrt{\lambda_k} c_k(\mathbf{x}) \xi_k(\omega)\Big). \tag{4}$$

Here, $(\lambda_k,\ c_k(\mathbf{x}))$ are the leading eigenpairs of the integral operator associated with $V(\mathbf{x}, \mathbf{y})$ in (2) and σ is the standard deviation of a. Let $\xi_k(\Omega) = \Gamma_k \subseteq \mathbb{R}$ and denote the probability density function of ξ_k by ρ_k. If the random variables are independent, then the joint density function is $\rho(\boldsymbol{\xi}) = \prod_{k=1}^{M} \rho_k(\xi_k)$ where $\boldsymbol{\xi} \in \Gamma = \prod_{k=1}^{M} \Gamma_k \subseteq \mathbb{R}^M$.

Replacing $a(\mathbf{x}, \omega)$ by $a_M(\mathbf{x}, \boldsymbol{\xi})$ in (1) results in an $M + 2$ dimensional deterministic PDE and the corresponding weak problem – which has been well studied – consists in finding $p(\mathbf{x}, \boldsymbol{\xi}) \in V = L^2_\rho(\Gamma, H^1_0(D))$ such that

$$\mathbb{E}\Big[\int_D a_M(\mathbf{x}, \boldsymbol{\xi}) \nabla p(\mathbf{x}, \boldsymbol{\xi}) \cdot \nabla v(\mathbf{x}, \boldsymbol{\xi}) \mathrm{d}\mathbf{x}\Big] = \mathbb{E}\Big[\int_D f(\mathbf{x}) v(\mathbf{x}, \boldsymbol{\xi}) \mathrm{d}\mathbf{x}\Big] \quad \forall\, v \in V \tag{5}$$

where $\mathbb{E}[\cdot] = \int_\Gamma \rho(\boldsymbol{\xi}) \cdot d\boldsymbol{\xi}$. Stochastic finite element methods proceed by discretizing the physical domain D in the usual way, leading to the semi-discrete problem: find $p_h \in V_h = L^2_\rho(\Gamma, X_h)$ with $X_h \subset H^1_0(D)$ and $\dim(X_h) = n_h$ such that

$$\mathbb{E}\Big[\int_D a_M(\mathbf{x}, \boldsymbol{\xi}) \nabla p_h(\mathbf{x}, \boldsymbol{\xi}) \cdot \nabla v(\mathbf{x}, \boldsymbol{\xi}) \mathrm{d}\mathbf{x}\Big] = \mathbb{E}\Big[\int_D f(\mathbf{x}) v(\mathbf{x}, \boldsymbol{\xi}) \mathrm{d}\mathbf{x}\Big] \quad \forall v \in V_h. \tag{6}$$

We can tackle (5) and (6) with Monte Carlo methods (MCMs), stochastic Galerkin methods (SGMs) [2, 7] and stochastic collocation methods (SCMs) [1, 10, 15]. MCMs approximate $\mathbb{E}[p_h]$ by the sample average at randomly chosen points $\boldsymbol{\xi}_r \in \Gamma$. If $a^r_M(\mathbf{x}) = a_M(\mathbf{x}, \boldsymbol{\xi}_r)$ is strictly positive then each $p^r_h(\mathbf{x}) = p_h(\mathbf{x}, \boldsymbol{\xi}_r) \in X_h$ satisfies

$$\int_D a^r_M(\mathbf{x}) \nabla p^r_h(\mathbf{x}) \cdot \nabla v(\mathbf{x}) \mathrm{d}\mathbf{x} = \int_D f(\mathbf{x}) v(\mathbf{x}) \mathrm{d}\mathbf{x} \qquad \forall\, v \in X_h, \tag{7}$$

leading to a sequence of decoupled, symmetric positive definite linear systems

$$A_r \mathbf{p}_r = \mathbf{b}, \qquad r = 1, 2, \ldots, \qquad A_r \in \mathbb{R}^{n_h \times n_h}. \tag{8}$$

SGMs, which seek $p_{hd} \in X_h \otimes S_d$ with $S_d \subset L^2_\rho(\Gamma)$, have a superior convergence rate [2], for low values of M but result in one system of dimension $n_h \times \dim(S_d)$. If S_d consists of complete polynomials, the equations must be solved simultaneously. Only if tensor product polynomials are used and $a_M(\mathbf{x}, \boldsymbol{\xi})$ is linear in ξ_k, as in

(3), does S_d possess a basis that decouples the equations. SCMs sample p_h and so handle (3) and (4) with equal ease. However, they converge as rapidly as SGMs.

The conjugate gradient method (CG) is an optimal solver for individual systems in (8) and algebraic multigrid (AMG, [13]) is a widely-used preconditioner for discretized elliptic PDEs that is highly robust with respect to variations in $a^r_M(\mathbf{x})$. When the number of systems is large, it may be infeasible to tune AMG to individual matrices and the one-preconditioner-fits-all approach has merit. That reduces set-up costs but the preconditioner may be so weak for some systems that no savings are made.

Jin et al. [8] and Ullmann [14] study the systems arising when (5) is discretized with a SGM. They only consider stochastically linear coefficients such as (3), which always lead to fairly well-conditioned matrices A_r. The systems are solved with recycled Krylov subspace methods [11], which are suboptimal for individual systems, but beneficial when applied to the whole sequence if a weak preconditioner is used. The domain decomposition preconditioner in [8] is optimal for most systems but weak for a certain subset. In [14], one V-cycle of AMG applied to the "mean" stiffness matrix (with diffusion coefficient $\mu_a(\mathbf{x})$) is used to precondition all systems. However, the efficiency deteriorates when σ in (2) is large relative to $\mu_a(\mathbf{x})$.

We focus on SCMs and investigate how computational savings can be made by recycling information between systems. The emphasis is on the reuse of preconditioner information and we propose an efficient way to solve the sequence (8) with CG using AMG preconditioning. Our method can handle (3) and (4) equally well, and is robust with respect to variations in all the discretization and statistical parameters.

2 Stochastic Collocation Methods

SCMs are derived by collocating (6) on a set of points $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_{n_c}$ in Γ. A global approximation p_{hd} is then obtained by performing Lagrange interpolation. That is,

$$p_{hd}(\mathbf{x}, \boldsymbol{\xi}) = \sum_{r=1}^{n_c} p^r_h(\mathbf{x}) L_r(\boldsymbol{\xi}), \tag{9}$$

where each $p^r_h(\mathbf{x}) = p_h(\mathbf{x}, \boldsymbol{\xi}_r) \in X_h$ satisfies (6) at $\boldsymbol{\xi}_r \in \Gamma$ and $L_r(\boldsymbol{\xi})$ is a multivariate Lagrange polynomial. By construction, $p_{hd}(\mathbf{x}, \boldsymbol{\xi}) \in X_h \otimes S_d$ where $S_d = \text{span}\{L_1(\boldsymbol{\xi}), \ldots, L_{n_c}(\boldsymbol{\xi})\} \subset L^2_\rho(\Gamma)$ and $\dim(S_d) = n_c$. Full tensor SCMs [1, 15] use Cartesian products of interpolation points on each Γ_k. Possibilities include Clenshaw–Curtis points and Gauss points. If $d_k + 1$ points are selected on Γ_k then $n_c = \prod_{k=1}^{M}(d_k + 1)$ which quickly becomes intractable as M increases.

Sparse grid SCMs [10, 12, 15] are based on interpolation rules for high-dimensional problems. Let Z_i be a set of points on Γ_i of size $m_i + 1$ where $m_0 = 1$

and $m_i = 2^{i-1}$ for $i \in \mathbb{N}$. For a given approximation level l, the sparse grid on Γ is then defined via

$$H(l, M) = \bigcup_{l \le \|\mathbf{i}\|_1 < l+M} Z_{i_1} \times \ldots \times Z_{i_M}, \qquad \mathbf{i} = (i_1, \ldots, i_M) \in \mathbb{N}^M.$$

The error incurred by approximating $p_h(\mathbf{x}, \boldsymbol{\xi})$ by $p_{hd}(\mathbf{x}, \boldsymbol{\xi})$ is due to interpolation. If d denotes the largest value for which polynomials of *total* degree d are interpolated exactly in (9) then sparse grid methods achieve total degree d accuracy with $l = d + 1$ [3] using far fewer points than full tensor methods.

3 Linear Systems

Each $p_h^r(\mathbf{x})$ in (9) solves (7) where $a_M^r(\mathbf{x})$ is (3) or (4) sampled at $\boldsymbol{\xi}_r$. If $X_h = \text{span}\{\phi_1(\mathbf{x}), \ldots, \phi_{n_h}(\mathbf{x})\}$ consists of piecewise polynomials then we have n_c *sparse* linear systems (8) where

$$\begin{aligned} [A_r]_{ij} &= \left(a_M^r(\mathbf{x}) \nabla \phi_i(\mathbf{x}),\, \nabla \phi_j(\mathbf{x})\right), \\ [\mathbf{b}]_i &= \left(f(\mathbf{x}),\, \phi_i(\mathbf{x})\right), \quad i, j = 1, \ldots, n_h. \end{aligned} \tag{10}$$

We assume for each r that $a_M(\mathbf{x}, \boldsymbol{\xi}_r)$ is strictly positive and bounded. That is,

$$0 < a_{1,r} \le a_M(\mathbf{x}, \boldsymbol{\xi}_r) \le a_{2,r} < \infty \quad \text{a.e. in } D \tag{11}$$

and so $\kappa(A_r) \lesssim a_{2,r} a_{1,r}^{-1} h^{-2}$. Note that $a_M(\mathbf{x}, \boldsymbol{\xi}_r)$ is not strictly positive for (3), if unbounded random variables are used. If piecewise linear polynomials are used for X_h and (11) holds, then each A_r is an M-matrix.

Definition 1 (M-matrix). An M-matrix is a symmetric positive definite matrix with positive diagonal entries and non-positive off-diagonal entries.

Theorem 1 gives insight into how ill-conditioned each A_r is with respect to the statistical parameters. We assume $\mu_a(\mathbf{x}) = \mu > 0$ and that the finite element meshes are shape regular and quasi-uniform. As usual, the largest edge length is denoted h.

Theorem 1. *The eigenvalues of A_r lie in $\left[ch^2(\mu - T_r), C(\mu + T_r)\right]$ for (3) and in $\left[ch^2 e^{\mu - T_r}, C e^{\mu + T_r}\right]$ for (4) where $c, C > 0$ are independent of h and $a_M^r(\mathbf{x})$ and*

$$T_r = \sigma\, S_M\, ||\boldsymbol{\xi}_r||_\infty, \qquad S_M = \sum_{k=1}^{M} \sqrt{\lambda_k}\, ||c_k||_{L^\infty(D)}.$$

Proof. Let $\mathbf{u} \in \mathbb{R}^{n_h} \setminus \{\mathbf{0}\}$ and define $v(\mathbf{x}) = \sum_{j=1}^{n_h} u_j \phi_j(\mathbf{x}) \in X_h$. Define the stiffness matrix A_0 via $[A_0]_{ij} = (\nabla \phi_i, \nabla \phi_j)$ and recall the standard result

$ch^2 \leq \frac{\mathbf{u}^T A_0 \mathbf{u}}{\mathbf{u}^T \mathbf{u}} \leq C$ (e.g., see [5]). If $a_M(\mathbf{x}, \boldsymbol{\xi})$ is defined as in (3),

$$|\mathbf{u}^T A_r \mathbf{u} - \mu \mathbf{u}^T A_0 \mathbf{u}| = \left| \int_D \left(\sigma \sum_{k=1}^{M} \sqrt{\lambda_k}\, c_k(\mathbf{x})\, [\boldsymbol{\xi}_r]_k \right) \nabla v(\mathbf{x})\, \nabla v(\mathbf{x}) \mathrm{d}\mathbf{x} \right|.$$

Hence $(\mu - T_r)\mathbf{u}^T A_0 \mathbf{u} \leq \mathbf{u}^T A_r \mathbf{u} \leq (\mu + T_r)\mathbf{u}^T A_0 \mathbf{u}$. Combining with the bound for the eigenvalues of A_0 gives the first result. The bound for (4) is similarly obtained.

As $M \to \infty$, S_M converges [6], at a rate that depends on $C_a(\mathbf{x}, \mathbf{y})$. Note that the bound is different for each system. T_r depends on σ and $||\boldsymbol{\xi}_r||_\infty$ which depends on d if unbounded random variables are used.

4 AMG Preconditioning

AMG is an iterative solver that combines smoothing and coarse grid correction. "Grids" are index sets of unknowns; no geometric information is needed. Consider the linear system $A^1 \mathbf{u} = \mathbf{v}$ with $\mathbf{u} \in \mathbb{R}^{n_1}$. Before iteration can begin, there is a set-up phase, during which the following information is generated.

1. Sequence of grids: $C_l \subset C_{l-1} \subset \cdots \subset C_2 \subset C_1 = \{1, \ldots, n_1\}$ with $|C_k| = n_k$.
2. Prolongation matrices: $P_k^{k-1} \in \mathbb{R}^{n_{k-1} \times n_k}$ for $k = 2, \ldots, l$.
3. Coarse grid matrices: $A^k = R_{k-1}^k A^{k-1} P_k^{k-1}$, $R_{k-1}^k = (P_k^{k-1})^T$ for $k = 2, \ldots, l$.

Coarse grids and prolongation matrices are constructed by exploiting algebraic information in A^1 and this results in a finely-tuned preconditioner for the given matrix.

Definition 2 (Optimal preconditioner). An optimal preconditioner for A_r is a matrix P_r, for which the action of P_r^{-1} can be computed in $O(n_h)$ work and the eigenvalues of $P_r^{-1} A_r$ are contained in $[\theta_r, \Theta_r]$ with $\theta_r, \Theta_r > 0$ independent of h.

For each A_r defined in (10) we employ one step of AMG as a preconditioner for CG. For M-matrices, this is a good strategy. Specifically, if P_r is the matrix for which $P_r^{-1}\mathbf{v}$ denotes the application of one AMG V-cycle to $A_r \mathbf{u} = \mathbf{v}$, with set-up information generated using A_r, then P_r is expected to be optimal. Rigorous convergence proofs are lacking for AMG but if the M-matrix property is not strongly violated, we expect θ_r and Θ_r to be quite insensitive to h and $a_M^r(\mathbf{x})$. If $a_M^r(\mathbf{x})$ is oscillatory, which can occur if $C_a(\mathbf{x}, \mathbf{y})$ has a small correlation length, then we expect some degradation [13]. However, MCMs are more appropriate than SCMs in that case.

The disadvantage of this *finely-tuned* AMG preconditioning strategy is that set-up information is required for n_c distinct matrices. Alternatively, we can employ one generic preconditioner. To this end, let A_μ be the stiffness matrix with coefficient $a_M(\mathbf{x}, \mathbf{0})$, and let P_μ be the matrix for which $P_\mu^{-1}\mathbf{v}$ denotes the application of one AMG V-cycle to $A_\mu \mathbf{u} = \mathbf{v}$, with set-up information generated from A_μ. Theorem 2 summarizes how the efficiency of P_μ varies from system to system.

Theorem 2. *Let $\mu_a(\mathbf{x}) = \mu > 0$. The eigenvalues of $P_\mu^{-1} A_r$ lie in $[c\ \theta_\mu e^{-T_r}, C\ \Theta_\mu e^{T_r}]$ for (4) and in $[c\ \theta_\mu(1-T_r\mu^{-1}), C\ \Theta_\mu(1+T_r\mu^{-1})]$ for (3), where c, $C > 0$ are independent of h and $a_M^r(\mathbf{x})$ and the eigenvalues of $P_\mu^{-1} A_\mu$ are contained in $[\theta_\mu, \Theta_\mu]$.*

We refer to this strategy as *mean-based* preconditioning. It is adequate for (3) as $T_r\mu^{-1}$ must be small for a well-posed problem. For (4), T_r can be arbitrarily large and in that case the bound is not a good one.

The costliest part of AMG set-up is the grid construction. Coarse grids must capture error not eliminated by smoothing and for M-matrices, such error varies slowly in the direction of strong dependence. If $a_M^r(\mathbf{x})$ is isotropic, strongly influencing points for A_μ are likely to be strongly influencing points for A_r, suggesting that coarse grids, and prolongation matrices can be computed once, and recycled.

Definition 3 (Strong influence). For an M-matrix A, the jth unknown strongly influences the ith unknown if for a given threshold $\alpha > 0$, $|A_{ij}| \geq \alpha \max_{k \neq i} |A_{ik}|$.

Formally, then, let $P_{\mu,r}$ be the matrix for which $P_{\mu,r}^{-1}\mathbf{v}$ denotes the application of one AMG V-cycle to $A_r\mathbf{u} = \mathbf{v}$, with coarse grids and prolongation matrices generated using A_μ. The coarse grid matrices should be computed using A_r and so the preconditioner is distinct for each system. We refer to this strategy as AMG preconditioning with *recycled setup*. As computing coarse grid matrices is relatively cheap, this strategy has set-up costs similar to mean-based AMG preconditioning. However, if A_r is an M-matrix we expect the eigenvalues of $P_{\mu,r}^{-1} A_r$ to lie in $[\theta_r^\mu, \Theta_r^\mu]$ with constants similar to θ_r and Θ_r obtained with the finely-tuned preconditioner P_r.

5 Numerical Results

Consider (1) on $D = (-1, 1) \times (-1, 1)$ with $f(\mathbf{x}) = 1$, $\mu_a(\mathbf{x}) = 1$ and covariance

$$C_a(\mathbf{x}, \mathbf{y}) = \sigma^2 \exp\big(-||\mathbf{x} - \mathbf{y}||_1\big).$$

We perform experiments with piecewise linear polynomials for (3) and (4) with $M = 6$. First, we apply a sparse grid SCM with Clenshaw–Curtis (CC) points. Next, we apply the full tensor SCM with Gauss points. We solve all systems using preconditioned CG with the zero vector as an initial guess. Computations are performed in serial on a dual-core laptop PC with 4GB of RAM using a MATLAB

Table 1 Average CG iterations for sparse grid SCM, uniform random variables, $l = 3$ and $n_c = 85$

Preconditioning strategy	h	Linear problem (3)			Nonlinear problem (4)		
		$\sigma^2 = 0.1$	0.2	0.27	$\sigma^2 = 1$	5	10
Finely-tuned	1/32	6.00	6.00	6.00	6.00	6.13	6.25
	1/128	6.01	6.15	6.32	6.60	6.91	7.01
Mean-based	1/32	8.16	9.38	10.44	13.96	32.64	61.14
	1/128	8.46	9.91	11.01	14.69	34.92	66.61
Recycled setup	1/32	6.00	6.00	6.00	6.02	6.42	6.69
	1/128	6.02	6.14	6.16	6.41	6.87	7.45

Table 2 AMG setup times, maximum CG iterations and total iteration times in seconds (in parentheses) for sparse grid SCM, uniform random variables, with $h = 1/128$, $l = 3$ and $n_c = 85$

Preconditioning strategy	AMG setup	Linear problem (3)				Nonlinear problem (4)			
		$\sigma^2 = 0.1$		$\sigma^2 = 0.27$		$\sigma^2 = 1$		$\sigma^2 = 10$	
Finely-tuned	3,212	7	(67)	7	(73)	7	(79)	8	(80)
Mean-based	41	12	(94)	25	(121)	23	(168)	224	(740)
Recycled setup	48	7	(67)	7	(68)	8	(76)	12	(84)

version of the AMG code [4]. The multigrid method is applied as a black-box with one pre and post Gauss–Seidel smoothing step. The stopping tolerance for CG is 10^{-6}.

Table 1 shows that the fine-tuned and recycled setup strategies are optimal with respect to variations in h and σ, for both (3) and (4). By recycling setup information, however, we obtain a finely-tuned preconditioner for each system at a fraction of the cost (see Table 2). Note that the exact benefits in terms of time depend on the coding environment. No systems arise which cannot be solved in an acceptably low number of iterations. There are considerable savings over mean-based preconditioning, whose performance, as Theorem 2 predicts, deteriorates as σ^2 increases. For (3), $\sigma^2 = 0.27$ is the largest value for which all subproblems are well-posed.

When uniform random variables are used, the collocation points lie in $\Gamma = [-\sqrt{3}, \sqrt{3}]^M$. Using unbounded Gaussian variables, which is permitted for (4), is more difficult. If the SCM uses d Gauss points in each dimension, the points are contained in $\Gamma = [-C_d, C_d]^M$ with $C_d = O(\sqrt{d})$. The results in Table 3 reveal how inefficient mean-based preconditioning then becomes with increasing d. AMG with recycled setup performs like finely-tuned AMG and is almost insensitive to d.

Additional savings can be made by reusing solutions to previous systems as initial guesses. For this, the collocation points need to be ordered so that successive samples of $a_M(\mathbf{x}, \boldsymbol{\xi})$ are as close as possible. We report only brief results. For full tensor SCMs, we ensure that successive points differ in only one component. Results are reported for a test problem in Table 4. The greatest savings occur for (3) when σ^2 is small and d is large. Indeed, samples of $a_M(\mathbf{x}, \boldsymbol{\xi})$ are all close to $\mu_a(\mathbf{x})$ and increasing d reduces the distance between distinct points. For (4), the benefits are negligible, even for large d. For sparse grid SCMs, after solving $A_r \mathbf{p}_r = \mathbf{b}$, we find the closest remaining point $\boldsymbol{\xi}_s$ to $\boldsymbol{\xi}_r$ with respect to the

Table 3 Average CG iterations for nonlinear problem using full tensor SCM with $h = 1/32$

Preconditioning strategy	σ^2	Uniform variables			Gaussian variables		
		$d = 2$	3	4	$d = 2$	3	4
		$n_c = 729$	4,096	15,625	$n_c = 729$	4,096	15,625
Finely-tuned	1	6.00	6.01	6.01	6.04	6.12	6.17
	10	6.41	6.40	6.44	6.53	6.61	6.68
Mean-based	1	17.33	17.99	18.27	22.48	29.12	35.97
	10	116.35	130.89	137.89	276.89	738.70	1777.32
Recycled setup	1	6.05	6.07	6.08	6.16	6.30	6.41
	10	7.04	7.10	7.13	7.38	7.65	7.83

Table 4 Average and range of CG iterations, AMG + recycled-setup, $h = 1/32$, uniform variables

		Linear problem (3)		Nonlinear problem (4)	
		$\sigma^2 = 0.01$	$\sigma^2 = 0.08$	$\sigma^2 = 1$	$\sigma^2 = 10$
Full tensor ($d = 2, n_c = 729$)	No sorting	6.00 [6, 6]	6.00 [6, 6]	6.05 [6, 7]	7.10 [6, 10]
	Sorting	4.24 [4, 6]	5.00 [5, 6]	5.55 [5, 7]	6.99 [6, 10]
Full tensor ($d = 4, n_c = 15,625$)	No sorting	6.00 [6, 6]	6.00 [6, 6]	6.08 [6, 7]	7.13 [6, 10]
	Sorting	4.13 [4, 6]	4.62 [4, 6]	5.28 [4, 7]	6.80 [5, 9]
Sparse grid ($l = 3, n_c = 85$)	No sorting	6.00 [6, 6]	6.00 [6, 6]	6.02 [6, 7]	6.69 [6, 9]
	Sorting	4.81 [4, 6]	5.00 [4, 6]	5.74 [5, 7]	6.86 [5, 10]
Sparse grid ($l = 7, n_c = 15,121$)	No sorting	6.00 [6, 6]	6.00 [6, 6]	6.07 [6, 8]	7.03 [6, 10]
	Sorting	4.08 [2, 6]	4.48 [2, 6]	5.15 [3, 7]	6.47 [4, 10]

measure $\|\boldsymbol{\xi}\|_1^w = \sum_{k=1}^{M} \sqrt{\lambda_k}\,|[\boldsymbol{\xi}]_k|$. Large savings are observed for the easier linear problems with large l. For the nonlinear problems (4), systems are simply "less similar."

In conclusion, we have demonstrated that is it feasible to use AMG preconditioning to solve the linear systems that arise when elliptic PDEs with random data are discretized via SCMs. Substantial computational savings are achieved over mean-based preconditioning for the stochastically nonlinear problem, if set-up information is recycled. The scheme is applicable for any sampling method, including MCMs and SGMs based on doubly-orthogonal polynomials. Recycled Krylov subspace solvers, as studied in [8] and [14], can also be employed. It remains to be seen, however, whether there are any real benefits when strong preconditioners, and ordering strategies, such as the ones we have suggested, are employed. In initial experiments, we found that using plain CG with a good preconditioner was cheaper overall.

References

1. Babuška I., Nobile F., Tempone R., *A stochastic collocation method for elliptic partial differential equations with random input data*. SIAM J. Numer. Anal. **45(3)**, 1005–1034 (2007)
2. Babuška I., Tempone R., Zouraris G. E., *Galerkin finite element approximations of stochastic elliptic partial differential equations*. SIAM J. Numer. Anal. **42(2)**, 800–825 (2004)

3. Barthelmann V., Novak E., Ritter K., *High dimensional polynomial interpolation on sparse grids*. Adv. Comput. Math. **12**, 273–288 (2000)
4. Boyle J., Mihajlović M. D., Scott J. A., *HSL_MI20: an efficient amg preconditioner*. Technical Report RAL-TR-2007-021, SFTC Rutherford Appleton Laboratory, Didcot (2007)
5. Elman H., Silvester D., Wathen A., *Finite Elements and Fast Iterative Solvers*. Oxford University Press, New York (2005)
6. Frauenfelder P., Schwab C., Todor R.A., *Finite elements for elliptic problems with stochastic coefficients*. Comput. Methods Appl. Mech. Eng. **194**, 205–228 (2005)
7. Ghanem R.G., Spanos P., *Stochastic Finite Elements: A Spectral Approach*. Springer, New York (1991)
8. Jin C., Cai X-C., Li C., *Parallel domain decomposition methods for stochastic elliptic equations*. SIAM J. Sci. Comput. **29(5)**, 2096–2114 (2007)
9. Loève M., *Probability Theory, Vol. II (4th ed.)*. Springer, New York (1978)
10. Nobile F., Tempone R., Webster C. G., *A sparse grid stochastic collocation method for partial differential equations with random input data*. SIAM J. Numer. Anal. **46(5)**, 2309–2345 (2008)
11. Parks M. L., de Sturler E., Mackey G., Johnson D. D., Maiti S., *Recycling Krylov subspaces for sequences of linear systems*. SIAM J. Sci. Comput. **28(5)**, 1651–1674 (2006)
12. Smolyak S., *Quadrature and interpolation formulas for tensor products of certain classes of functions*. Dokl. Akad. Nauk. SSSR **4**, 240–243 (1963)
13. Stüben K., *Algebraic multigrid (AMG): an introduction with applications*. In: Trottenberg U., et al. (eds.) Multigrid. Academic Press, New York (2000)
14. Ullmann E., *Krylov subspace recycling methods in stochastic finite element computations*. Technical report 2008–02 (ISSN 1433–9307), University of Freiberg, Germany (2008)
15. Xiu D., Hesthaven J. S., *High-order collocation methods for differential equations with random inputs*. SIAM J. Sci. Comput. **27(3)**, 1118–1139 (2005)

Adaptive Two-Step Peer Methods for Incompressible Navier–Stokes Equations

B. Gottermeier and J. Lang

Abstract The paper presents a numerical study of two-step peer methods up to order six, applied to the non-stationary incompressible Navier–Stokes equations. These linearly implicit methods show good stability properties, but the main advantage over one-step methods lies in the fact that even for PDEs no order reduction is observed. To investigate whether the higher order of convergence of the two-step peer methods equipped with variable time steps pays off in practically relevant CFD computations, we consider typical benchmark problems. Higher accuracy and better efficiency of the two-step peer methods compared to classical third-order one-step methods of Rosenbrock-type can be observed.

1 Introduction

In industrial and scientific applications, incompressible flows are modelled by the well-known Navier–Stokes equations, for which the time-dependent system on the domain $[0, T] \times \Omega, \Omega \subset \mathbb{R}^2$ is given by the following nonlinear equations

$$\partial_t u - Re^{-1} \Delta u + (u \cdot \nabla) u + \nabla p = F \qquad \text{in } (0, T] \times \Omega \tag{1}$$

$$\nabla \cdot u = 0 \qquad \text{in } [0, T] \times \Omega \tag{2}$$

$$u = G \qquad \text{on } [0, T] \times \partial\Omega \tag{3}$$

$$u(0, x) = u_0 \qquad x \in \Omega. \tag{4}$$

The vector $u = (u_1, u_2)^T \in \mathbb{R}^2$ represents the velocity field, the scalar p the pressure function and F denotes external forces. The Reynolds number Re will be limited to laminar flows in this paper. The function G and u_0 are given by boundary and initial conditions, respectively.

B. Gottermeier and J. Lang (✉)
Technische Universität Darmstadt, Department of Mathematics, Dolivostr. 15, 64293 Darmstadt, DFG Cluster of Excellence Smart Interfaces, Petersenstr. 32, 64287 Darmstadt, Germany
e-mail: gottermeier@mathematik.tu-darmstadt.de, lang@mathematik.tudarmstadt.de

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_41,

In recent time, intensive research on numerical methods has been made for an accurate and efficient numerical solution of the Navier–Stokes equations. Unfortunately, classical one-step methods, such as Runge–Kutta and Rosenbrock methods, suffer from order reduction when they are applied to partial differential equations (PDEs). In this paper, linearly implicit two-step peer methods are used to solve the nonlinear Navier–Stokes equations. They are based on a linear combination of approximations of equal order to the exact solution at intermediate points. For having the same accuracy and stability properties, these variables are called "peer." There exists methods up to order six, which provide good stability properties, i.e., optimal zero-stability and $L(\alpha)$-stability with an angle α of at least 85°. Strong damping properties at infinity are given without further restrictions, which leads to robust methods with respect to stepsize changes. Because of the embedding of the Jacobian matrix directly into the integration formula, they require only the solution of linear systems in each time step, making them very attractive for practical computations. The main advantage over one-step methods lies in the fact that even the higher-order peer methods have shown no order reduction when they are applied to PDEs. Additionally, they are competitive compared to the applied one-step methods and sometimes even more efficient [2].

The present paper is organized as follows: In Sect. 2 we begin with the time and space discretizations of the Navier–Stokes equations and explain the strategy for the time adaptivity in KARDOS [1], the finite element software package used for the numerical computations. Results of the numerical simulations are contained in Sect. 3. An analytical example to validate the higher orders of convergence as well as a typical benchmark problem are presented. Finally, we summarize our results and conclusions in Sect. 4.

2 Discretization of the Navier–Stokes Equations

For a higher-order temporal discretization of the instationary Navier–Stokes equations, we apply an s-stage linearly implicit two-step peer method [2] to (4).

Let $\tau_m > 0$ be a variable time step and $V_{mi} = (P_{mi}, U_{mi})^T$ the approximation to the exact solution at time $t_{mi} := t_m + c_i \tau_m$ with $t_m = t_{m-1,s}$ for $m \geq 1$ and $c_i \in [-1, 1]$, $c_s = 1$. Then the system of linear equations which has to be solved for each time step reads as

$$\left(\frac{I}{\tau_m \gamma} - Re^{-1}\Delta + U_{m-1,s} \cdot \nabla\right)(U_{mi} - U^0_{mi}) + ((U_{mi} - U^0_{mi}) \cdot \nabla) U_{m-1,s} + \nabla(P_{mi} - P^0_{mi})$$
$$= (Re^{-1}\Delta - (U^0_{mi} \cdot \nabla)) U^0_{mi} - \nabla P^0_{mi} + \frac{1}{\tau_m \gamma}(Q w_i - U^0_{mi}) + F(t_{mi}) \quad (5)$$

$$\nabla \cdot (U_{mi} - U^0_{mi}) = -\nabla \cdot U^0_{mi} \quad (6)$$

for $i = 1, ..., s$, with the boundary conditions $U_{mi} = G(t_{mi})$. The internal values and the predictors are given by

$$w_i = \sum_{j=1}^{i-1} \frac{1}{\gamma} a_{ij}(V_{mj} - w_j) + \sum_{j=1}^{s} u_{ij}(\sigma_m) V_{m-1,j},$$
$$V_{mi}^0 = \sum_{j=1}^{i-1} \frac{1}{\gamma} a_{ij}^0(V_{mj} - w_j) + \sum_{j=1}^{s} u_{ij}^0(\sigma_m) V_{m-1,j}.$$

The matrix Q is defined such that only the second component of the vector w_i is selected. The values for the abscissa $c \in \mathbb{R}^s$ are stretched Chebychev nodes

$$c_i := -\frac{\cos\left(\left(i - \frac{1}{2}\right)\frac{\pi}{s}\right)}{\cos\left(\frac{\pi}{2s}\right)}, \quad i = 1, ..., s,$$

with $c_s = 1$. The remaining coefficients of the method are combined in a lower triangular matrix $A = (a_{ij}) \in \mathbb{R}^{s\times s}$ with constant diagonal elements $a_{ii} = \gamma > 0$ and in a possibly full matrix $U = (u_{ij}) \in \mathbb{R}^{s\times s}$ which depends on the step size ratio $\sigma_m := \tau_m/\tau_{m-1}$. For the predictor V_{mi}^0, the real coefficient matrices have similar properties:

$$A^0 = (a_{ij}^0) \quad \text{with} \quad a_{ij}^0 = 0 \quad \text{for} \quad i \le j \quad \text{and} \quad U^0 = (u_{ij}^0(\sigma_m)).$$

The coefficients are chosen in such a way that the method has order $p = s$ for constant step size and order $p = s - 1$ for variable step size.

Adaptivity in time is gained with an embedding strategy. A second solution $\widetilde{V}_{ms}$ of inferior order $\tilde{p} = s - 2$ is computed by a linear combination of the V_{mi}, $i = 1, ..., s - 1$. The new time step size is then defined by

$$\tau_{\text{new}} = \min\{\tau_{\max}, \min\{2, \max\{0.2, (TOL_t/ERR_t)^{1/(\tilde{p}+1)}\}\} \times 0.9\tau_m\},$$

where

$$ERR_t := \left(\frac{1}{n} \sum_{i=1}^{n} \frac{\|V_{ms} - \widetilde{V}_{ms}\|_{L_2}^2}{\left(Scal\, R_i\, \|e_i^T V_{ms}\|_{L_2} + Scal\, A_i \sqrt{|\Omega|}\right)^2} \right)^{\frac{1}{2}}$$

and $Scal\, R_i$, $Scal\, A_i$ and TOL_t are user-prescribed parameters for the relative and absolute scaling factors and the desired time tolerance, respectively. For a more detailed description we refer to [2].

In the next step, the arising spatial problems (5-6), now independent of time, are solved by a multilevel finite element method [5]. We select finite dimensional subspaces S_h^q of the finite element meshes $\mathcal{T}_m^h$ at time $t = t_m$ with refinement level h, where the continuous functions of S_h^q are chosen to be polynomials of order q on each finite element $T \in \mathcal{T}_m^h$. Defining $\hat{P}_{mi} = P_{mi} - P_{mi}^0$ and $\hat{U}_{mi} = U_{mi} - U_{mi}^0$,

the standard Galerkin finite element solutions $V_{mi}^h \in S_h^q$ can then be computed from the equations

$$\begin{aligned}\frac{1}{\tau_m \gamma}\left(\hat{U}_{mi}^h, \varphi\right) - Re^{-1}\left(\Delta \hat{U}_{mi}^h, \varphi\right) + \left((U_{m-1,s}^h \cdot \nabla)\hat{U}_{mi}^h, \varphi\right) \\ + \left((\hat{U}_{mi}^h \cdot \nabla)U_{m-1,s}^h, \varphi\right) + \left(\nabla \hat{P}_{mi}^h, \varphi\right) = \left(\hat{F}^h(t_{mi}, P_{mi}^{0,h}, U_{mi}^{0,h}), \varphi\right) \\ \left(\nabla \cdot \hat{U}_{mi}^h, \varphi\right) = -\left(\nabla \cdot U_{mi}^{0,h}, \varphi\right), \quad \forall \varphi \in S_h^q,\end{aligned}$$

where $\hat{F}^h(t_{mi}, P_{mi}^{0,h}, U_{mi}^{0,h})$ is the right hand side of (5). Because of numerical oscillations in V_{mi}^h due to advection-dominated terms, we choose a Galerkin/least-squares method to stabilize the discretization by adding locally weighted residuals as described in [5]. We use the same finite element functions for the pressure and the velocity. To avoid spurious pressure modes of the numerical solution, a relaxation of the incompressibility condition

$$\nabla \cdot u = \delta \nabla \cdot \left(\partial_t u - Re^{-1}\Delta u + (u \cdot \nabla)u + \nabla p - F\right)$$

is applied to get a stable discretization, where δ is defined by

$$\delta = c\frac{h_b}{2u_{\text{ref}}}\frac{\hat{Re}}{\sqrt{1+\hat{Re}^2}}, \quad \hat{Re} = h_b u_{\text{ref}} Re, \quad c = 0.4,$$

with a global reference velocity u_{ref} and the diameter h_b of the two-dimensional ball which is area-equivalent to the element $T \in \mathscr{T}^h$.

3 Numerical Results

We first apply the two-step methods PEER4, PEER5 and PEER6 [2] to (4) with given analytical solution to validate their classical orders 4, 5 and 6, respectively. Then, a typical benchmark problem is considered to study the accuracy and efficiency of these methods equipped with variable time steps. Comparisons are made with linearly implicit one-step Rosenbrock methods ROS3P [7] and ROS3PL [6] of classical order three.

3.1 Analytical Example

We choose a very stiff test case [4], where the computational domain is chosen to be the unit square $\Omega = (0,1)^2$, the final time is T and the Reynolds number Re is set to 1. The functions F, G and u_0 in (4) are computed with the help of the exact solution

$$
\begin{aligned}
p(t,x,y) &= (10+t)\,\mathrm{E}^{-t}(x+y-1)\\
u_1(t,x,y) &= t^3 y^2\\
u_2(t,x,y) &= e^{2^{-50tx}}.
\end{aligned}
$$

We implement Dirichlet boundary conditions and use a quadratic ansatz for the finite elements on a fixed spatial mesh consisting of 2,048 triangles. In this way, the spatial discretization is solved exactly and the arising errors consist of the time integration errors only.

We consider the global error for the solution vector $v = (p, u_1, u_2)^T$ in the norm $L^2(0,1;L^2(\Omega))$, i.e.,

$$\|v - v_h\|_{L^2(0,1;L^2(\Omega))} = \left(\int_0^1 \|v - v_h\|^2_{L^2(\Omega)} dt\right)^{1/2}.$$

The simulations are performed with fixed and variable time steps and starting values taken from the exact solution.

3.1.1 Validation of Higher Orders of Convergence

Fixed time steps are chosen to validate the higher orders of convergence of the tested methods. The global errors for the two-step peer methods as well as for the one-step Rosenbrock solvers are presented in Fig. 1 for several numbers of time steps. The peer methods achieve their higher orders and show the super-convergence property. Likewise, the one-step Rosenbrock solvers show order three as expected. The advantage of the peer methods becomes obvious not only because of the higher order of convergence but also when considering the computed errors. At least one

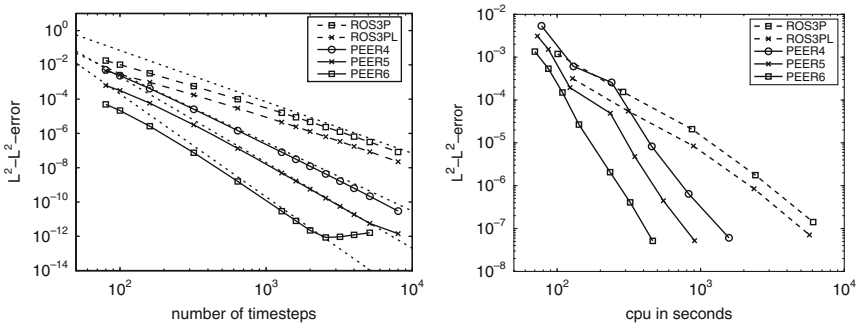


Fig. 1 In the *left* picture, where the computed error is drawn against the number of timesteps, the expected order can be observed for peer and Rosenbrock solvers. The picture on the *right* shows the higher efficiency of the two-step methods compared to the tested one-step methods. The requested time tolerances are $10^{-i}, i = 0, \ldots, 6$

order of magnitude for each time step size can be gained choosing a method with a higher order.

3.1.2 Efficiency

Time-adaptive simulations are considered to show the good performance of the peer methods compared to the tested Rosenbrock solvers. The requested time tolerances are set to 10^{-i}, $i = 0, \ldots, 6$, and the new time step is chosen according to the adaptive strategy in Sect. 2. The results are shown in the right picture of Fig. 1, where the global error is drawn against CPU time in seconds. The higher efficiency of the peer methods compared to the Rosenbrock methods can be observed clearly, especially for more stringent time tolerances. Considering shorter CPU times the two-step methods are still comparable with the one-step Rosenbrock solvers.

3.2 Flow Around a Cylinder

We consider the benchmark problem of a laminar flow around a cylinder in two dimensions, which was defined within the DFG Priority Research Programme "Flow Simulation on High Performance Computers" [8]. We choose the third case therein, which uses the instationary Navier–Stokes equations (4) combined with a time-dependent parabolic inflow profile

$$u_1(t, 0, y) = 4u_m y(H - y)\sin(\pi t/8)/H^2, \quad u_2 = 0$$

with $u_m = 1.5\,\mathrm{m\,s^{-1}}$ in the time interval $0 \le t \le 8\,\mathrm{s}$. The computational domain Ω is shown in Fig. 2 with $H = 0.41\,\mathrm{m}$ and the diameter $D = 0.1\,\mathrm{m}$ of the cylinder. The characteristic values of the fluid are the kinematic viscosity $\nu = 10^{-3}\,\mathrm{m^2\,s^{-1}}$ and the density $\rho = 1.0\,\mathrm{kg\,m^{-3}}$. The time-dependent Reynolds number is defined by $Re = \bar{u}D/\nu$ with mean velocity $\bar{u} = 2u_{\max}/3 = 2u(t, 0, H/2)/3$ yielding values in the interval $[0, 100]$.

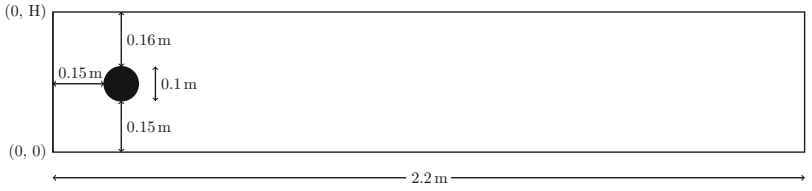


Fig. 2 Computational domain Ω

The initial conditions are $u_1 = u_2 = 0$. For the velocity vector, we prescribe the given parabolic inflow profile for $x = 0$ and a non-flux boundary condition for $x = 2.2$. On all remaining parts, no-slip boundary conditions are imposed.

The performance of the different solvers which are applied to the benchmark problem are compared by means of the drag and lift coefficients as well as the pressure difference between the front and the back of the cylinder

$$\Delta p(t) = p(t,\ 0.15,\ 0.2) - p(t,\ 0.25,\ 0.2).$$

We use a volume integral formulation for the drag and lift coefficients

$$c_d(t) = -k \int_\Omega \left[\partial_t u(t) \cdot v_d + \nu \nabla u(t) : \nabla v_d + (u(t) \cdot \nabla) u(t) \cdot v_d - p(t)(\nabla \cdot v_d)\right] dx dy,$$
$$c_l(t) = -k \int_\Omega \left[\partial_t u(t) \cdot v_l + \nu \nabla u(t) : \nabla v_l + (u(t) \cdot \nabla) u(t) \cdot v_l - p(t)(\nabla \cdot v_l)\right] dx dy,$$

$k := 2/(\rho D u_{\max}^2)$, as these seem to be more accurate and less susceptible to approximations of the cylinder boundary than a line integral formulation [3]. The formulas are valid for all functions $v_d, v_l \in (H^1(\Omega))^2$ with $(v_d)|_S = (1,0)^T, (v_l)|_S = (0,1)^T$ on the cylinder boundary S and vanishing v_d, v_l on all other boundaries. We take the maximal values for the drag and lift coefficients in the time interval $[0, T]$ and the pressure difference at the final time $T = 8\,\mathrm{s}$ and compare them with the following reference values:

$$c_{d,\max} = 2.952003, \qquad c_{l,\max} = 0.4773925, \qquad \Delta p = -0.1116111.$$

These values are obtained by performing the simulations with the peer methods and decreasing constant stepsizes until the convergence up to the sixth decimal of the values. The obtained reference values equal for all peer methods. They lie in the intervals computed in [8] and are comparable to the ones obtained in [3].

A very fine mesh, particularly at the cylinder boundary, which consists of 119,918 triangles, is used because of the sensitivity of the lift coefficient to the discretization in space. Linear finite elements are chosen to perform the time-adaptive simulations.

The good performance of the two-step peer methods compared to the one-step methods can be observed in Fig. 3 considering CPU times needed for comparable accuracy. The peer methods, in particular PEER5, are highly accurate and much more efficient than the tested one-step methods ROS3P and ROS3PL. Note that although PEER5 and PEER6 show an irregular behaviour for the drag and lift coefficient, they still remain efficient with regard to computing time.

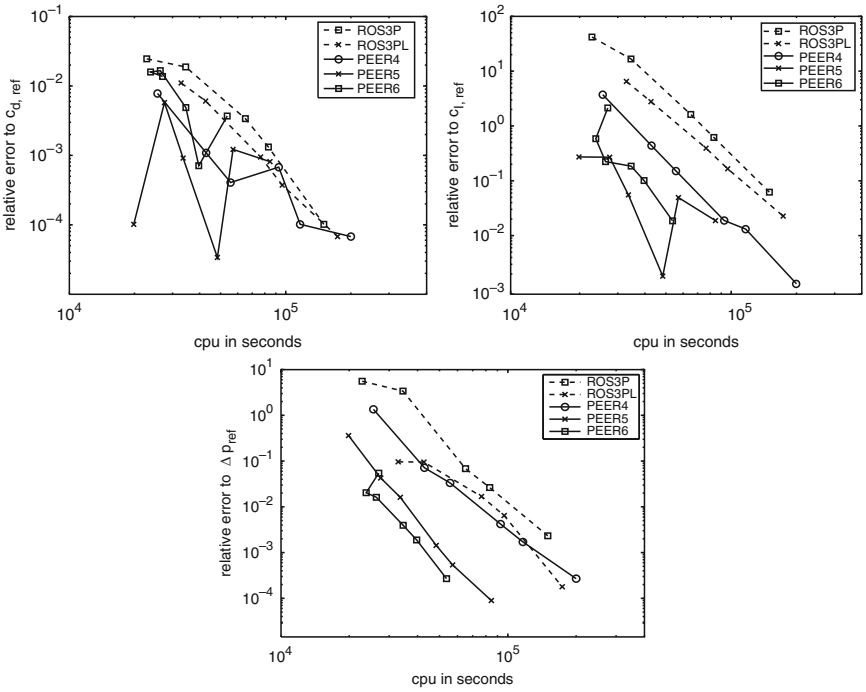


Fig. 3 Relative error of drag (*top left*) and lift (*top right*) coefficients and pressure difference (*bottom*) for peer and Rosenbrock solvers. The requested time tolerances are $5\times10^{-3}, 10^{-3}, \ldots, 10^{-5}$ for the peer methods and $10^{-4}, 5\times10^{-5}, \ldots, 10^{-6}$ for the Rosenbrock methods

4 Conclusions

We have applied linearly implicit two-step peer methods and a multilevel finite element method based on a Galerkin/least-squares stabilization to solve the non-stationary incompressible Navier–Stokes equations. The two-step methods have shown their expected classical orders of convergence and even the super-convergence property can be observed. Compared to Rosenbrock-type one-step methods, the peer methods are more accurate and provide an efficient solution for incompressible flows.

References

1. Erdmann, B., Lang, J., Roitzsch, R.: KARDOS user's guide. Tech. Rep. ZR 02–42, Konrad-Zuse-Zentrum, Berlin, 2002
2. Gerisch, A., Lang, J., Podhaisky, H., Weiner, R.: High-order linearly implicit two-step peer-finite element methods for time-dependent PDEs. Appl. Numer. Math. **59**, 624–638 (2009)

3. John, V.: Reference values for drag and lift of a two-dimensional time-dependent flow around a cylinder. Int. J. Numer. Meth. Fluids **44**, 777–788 (2004)
4. John, V., Matthies, G., Rang, J.: A comparison of time-discretization/linearization approaches for the incompressible Navier–Stokes equations. Comput. Methods Appl. Mech. Eng. **195**, 5995–6010 (2006)
5. Lang, J.: Adaptive incompressible flow computations with linearly implicit time discretization and stabilized finite elements. In: Papailiou, K.D., Tsahalis, D., Periaux, J., Hirsch, C., Pandolfi, M. (eds.) Computational Fluid Dynamics, pp. 200–204. Wiley, New York (1998)
6. Lang, J., Teleaga, D.: Towards a fully space-time adaptive FEM for magnetoquasistatics. IEEE Trans. Magn. **44**, 1238–1241 (2008)
7. Lang, J., Verwer, J.: ROS3P – an accurate third-order Rosenbrock solver designed for parabolic problems. BIT **41**, 731–738 (2001)
8. Schäfer, M., Turek, S.: Benchmark computations of laminar flow around a cylinder. In: Hirschel, E.H. (ed.) Flow Simulation with High-Performance Computers II, pp. 547–566. Vieweg (1996)

On Hierarchical Error Estimators for Time-Discretized Phase Field Models

Carsten Gräser, Ralf Kornhuber, and Uli Sack

Abstract We suggest hierarchical a posteriori error estimators for time-discretized Allen–Cahn and Cahn–Hilliard equations with logarithmic potential and investigate their robustness numerically. We observe that the associated effectivity ratios seem to saturate for decreasing mesh size and are almost independent of the temperature.

1 Introduction

Hierarchical a posteriori error estimators are based on the extension of the given finite element space $\mathcal{S}$ by an incremental space $\mathcal{V}$. After discretization of the actual defect problem with respect to the extended space $\mathcal{Q} = \mathcal{S} + \mathcal{V}$, hierarchical preconditioning and subsequent localization give rise to local defect problems associated with low-dimensional subspaces of $\mathcal{V}$. The resulting local contributions to the desired global error estimate are often used as error indicators in an adaptive refinement process. We refer to the pioneering work of Zienkiewicz et al. [23] and Deuflhard et al. [10] or to the monograph of Ainsworth and Oden [1].

Local lower bounds by hierarchical error estimators typically come without unknown constants, e.g., for linear self-adjoint problems. In this sense, hierarchical error estimators are properly scaled by construction. In early papers, upper bounds are often derived from the so-called saturation assumption that the extended space $\mathcal{Q}$ provides a more accurate approximation than $\mathcal{S}$. It turned out later that local equivalence to residual estimators provides upper bounds up to data oscillation and, conversely, that small data oscillation implies the saturation assumption [7, 12]. For a direct proof based on local L^2-projections we refer to [24].

Another attractive feature of hierarchical error estimators is their intriguing simplicity, particularly as applied to nonlinear, non-smooth problems [2, 17, 19–22, 24]. In this numerical study, we consider hierarchical error estimators for semilinear elliptic problems as arising from the time discretization of Allen–Cahn and

C. Gräser (✉), R. Kornhuber, and U. Sack
Freie Universität Berlin, Germany
e-mail: graeser@mi.fu-berlin.de

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_42, © Springer-Verlag Berlin Heidelberg 2010

Cahn–Hilliard equations. While previous work concentrates on quartic shallow quench approximations [4,13,18] or on heuristic strategies for obstacle potentials [5] we consider the logarithmic potential here. In particular, we investigate the robustness of the effectivity ratios as temperature is approaching the deep quench limit. In our numerical experiments, we found that both for Allen–Cahn- and Cahn–Hilliard-type problems the associated effectivity ratios seem to saturate with decreasing mesh size and are hardly influenced by temperature. Moreover, the local contributions to the global estimator were used successfully for adaptive refinement.

2 Hierarchical Error Estimators

In this section, we derive hierarchical error estimators in an abstract setting. Special cases will be considered later on. For ease of presentation, we assume that all occurring problems and subproblems are uniquely solvable. Let H denote a Hilbert space with the norm $\|\cdot\|_H$. We consider the variational inequality

$$u \in H: \qquad a(u, v-u) + \phi(v) - \phi(u) \geq \ell(v-u) \qquad \forall v \in H \tag{1}$$

with $a(\cdot,\cdot)$, $\phi : H \to \mathbb{R} \cup \{+\infty\}$, and ℓ denoting a symmetric bilinear form, a convex functional, and a bounded linear functional on H, respectively. The additional conditions that $a(\cdot,\cdot)$ is H-elliptic and that ϕ is lower semi-continuous and proper are sufficient but not necessary to ensure existence and uniqueness [14]. Let $\mathcal{S}$ denote a finite-dimensional subspace of H and let the symmetric bilinear form $a_{\mathcal{S}}(\cdot,\cdot)$ and the functional $\phi_{\mathcal{S}} : \mathcal{S} \to \mathbb{R} \cup \{+\infty\}$ be approximations of $a(\cdot,\cdot)$ and ϕ on $\mathcal{S}$, respectively, e.g., by numerical quadrature like mass lumping. Then the associated Ritz–Galerkin discretization reads

$$u_{\mathcal{S}} \in \mathcal{S}: \qquad a_{\mathcal{S}}(u_{\mathcal{S}}, v-u_{\mathcal{S}}) + \phi_{\mathcal{S}}(v) - \phi_{\mathcal{S}}(u_{\mathcal{S}}) \geq \ell(v-u_{\mathcal{S}}) \qquad \forall v \in \mathcal{S}. \tag{2}$$

We want to derive a posteriori estimates of the error $\|u - u_{\mathcal{S}}\|_H$. To this end, we consider the *defect problem*

$$e \in H: \qquad a(e, v-e) + \psi(v) - \psi(e) \geq r(v-e) \qquad \forall v \in H \tag{3}$$

involving the shifted nonlinearity ψ and the residual r, defined by

$$\psi(v) = \phi(u_{\mathcal{S}} + v), \qquad r(v) = \ell(v) - a(u_{\mathcal{S}}, v), \qquad v \in H,$$

respectively. Obviously, $u = u_{\mathcal{S}} + e$. To approximate (3), we select an incremental space $\mathcal{V} \subset H$ with the property $\mathcal{V} \cap \mathcal{S} = \{0\}$ and consider the hierarchical extension

$$\mathcal{Q} = \mathcal{S} \oplus \mathcal{V}$$

of $\mathcal{S}$. The subspace $\mathcal{Q} \subset H$ is equipped with the discrete norm $\|\cdot\|_{\mathcal{Q}}$ which intentionally is an equivalent approximation of $\|\cdot\|_H$. The associated *discretized defect problem* is given by

$$e_{\mathcal{Q}} \in \mathcal{Q}: \qquad a_{\mathcal{Q}}(e_{\mathcal{Q}}, v - e_{\mathcal{Q}}) + \psi_{\mathcal{Q}}(v) - \psi_{\mathcal{Q}}(e_{\mathcal{Q}}) \geq r(v - e_{\mathcal{Q}}) \qquad \forall v \in \mathcal{Q} \tag{4}$$

with $a_{\mathcal{Q}}(\cdot,\cdot)$ and $\psi_{\mathcal{Q}} : \mathcal{Q} \to \mathbb{R} \cup \{+\infty\}$ denoting approximations of $a(\cdot,\cdot)$ and ψ. In order to avoid the computational effort for the computation of $e_{\mathcal{Q}}$, we now modify (4) in a way that allows for a decomposition into a number of independent, low-dimensional subproblems. This modification is based on the assumption that the given nonlinearity is local in the sense that there is a direct splitting

$$\mathcal{V} = \mathcal{V}_1 \oplus \cdots \oplus \mathcal{V}_m$$

of $\mathcal{V}$ into low-dimensional subspaces $\mathcal{V}_i$ such that for $v \in \mathcal{V}$ the representation

$$\psi_{\mathcal{Q}}(v) = \sum_{i=1}^{m} \Psi_i(v_i) \tag{5}$$

holds with certain convex functionals $\Psi_i : \mathcal{V}_i \to \mathbb{R} \cup \{+\infty\}$ and the uniquely determined decomposition $v = \sum v_i$, $v_i \in \mathcal{V}_i$. Then, in the first step, we replace the bilinear form $a(\cdot,\cdot)$ by the hierarchical preconditioner

$$b(v, w) = a_{\mathcal{Q}}(v_{\mathcal{S}}, w_{\mathcal{S}}) + \sum_{i=1}^{m} a_{\mathcal{Q}}(v_i, w_i), \qquad v, w \in \mathcal{Q},$$

based on the uniquely determined decompositions of $v = v_{\mathcal{S}} + v_{\mathcal{V}} \in \mathcal{Q}$ into $v_{\mathcal{S}} \in \mathcal{S}$, $v_{\mathcal{V}} \in \mathcal{V}$ and of $v_{\mathcal{V}} = \sum v_i$, $w_{\mathcal{V}} = \sum w_i$ into $v_i, w_i \in \mathcal{V}_i$. It can be shown under certain conditions [19] that the solution $\tilde{e}_{\mathcal{Q}}$ of the resulting preconditioned defect problem provides an efficient and reliable error estimate $b(\tilde{e}_{\mathcal{Q}}, \tilde{e}_{\mathcal{Q}})^{1/2}$. However, the exact evaluation of $\tilde{e}_{\mathcal{Q}} = \tilde{e}_{\mathcal{S}} + \tilde{e}_{\mathcal{V}}$ is still too costly: In contrast to linear situations, we cannot expect $\tilde{e}_{\mathcal{S}} = 0$, because $\tilde{e}_{\mathcal{S}} \in \mathcal{S}$ and $\tilde{e}_{\mathcal{V}} \in \mathcal{V}$ are still coupled with respect to the nonlinearity $\psi_{\mathcal{Q}}(v) = \psi_{\mathcal{Q}}(v_{\mathcal{S}} + v_{\mathcal{V}})$. As a remedy, we simply *assume* that the low-frequency part $\tilde{e}_{\mathcal{S}}$ of our error estimate can be neglected. In this way, we finally obtain the *localized defect problem*

$$e_{\mathcal{V}} \in \mathcal{V}: \qquad b(e_{\mathcal{V}}, v - e_{\mathcal{V}}) + \psi_{\mathcal{Q}}(v) - \psi_{\mathcal{Q}}(e_{\mathcal{V}}) \geq r(v - e_{\mathcal{V}}) \qquad \forall v \in \mathcal{V}. \tag{6}$$

It has been shown for obstacle problems that reliability might get lost by this localization step but can be reestablished by a suitable higher order term [20]. Exploiting assumption (5), the evaluation of $e_{\mathcal{V}} = \sum e_i$ amounts to the solution of m independent subproblems

$$e_i \in \mathcal{V}_i: \quad a_{\mathcal{Q}}(e_i, v - e_i) + \Psi_i(v) - \Psi_i(e_i) \geq r(v - e_i) \quad \forall v \in \mathcal{V}_i. \tag{7}$$

The quantity

$$\eta = \left(\sum_{i=1}^{m} \|e_i\|_{\mathcal{Q}}^2 \right)^{1/2} \tag{8}$$

is our hierarchical error estimator. If $a_{\mathcal{Q}}(\cdot,\cdot)$ is $\mathcal{Q}$-elliptic and $\mathcal{Q}$ is equipped with the energy norm $\|\cdot\|_{\mathcal{Q}} = a_{\mathcal{Q}}(\cdot,\cdot)^{1/2}$, then (8) takes the form $\eta = b(e_{\mathcal{V}}, e_{\mathcal{V}})^{1/2}$.

3 Allen–Cahn Equations

Implicit time discretization of the Allen–Cahn equation with logarithmic potential (see, e.g., [9, Sect. 7] for an overview) gives rise to spatial problems of the form (1) with $H = H^1(\Omega)$ equipped with the energy norm induced by the bilinear form

$$a(v,w) = \gamma(v,w)_{L^2(\Omega)} + \tau(\nabla v, \nabla w)_{L^2(\Omega)}, \qquad \gamma = 1 - \tau\theta_c/\varepsilon^2,$$

the right hand side $\ell(v) = (u_0, v)_{L^2(\Omega)}$, and the convex, lower semi-continuous, proper functional $\phi(v) = \int_\Omega \Phi^\theta(v)\, dx$ with

$$\Phi^\theta(v) = \begin{cases} \frac{\tau}{2\varepsilon^2}\theta\big((1+v)\log(1+v) + (1-v)\log(1-v)\big), & \text{if } \theta > 0 \\ \chi_{[-1,1]}(v), & \text{if } \theta = 0 \end{cases}. \tag{9}$$

Here, $\Omega \subset \mathbb{R}^2$ denotes a polygonal domain, $\varepsilon > 0$ is an interface parameter, $\theta \geq 0$ and $\theta_c > 0$ stand for the temperature and the critical temperature, respectively, $\chi_{[-1,1]}$ is the characteristic function of $[-1,1]$, $u_0 \in L^2(\Omega)$ is an approximation from the preceding time step, and $\tau > 0$ is the time step size. We assume $\gamma > 0$ or, equivalently, $\tau < \varepsilon^2/\theta_c$ so that $a(\cdot,\cdot)$ is H-elliptic. With these definitions, (1) can be rewritten as a semi-linear elliptic problem for positive temperature $\theta > 0$ and as an elliptic obstacle problem for $\theta = 0$.

Let $\mathcal{S} = \mathcal{S}_h$ denote the space of piecewise linear finite elements with respect to a regular triangulation $\mathcal{T}_h$ with mesh size h and interior vertices $\mathcal{N}_h$. Then, $a_{\mathcal{S}}(\cdot,\cdot) = a_h(\cdot,\cdot)$ is defined by replacing $(v,w)_{L^2(\Omega)}$ with the lumped L^2-scalar product $\langle v, w\rangle_{\mathcal{S}} = \int_\Omega I_h(vw)\, dx$, where $I_h : C(\overline{\Omega}) \to \mathcal{S}_h$ denotes nodal interpolation. Similarly, we set $\phi_{\mathcal{S}}(v) = \phi_h(v) = \int_\Omega I_h(\Phi^\theta(v))\, dx$. Connecting the midpoints p_i of the edges of all triangles $t \in \mathcal{T}_h$, we obtain the uniformly refined triangulation $\mathcal{T}_{h/2}$ with interior vertices $\mathcal{N}_{h/2}$. The local incremental spaces $\mathcal{V}_i = \operatorname{span}\{\mu_i\}$ are spanned by the piecewise linear edge bubble functions satisfying $\mu_i(p_i) = 1$ and vanishing on all other vertices $p \in \mathcal{N}_{h/2}$. This choice leads to $\mathcal{Q} = \mathcal{S}_{h/2}$. It is motivated by the lack of stability of piecewise quadratic approximations for obstacle problems [14, 20]. We select the discrete $\mathcal{Q}$-elliptic bilinear form

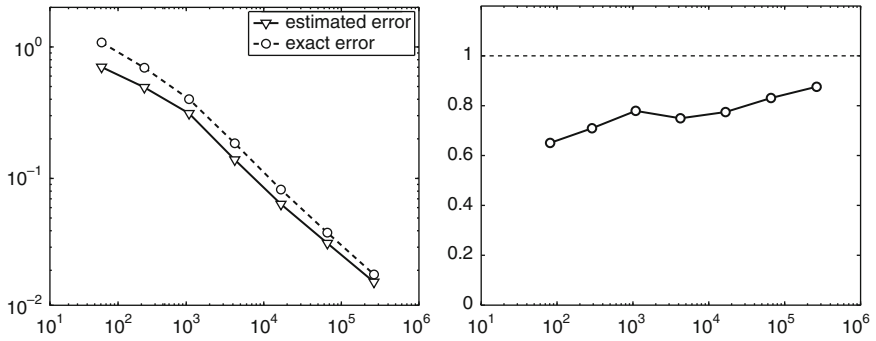


Fig. 1 Estimated and "exact" error (*left*) and effectivity ratio (*right*) over number of unknowns

$a_{\mathcal{Q}}(\cdot,\cdot) = a_{h/2}(\cdot,\cdot)$ with the associated energy norm $\|\cdot\|_{\mathcal{Q}} = a_{h/2}(\cdot,\cdot)^{1/2}$. Then the locality condition (5) is satisfied with $\Psi_i(v) = \Phi^\theta((u_S + v)(p_i)) \int_\Omega \mu_i \, dx$, $v \in \mathcal{V}_i$. For obstacle problems, i.e., for $\theta = 0$, the resulting error estimator (8) was proposed in [17] and later analyzed in [20, 22, 24]. Here we concentrate on $\theta > 0$ and investigate robustness for $\theta \to 0$.

In our numerical experiments, we consider the first spatial problem of the semi-discrete Allen–Cahn equation with parameters $\varepsilon = 2 \cdot 10^{-2}$, $\theta_c = 1.0$, time step size $\tau = 10^{-4}$, and the initial condition u_0 as depicted in the left picture of Fig. 3. We first compare the a posteriori error estimator with the "exact" error for a fixed temperature $\theta = 0.1$ and a sequence of triangulations $\mathcal{T}_{h_j}$ with decreasing mesh size h_j. The triangulations $\mathcal{T}_j = \mathcal{T}_{h_j}$ are obtained by $j = 1, \ldots, 9$ uniform refinements of the initial triangulation $\mathcal{T}_0$ which is a partition of $\Omega = (-1, 1) \times (-1, 1)$ into two congruent triangles. The "exact" error $\tilde{e}_j = \|\tilde{u} - u_j\|_{\widetilde{H}}$ is obtained by approximating H with $\widetilde{H} = \mathcal{S}_{11}$, i.e., by an approximation $\tilde{u}$ of u based on two further uniform refinement steps. The left picture in Fig. 1 shows η_j and $\tilde{e}_j$ over the number of unknowns. We observe asymptotic first order convergence and a good agreement of η_j and $\tilde{e}_j$. More precisely, the effectivity ratios $\eta_j / \tilde{e}_j$ seem to saturate at about 0.9 (Fig. 1 right). In our next experiment, we fix the mesh $\mathcal{T}_9$ and vary the temperature θ. The left picture in Fig. 2 shows that the effectivity ratios are hardly affected by the transition from a shallow to a deep quench and even seem to converge in the deep quench limit. In the last experiment, we use the edge-oriented local error indicators $\|e_i\|_{\mathcal{Q}}$ occurring in the global estimate (8) and a classical marking strategy [11] for adaptive mesh refinement. Figure 3 illustrates that refinement nicely follows the diffuse interface. Moreover, the zoom in Fig. 2 shows that refinement concentrates on the strong variation of the solution at the boundary of the diffuse interface and not on the interior where steep gradients are resolved sufficiently well. Finally, note that optimal order of convergence is preserved by adaptivity.

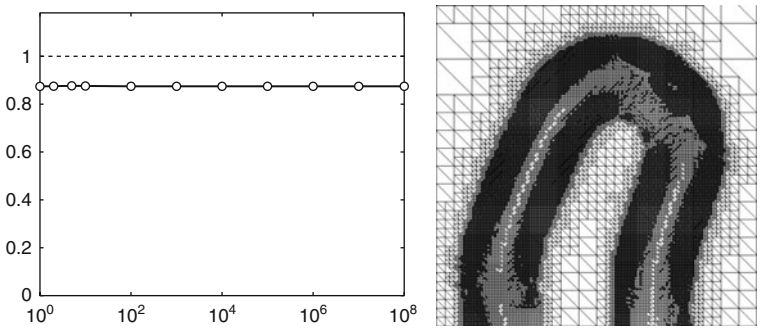


Fig. 2 Effectivity ratios over inverse temperature (*left*) and detail of adaptively refined mesh (*right*)

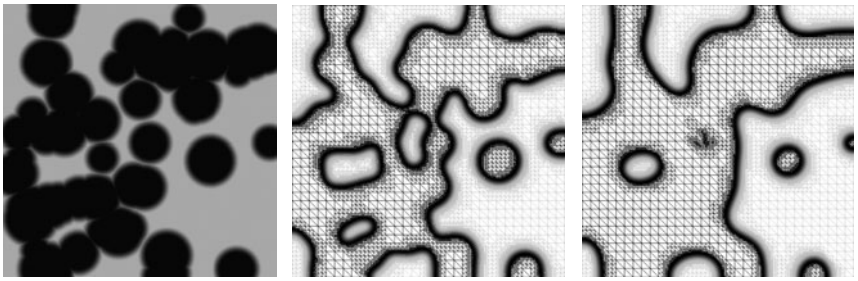


Fig. 3 Initial condition and approximations at time 20τ, 100τ

4 Cahn–Hilliard Equations

Semi-implicit time discretization [6,16] of the Cahn–Hilliard equation with logarithmic potential leads to spatial problems of the form (1) with $H = H^1(\Omega) \times H^1(\Omega)$ equipped with norm

$$\|(v_1, v_2)\|_H^2 = \varepsilon^2 \left(\|\nabla v_1\|_{L^2(\Omega)}^2 + (v_1, 1)_{L^2(\Omega)}^2 \right) + \tau \left(\|\nabla v_2\|_{L^2(\Omega)}^2 + \|v_2\|_{L^2(\Omega)}^2 \right),$$

the indefinite bilinear form

$$a(v, w) = \varepsilon^2 \left((\nabla v_1, \nabla w_1)_{L^2(\Omega)} + (v_1, 1)_{L^2(\Omega)} (w_1, 1)_{L^2(\Omega)} \right) - (v_2, w_1)_{L^2(\Omega)} \\ - (v_1, w_2)_{L^2(\Omega)} - \tau (\nabla v_2, \nabla w_2)_{L^2(\Omega)},$$

the right hand side $\ell(v) = (u_0, v_1 - v_2)_{L^2(\Omega)} + \varepsilon^2 (u_0, 1)_{L^2(\Omega)} (v_1, 1)_{L^2(\Omega)}$ and the convex functional $\phi(v) = \int_\Omega \Phi^\theta(v_1)\, dx$ with Φ^θ defined in (9) for temperature $\theta \geq 0$. Here, ε is an interface parameter, τ is the time step size and u_0 is an approximation from the preceding time step. Utilizing the notation of Sect. 3, the approximation (2) is based on $\mathcal{S} = \mathcal{S}_h \times \mathcal{S}_h$, and on $a_{\mathcal{S}}(\cdot, \cdot)$ and $\phi_{\mathcal{S}}$ as obtained

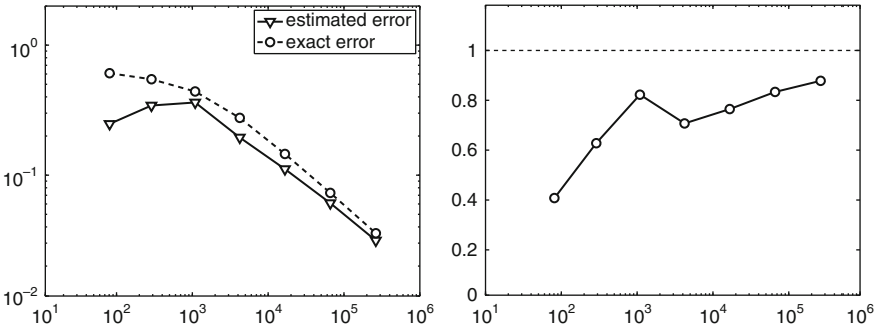


Fig. 4 Estimated and "exact" error (*left*) and effectivity ratios over number of unknowns

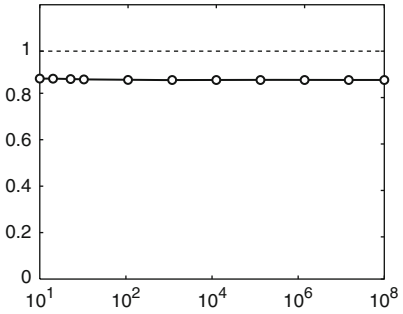


Fig. 5 Effectivity ratios over inverse temperature.

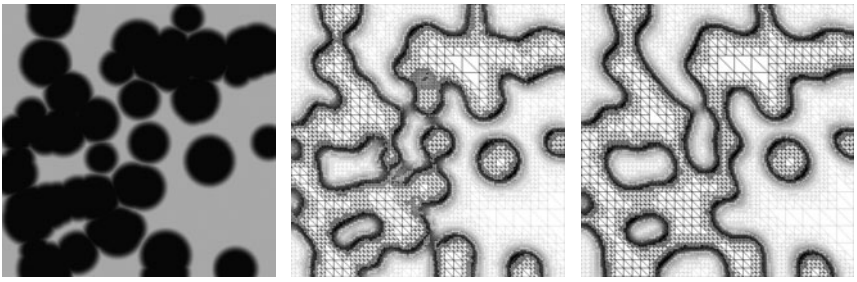


Fig. 6 Initial condition and approximations at time 20τ, 100τ

by mass lumping. Existence, uniqueness, and convergence results have been established in [6] for the double obstacle case $\theta = 0$ and in [3, 8] for $\theta > 0$. Fast solvers for the resulting algebraic problems are described in [15]. The components u_1 and u_2 of the solution $u = (u_1, u_2)$ are often called order parameter and chemical potential, respectively. Similar to Sect. 3, we select the incremental spaces $\mathcal{V}_i = \text{span}\{(\mu_i, 0), (0, \mu_i)\}$ providing $\mathcal{Q} = \mathcal{S}_{h/2} \times \mathcal{S}_{h/2}$. Again (5) is satisfied with $\Psi_i(v) = \Phi^\theta((u_{\mathcal{S},1} + v_1)(p_i)) \int_\Omega \mu_i \, dx$, $v \in \mathcal{V}_i$. In this setting the localized defect problem (6) admits a unique solution. The discrete norm $\|\cdot\|_\mathcal{Q}$ is obtained by (spectrally equivalent) mass lumping of the zero order terms in $\|\cdot\|_H$.

We consider the first spatial problem of the semi-discrete Cahn-Hilliard equation with parameters, time step size, and initial condition given in Sect. 3. In our numerical experiments, we proceed in complete analogy to the previous section. We begin with a comparison of the error estimators η_j with an "exact" error $\tilde{e}_j$ for fixed temperature $\theta = 0.1$ and decreasing mesh size h_j. Figure 4 shows optimal order of convergence and asymptotic saturation of the effectivity ratios $\eta_j / \tilde{e}_j$ at about 0.9. For fixed mesh $\mathcal{T}_9$ effectivity is hardly affected by strongly varying temperature θ as depicted in Fig. 5. Adaptive mesh refinement based on the local error indicators $\|e_i\|_{\mathcal{Q}}$ nicely captures strong variation of the order parameter as illustrated by Fig. 6. Strong variation of the chemical potential as occurring, e.g., after topological changes, is also reflected by adaptive refinement. Finally, it turned out that optimal order of convergence is preserved by adaptivity.

Acknowledgement This work was supported by the DFG Research Center MATHEON

References

1. M. Ainsworth and J.T. Oden. *A posteriori error estimation in FE analysis*. Wiley, NY, 2000
2. R.E. Bank and R.K. Smith. A posteriori error estimates based on hierarchical bases. *SIAM J. Numer. Anal.*, 30:921–935, 1993
3. J.W. Barrett and J.F. Blowey. Finite element approximation of an Allen-Cahn/Cahn-Hilliard system. *IMA J. Numer. Anal.*, 22(1):11–71, 2002
4. S. Bartels and R. Müller. Optimal and robust a posteriori error estimates in $L^\infty(L^2)$ for the approximation of Allen-Cahn equations past singularities. Preprint, Uni Bonn, 2009
5. L. Baňas and R. Nürnberg. A posteriori estimates for the Cahn-Hilliard equation with obstacle free energy. *Math. Model. Numer. Anal.*, 43:1003–1026, 2009
6. J.F. Blowey and C.M. Elliott. The Cahn-Hilliard gradient theory for phase separation with non–smooth free energy II: Numerical analysis. *Euro. J. Appl. Math.*, 3:147–179, 1992
7. 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.*, 33:1188–1204, 1996
8. M.I.M. Copetti and C.M. Elliott. Numerical analysis of the Cahn-Hilliard equation with a logarithmic free energy. *Numer. Math.*, 63:39–65, 1992
9. K. Deckelnick, G. Dziuk, and C.M. Elliott. Computation of geometric partial differential equations and mean curvature flow. *Acta Numer.*, 14:139–232, 2005
10. P. Deuflhard, P. Leinen, and H. Yserentant. Concepts of an adaptive hierarchical finite element code. *IMPACT Comput. Sci. Eng.*, 1:3–35, 1989
11. W. Dörfler. A convergent adaptive algorithm for Poisson's equation. *SIAM J. Numer. Anal.*, 33:1106–1124, 1996
12. W. Dörfler and R.H. Nochetto. Small data oscillation implies the saturation assumption. *Numer. Math.*, 91:1–12, 2002
13. X. Feng and H. Wu. A posteriori error estimates for finite element approximations of the Cahn-Hilliard equation and the Hele-Shaw flow. *J. Comput. Math.*, 26:767–796, 2008
14. R. Glowinski. *Numerical Methods for Nonlinear Variational Problems*. Springer, Berlin, 1984
15. C. Gräser. *Convex Minimization and Phase Field Models.* PhD thesis, FU Berlin, to appear
16. C. Gräser and R. Kornhuber. On preconditioned Uzawa-type iterations for a saddle point problem with inequality constraints. In O.B. Widlund and D.E. Keyes, editors, *Domain Decomposition Methods in Science and Engineering XVI*, pages 91–102. Springer, Berlin, 2006
17. R.H.W. Hoppe and R. Kornhuber. Adaptive multilevel-methods for obstacle problems. *SIAM J. Numer. Anal.*, 31(2):301–323, 1994

18. D. Kessler, R. Nochetto, and A. Schmidt. A posteriori error control for the Allen-Cahn problem: circumventing Gronwal's inequality. *M2AN*, 38:129–142, 2004
19. R. Kornhuber. A posteriori error estimates for elliptic variational inequalities. *Comput. Math. Appl.*, 31:49–60, 1996
20. R. Kornhuber and Q. Zou. Efficient and reliable hierarchical error estimates for the discretization error of elliptic obstacle problems. Math. Comp., to appear
21. O. Sander. *Multi-dimensional coupling in a human-knee model.* PhD thesis, FU Berlin, 2008
22. K.G. Siebert and A. Veeser. A unilaterally constrained quadratic minimization with adaptive finite elements. *SIAM J. Optim.*, 18:260–289, 2007
23. O.C. Zienkiewicz, J.P. De S.R. Gago, and D.W. Kelly. The hierarchical concept in finite element analysis. *Comput. Struct.*, 16:53–65, 1983
24. Q. Zou, A. Veeser, R. Kornhuber, and C. Gräser. Hierarchical error estimates for the energy functional in obstacle problems. Preprint 575, Matheon Berlin, 2009

Nonlinear Decomposition Methods in Elastodynamics

Christian Groß, Rolf Krause, and Mirjam Walloth

Abstract For the stable numerical solution of nonlinear elastodynamic contact problems implicit discretization schemes are required. Here, we discuss a temporal discretization scheme where in each time step spatial displacements are computed as the solution of an optimization problem allowing for the application of globalization strategies. Moreover, in order to treat the solution of these optimization problems more efficiently, a novel abstract nonlinear preconditioning framework for globalization strategies is presented.

1 Introduction

The numerical simulation of dynamic contact problems obeying a nonlinear elastic material law is a demanding task, as the contact conditions as well as the nonlinear material behavior have to be handled carefully. For the stability of the contact forces, their fully implicit treatment is clearly worthwhile: even in the case of linear elasticity, the straightforward application of common time discretizations as for example the classical Newmark scheme [11], may lead to energy blow-ups and, in consequence, to a loss of accuracy, see e.g., [10]. A possible remedy for this can be found in the purely implicit treatment of the contact forces, as first introduced in [8] and further analyzed in [3].

It is well known that for stability reasons an treatment of the nonlinear material behavior is favorable. We note that transferring stability results from linear to

C. Groß (✉) and R. Krause
Institute of Computational Science, University of Lugano, Switzerland
e-mail: christian.gross@usi.ch, rolf.krause@usi.ch

M. Walloth
Institute for Numerical Simulation, University of Bonn, Germany
e-mail: walloth@ins.uni-bonn.de

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_43,

nonlinear materials is not straightforward, see e.g., [1]. As a matter of fact, larger timesteps can usually only be realized employing an implicit discretization scheme, with the cost of solving (non-)linear problems in space in each timestep.

In order to allow for the efficient numerical simulation of elastodynamic contact problems, fast and robust solution methods for the arising constrained nonlinear and nonconvex minimization problems have to be provided. In the present work, we consider the application of recently developed nonlinear preconditioning techniques, cf. [7], in the context of dynamic contact problems in nonlinear mechanics.

With respect to the robustness of the solution of the arising nonlinear programming problems (especially for large time steps), globalization strategies, such as Linesearch or Trust-Region methods, are of particular interest. As a matter of fact, these strategies provably converge to first-order critical points with modest assumptions on the objective function. In particular, the paradigm of these iterative strategies is to damp the computed corrections in order to ensure a sufficient decrease of the objective function. Though, in return, the damping might slow down the convergence of the strategy.

Therefore, nonlinear multigrid methods such as RMTR [5] and MLS [12] were developed as efficient globalization strategies. As a more recent development, also a fully nonlinear domain decomposition method called APTS, see [7], has proven to be an efficient and reliable solution strategy for large-scale minimization problems. Basically, these strategies aim on a better resolution of low-frequency or "local" contributions of the solution in addition to the sole computation of Quasi-Newton steps.

To generalize these approaches, here we will employ an abstract nonlinear preconditioning framework, encompassing the concepts employed in the RMTR, MLS and APTS algorithms. Within this nonlinear preconditioning framework, local nonlinearities are resolved by a particular nonlinear update operator yielding an updated iterate. This nonlinear update operator may be, for instance, the result of an RMTR cycle. In a second step, a globalization strategy is employed to compute a "global" correction, which often is necessary to resolve high-frequent nonlinearities. Thus, nonlinearly preconditioned globalization strategies are well-suited to efficiently solve the nonlinear programming problems arising in elastostatic or elastodynamic simulations.

We will apply these nonlinear decomposition methods as spatial solver within the contact-implicit Newmark scheme, which here is extended to the case of non-linear material laws (see Sect. 2). For simplicity, we here have considered only linearized non-penetration conditions. Incorporating the exact geometric non-penetration condition, see, e.g., [9], or possible contact stabilizations, see, e.g., [3], into the new nonlinear framework employed here is out of the scope of this paper and will be subject of further research.

2 Dynamic Contact Problems for Non-Linear Materials and Their Time-Discretization

We are interested in the displacements $\boldsymbol{u} : [t_0, t_{\text{end}}] \times \Omega \to \mathbb{R}^3$ of a non-linear elastic body $\Omega \subset \mathbb{R}^3$ identified with a polyhedral domain which might be subjected to volume forces $\boldsymbol{F} : [t_0, t_{\text{end}}] \times \Omega(t) \to \mathbb{R}^3$, like for instance gravity. The non-linear, hyperelastic material behavior is characterized by a stored energy function $W : \Omega \times \mathbb{R}^{3\times 3} \to \mathbb{R}$ 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 $W(x, \cdot)$ is differentiable, the first Piola–Kirchhoff stress tensor is given by $\hat{\boldsymbol{T}}(x, \boldsymbol{C}) = \frac{\partial}{\partial \boldsymbol{C}} W(x, \boldsymbol{C})$. We also define $\hat{W}(x, \nabla \boldsymbol{u}) := W(x, \boldsymbol{C})$, which is the stored energy function depending on the space derivative of the displacements instead of the (right) Cauchy-Green strain tensor. 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, where the body is exposed to the contact constraints

$$\boldsymbol{u}(t, \cdot) \cdot \boldsymbol{n}(\cdot) \leq \phi(t, \cdot) \tag{1}$$

where $\phi : [t_0, t_{\text{end}}] \times \Gamma_C(t) \to \mathbb{R}^3$. Here and in the remainder we denote with $\boldsymbol{n}(x)$ the outer normal at $x \in \partial\Omega$. Then, the strong formulation for the dynamic contact problem to be considered here reads as

$$\rho \ddot{\boldsymbol{u}} - \operatorname{div} \hat{\boldsymbol{T}}(\boldsymbol{u}) = \boldsymbol{F} \text{ in } \Omega \tag{2a}$$

$$\hat{\boldsymbol{T}}(\boldsymbol{u}) \cdot \boldsymbol{n} = \boldsymbol{p} \text{ a.e. on } \Gamma_N \tag{2b}$$

$$\boldsymbol{u} \cdot \boldsymbol{n} \leq \phi \text{ a.e. on } \Gamma_C \tag{2c}$$

$$\boldsymbol{u}(t_0, x) = \boldsymbol{u}^0(x) \text{ a.e. in } \Omega \tag{2d}$$

$$\dot{\boldsymbol{u}}(t_0, x) = \dot{\boldsymbol{u}}^0(x) \text{ a.e. in } \Omega \tag{2e}$$

For simplicity, in the remainder we will set $\rho = 1$ for the mass density. Moreover, we define the external forces $\boldsymbol{F}_{\text{ext}} = \boldsymbol{F}_{\text{ext}}(t, x)$ as the sum of the volume forces and surface traction $\int_\Omega \boldsymbol{F}_{\text{ext}} \cdot \boldsymbol{u} := \int_\Omega \boldsymbol{F}(t, x) \cdot \boldsymbol{u} + \int_{\Gamma_N} \boldsymbol{p}(t, x) \cdot \boldsymbol{u}$ for any displacement $\boldsymbol{u}$. The internal forces $\boldsymbol{F}_{\text{int}} = \boldsymbol{F}(t, x, \boldsymbol{u})$ are given by $\int_\Omega \boldsymbol{F}_{\text{int}} \cdot \boldsymbol{u} := \int_\Omega \operatorname{div} \hat{\boldsymbol{T}}(\boldsymbol{u}) \cdot \boldsymbol{u}$ for any displacement $\boldsymbol{u}$. With this notation it is easy to see that the system to solve is exactly Newton's equation of motion under additional constraints at the contact boundary.

In order to discretize (2), we use Rothe's method which means that we discretize first in time then in space. The discretization in space will be carried out using linear finite elements. As already mentioned in the introduction, the discretization in time is not straightforward because even in the case of linear elastic materials classical time discretizations fail to handle the contact constraints appropriately. The classical Newmark scheme [11] is very common in continuum mechanics where the acceleration can be expressed in terms of displacements by means of Newton's equation of motion and the material law. But unfortunately it may evoke energy blow-ups and

oscillations at the contact boundary, see e.g., [10]. In [8], a purely implicit treatment of the contact forces was introduced for the case of linear elasticity. Here, we use an extension of this contact-implicit Newmark scheme to non-linear elastic materials.

Let $\tau > 0$ be the time-step size and let $t_i = t_0 + i\,\tau$ be the ith time step. Then, by denoting the approximation at $\boldsymbol{u}(t_i)$ with $\boldsymbol{u}^i$, we can write the classical Newmark scheme as

$$\begin{aligned}\boldsymbol{u}^{i+1} &= \boldsymbol{u}^i + \tau\dot{\boldsymbol{u}}^i + \frac{\tau^2}{2}((1-2\beta)\ddot{\boldsymbol{u}}^i + 2\beta\ddot{\boldsymbol{u}}^{i+1})\\ \dot{\boldsymbol{u}}^{i+1} &= \dot{\boldsymbol{u}}^i + \tau((1-\gamma)\ddot{\boldsymbol{u}}^i + \gamma\ddot{\boldsymbol{u}}^{i+1})\,.\end{aligned}$$

In the remainder we will choose $2\beta = \gamma = 1/2$. Here, $\dot{\boldsymbol{u}}^i$ denotes the approximation of the velocity at t_i; we furthermore define $\boldsymbol{F}^i_{\text{int}} = \boldsymbol{F}_{\text{int}}(t_i, x, \boldsymbol{u}^i)$, $\boldsymbol{F}^i_{\text{ext}} = \boldsymbol{F}_{\text{ext}}(t_i, x)$. In the unconstrained case, based on Newton's equation of motion, the accelerations $\ddot{\boldsymbol{u}}^i$ and $\ddot{\boldsymbol{u}}^{i+1}$ are replaced by the sum of the external and internal forces $\boldsymbol{F}^i_{\text{int}} + \boldsymbol{F}^i_{\text{ext}}$, and $\boldsymbol{F}^{i+1}_{\text{int}} + \boldsymbol{F}^{i+1}_{\text{ext}}$, respectively. In the constrained case, the second derivative of $\boldsymbol{u}$ in time may not exist everywhere, but the whole forces are known. They consist of the external, internal forces and the contact forces, which are the constraining forces to the contact constraints. Therefore when using the classical Newmark scheme for contact problems the accelerations are formally replaced by $\boldsymbol{F}^i_{\text{int}} + \boldsymbol{F}^i_{\text{ext}} + \boldsymbol{F}^i_{\text{con}}$ and $\boldsymbol{F}^{i+1}_{\text{int}} + \boldsymbol{F}^{i+1}_{\text{ext}} + \boldsymbol{F}^{i+1}_{\text{con}}$ for the time steps i and $i+1$. In contrast, in the case of the contact-implicit Newmark scheme we replace $\frac{1}{2}\ddot{\boldsymbol{u}}^i$ by $\frac{1}{2}(\boldsymbol{F}^i_{\text{int}} + \boldsymbol{F}^i_{\text{ext}})$ and $\frac{1}{2}\ddot{\boldsymbol{u}}^{i+1}$ by $\frac{1}{2}(\boldsymbol{F}^{i+1}_{\text{int}} + \boldsymbol{F}^{i+1}_{\text{ext}}) + \boldsymbol{F}^{i+1}_{\text{con}}$. A motivation for the use of this contact-implicit Newmark scheme can be found in [3], where it is shown that no energy blow-ups occur at least in the case of linear elasticity.

In order to solve the arising system, e.g., with the APTS scheme, cf. [7], we provide a formulation as minimization problem. On the basis of the stored energy function $\hat{W}$, we therefore define the energy function

$$\mathscr{E}^{i+1}(\boldsymbol{u}) = \int_\Omega \left(\hat{W}(x, \nabla\boldsymbol{u}) - \boldsymbol{F}^{i+1}_{\text{ext}} \cdot \boldsymbol{u}\right). \tag{3}$$

The derivative of the total energy $\frac{\partial}{\partial \boldsymbol{u}}\mathscr{E}^{i+1}(\boldsymbol{u})$ gives the sum of the external and internal forces. Furthermore, we define the set of admissible displacements $\mathscr{K} = \{\boldsymbol{u} \,|\, \boldsymbol{u}\cdot\boldsymbol{n} \le \phi \text{ a.e. on } \Gamma_C\}$.

In timestep t_{i+1}, the displacements $\boldsymbol{u}^{i+1}$ are computed as the solution of the following nonlinear and perhaps nonconvex minimization problem

$$\boldsymbol{u}^{i+1} \in \mathscr{K}: \qquad J^{i+1}(\boldsymbol{u}^{i+1}) = \min! \tag{4}$$

where

$$\begin{aligned}J^{i+1}(\boldsymbol{u}^{i+1}) := \int_\Omega &(\frac{1}{2}\boldsymbol{u}^{i+1} - \boldsymbol{u}^i - \tau\dot{\boldsymbol{u}}^i)\cdot\boldsymbol{u}^{i+1}\\ &+ \frac{\tau^2}{4}\left[\frac{\partial}{\partial \boldsymbol{u}^i}\mathscr{E}^i(\boldsymbol{u}^i)\right](\boldsymbol{u}^{i+1}) + \frac{\tau^2}{4}\mathscr{E}^{i+1}(\boldsymbol{u}^{i+1}).\end{aligned}$$

Therefore, the solution $\boldsymbol{u}^{i+1}$ of the minimization problem (4) satisfies

$$\boldsymbol{u}^{i+1} = \boldsymbol{u}^i + \tau \dot{\boldsymbol{u}}^i - \frac{\tau^2}{2}\left(\frac{1}{2}\frac{\partial}{\partial \boldsymbol{u}^i}\mathscr{E}^i(\boldsymbol{u}^i) + \frac{1}{2}\frac{\partial}{\partial \boldsymbol{u}^{i+1}}\mathscr{E}^{i+1}(\boldsymbol{u}^{i+1}) - \boldsymbol{F}_{\text{con}}(\boldsymbol{u}^{i+1})\right) \tag{5}$$

which is the first equation of the contact-implicit Newmark scheme; we refer also to [1].

To sum up, the above considerations lead to the following contact-implicit time discretization scheme

$$\boldsymbol{u}^{i+1} \in \mathscr{K} : \qquad J^{i+1}(\boldsymbol{u}^{i+1}) = \min! \tag{6a}$$

$$\dot{\boldsymbol{u}}^{i+1} = \dot{\boldsymbol{u}}^i - \tau\left(\frac{1}{2}\frac{\partial}{\partial \boldsymbol{u}^i}\mathscr{E}^i(\boldsymbol{u}^i) + \frac{1}{2}\frac{\partial}{\partial \boldsymbol{u}^{i+1}}\mathscr{E}^{i+1}(\boldsymbol{u}^{i+1}) - \boldsymbol{F}_{\text{con}}(\boldsymbol{u}^{i+1})\right) \tag{6b}$$

3 Efficient Solution of the Nonlinear Programming Problems

As pointed out before, the objective function connected to the spatial minimization problem (6a) in Sect. 2 might be nonconvex. In this case, a globalization strategy, such as a Trust-Region or a Linesearch strategy, might be employed in order to ensure the convergence for almost arbitrary initial iterates. Unfortunately, in particular in the context of large-scale problems arising from nonlinear elasticity, these strategies often tend to converge slowly, cf. [6]. Here, we therefore employ a novel nonlinear preconditioning strategy, which is designed to enhance the rates of convergence of these globalization strategies.

To this end, we consider a cheaply computable, nonlinear operator $\mathscr{F} : \mathbb{R}^n \to \mathbb{R}^n$ with the following property

$$\mathscr{F}(\boldsymbol{u}) \neq \boldsymbol{u} \qquad \Rightarrow \qquad J(\mathscr{F}(\boldsymbol{u})) < J(\boldsymbol{u})\,. \tag{7}$$

Let us emphasize that the nonlinear multigrid cycles in the RMTR method [6], in the MLS method [12] and also the asynchronous solution phase in the APTS method [7] or the PVD scheme [4] can be regarded as instances of our nonlinear preconditioning operator $\mathscr{F}$.

We remark that the following considerations are valid for arbitrary functions $\boldsymbol{u} : \mathbb{R}^n \to \mathbb{R}^n$. However, in this paper we only consider systems in elasticity, which after discretization with finite elements give rise to coefficient vectors in $\mathbb{R}^{3N}$, where N is the number of nodes, i.e., we have the special case $n = 3N$. For the ease of presentation, we thus again use boldface symbols for our solution $\boldsymbol{u}$ (as in the previous chapters).

In (7), J is a nonlinear objective function, such as J^i from (6a) in Sect. 2. In order to derive our nonlinear preconditioning scheme, we now consider the first-order conditions

$$\overline{\boldsymbol{u}} \in \mathbb{R}^n : \nabla J(\overline{\boldsymbol{u}}) = 0$$

of the following problem

$$\overline{\boldsymbol{u}} \in \mathbb{R}^n : J(\overline{\boldsymbol{u}}) = \min!$$

where J is a twice continuously differentiable and arbitrary nonlinear objective function. Similar to linear right preconditioning, we can reformulate the original first-order conditions to the following problem: for a given $\boldsymbol{u} \in \mathbb{R}^n$ compute $\boldsymbol{s} \in \mathbb{R}^n$ such that

$$\nabla J(\mathscr{F}(\boldsymbol{u} + \boldsymbol{s})) = 0 \tag{8a}$$

$$\overline{\boldsymbol{u}} = \mathscr{F}(\boldsymbol{u} + \boldsymbol{s}) \tag{8b}$$

which yields a *nonlinear right preconditioning strategy*. If we assume that $\mathscr{F}'$ exists, we might compute the search direction $\boldsymbol{s}$ employing Newton's method which gives rise to the following iterative scheme

$$\nabla^2 J(\mathscr{F}(\boldsymbol{u}^\nu))\mathscr{F}'(\boldsymbol{u}^\nu)\boldsymbol{s}^\nu = -\nabla J(\mathscr{F}(\boldsymbol{u}^\nu)) \tag{9a}$$

$$\boldsymbol{u}^{\nu+1} = \mathscr{F}(\boldsymbol{u}^\nu + \boldsymbol{s}^\nu) \tag{9b}$$

However, even if $\boldsymbol{s}^\nu$ is computed employing a globalization strategy convergence cannot be ensured, since

$$J(\mathscr{F}(\boldsymbol{u}^\nu)) \leq J(\boldsymbol{u}^\nu) \text{ and } J(\mathscr{F}(\boldsymbol{u}^\nu) + \boldsymbol{s}^\nu) \leq J(\mathscr{F}(\boldsymbol{u}^\nu)) \not\Rightarrow J(\mathscr{F}(\boldsymbol{u}^\nu + \boldsymbol{s}^\nu)) \leq J(\boldsymbol{u}^\nu)$$

In order to overcome this difficulty, we could compute a damping parameter $\alpha^\nu \in (0, 1]$ such that

$$J(\mathscr{F}(\boldsymbol{u}^\nu + \alpha^\nu \boldsymbol{s}^\nu)) \leq J(\boldsymbol{u}^\nu)\,. \tag{10}$$

This is possible, if both J and $\mathscr{F}$ are continuous, and if (7) and $\mathscr{F}(\boldsymbol{u}^\nu) \neq \boldsymbol{u}^\nu$ holds. In order to avoid the possibly costly use of a backtracking algorithm to determine α in (10), we linearize the update in (9b) leading to the following iterative scheme

$$\nabla^2 J(\mathscr{F}(\boldsymbol{u}^\nu))\mathscr{F}'(\boldsymbol{u}^\nu)\boldsymbol{s}^\nu = -\nabla J(\mathscr{F}(\boldsymbol{u}^\nu)) \tag{11a}$$

$$\boldsymbol{u}^{\nu+1} = \mathscr{F}(\boldsymbol{u}^\nu) + \mathscr{F}'(\boldsymbol{u}^\nu)\boldsymbol{s}^\nu \tag{11b}$$

This has the important advantage that if $\mathscr{F}'(\boldsymbol{u}^\nu)\boldsymbol{s}^\nu$ was computed employing a globalization strategy, e.g., as a Trust-Region step starting from $\mathscr{F}(\boldsymbol{u}^\nu)$, then the following inequalities hold

$$\begin{aligned} J(\mathscr{F}(\boldsymbol{u}^\nu)) \leq J(\boldsymbol{u}^\nu) \;\text{ and }\; & J(\mathscr{F}(\boldsymbol{u}^\nu) + \mathscr{F}'(\boldsymbol{u}^\nu)\boldsymbol{s}^\nu) \leq J(\mathscr{F}(\boldsymbol{u}^\nu)) \\ \Rightarrow\; & J(\mathscr{F}(\boldsymbol{u}^\nu) + \mathscr{F}'(\boldsymbol{u}^\nu)\boldsymbol{s}^\nu) \leq J(\boldsymbol{u}^\nu) \end{aligned}$$

Moreover, one can approximate $\mathscr{F}'(\boldsymbol{u}^\nu)\boldsymbol{s}^\nu$ by a quasi-Newton correction $\tilde{\boldsymbol{s}}^\nu$ as follows

$$B(\boldsymbol{u}^\nu)\tilde{\boldsymbol{s}}^\nu = -\nabla J(\mathscr{F}(\boldsymbol{u}^\nu)) \tag{12a}$$

$$\boldsymbol{u}^{\nu+1} = \mathscr{F}(\boldsymbol{u}^\nu) + \tilde{\boldsymbol{s}}^\nu \tag{12b}$$

where $B(\boldsymbol{u}^\nu)$ is a symmetric matrix. As a matter of fact, this *nonlinear preconditioning strategy* with *linearized update* can be regarded as a globalization strategy, as far as (7) holds and $\tilde{\boldsymbol{s}}^\nu$ in (12a) is computed employing a globalization strategy. Let us remark that for the sake of simplicity the above considerations have been made for unconstrained nonlinear programming problems. Similar results can also be obtained in the constrained case; the numerical results given here were computed employing the corresponding solution strategy for constrained problems.

4 Dynamic Simulation of a Can

In this section, we are interested in the simulation of an actual elastodynamic contact problem, as shown in Fig. 1. This is an elastodynamic contact problem subject to a nonlinear material law. For the material response, we employ the following polyconvex stored energy function, see e.g., [2],

$$W(x, \boldsymbol{C}) = 3(a+b) + (2a+4b)\cdot \operatorname{tr} \mathrm{E} + 2b\cdot(\operatorname{tr} \mathrm{E})^2 - 2b\cdot\operatorname{tr}(\mathrm{E}^2) + \Gamma(\det(\nabla\boldsymbol{\varphi})) \tag{13}$$

Here, we have set $\boldsymbol{C} = \boldsymbol{C}(\boldsymbol{u}) = (I + \nabla\boldsymbol{u})^T(I + \nabla\boldsymbol{u})$ (the right Cauchy-Green strain tensor), $\mathrm{E} = \mathrm{E}(\boldsymbol{u}) = \frac{1}{2}(\boldsymbol{C}(\boldsymbol{u}) - I)$ is the Green-St. Venant strain tensor, $\nabla\boldsymbol{\varphi} = I + \nabla\boldsymbol{u}$ is the deformation tensor and $\Gamma(\delta) = c\delta^2 - d\log\delta$ a logarithmic barrier function. For our example, the constants are chosen as follows

$$a = \mu + \frac{1}{2}\Gamma'(1),\ b = -\frac{\mu}{2} - \frac{1}{2}\Gamma'(1),\ c = -\frac{\lambda}{4} - \mu \text{ and } d = \frac{3\lambda}{4} + \mu \tag{14}$$

Here, we employed $\Gamma_D = \emptyset$, $\Gamma_N = \partial\Omega - \Gamma_C$ where $\Gamma_C = \{(x, y, z) | z = -0.5\}$ with natural boundary conditions. On the other hand, the initial velocity is given by $(\dot{\boldsymbol{u}}_0)_k = (0, 0, -0.05)$ for all nodes k, yielding a movement in direction of the obstacle. Initially the displacements are given by $\boldsymbol{u}_0 = \boldsymbol{0}$ and the gap between geometry and obstacle is slightly larger than zero. At all unknowns which are not related to Γ_C, we choose $\underline{\phi}_k = -10^6$ and $\overline{\phi}_k = 10^6$.

The simulation itself was carried out computing 1,000 timesteps with $\tau = 0.01$. The geometry is uniformly refined once giving rise to a nonlinear programming problem, equation (6), with approximately 54,000 unknowns for each timestep. The employed material parameters are $E = 1000$[MPa] and $\nu = 0.3$. Finally, the solution of the arising minimization problems (6a) in Sect. 2 was carried out efficiently

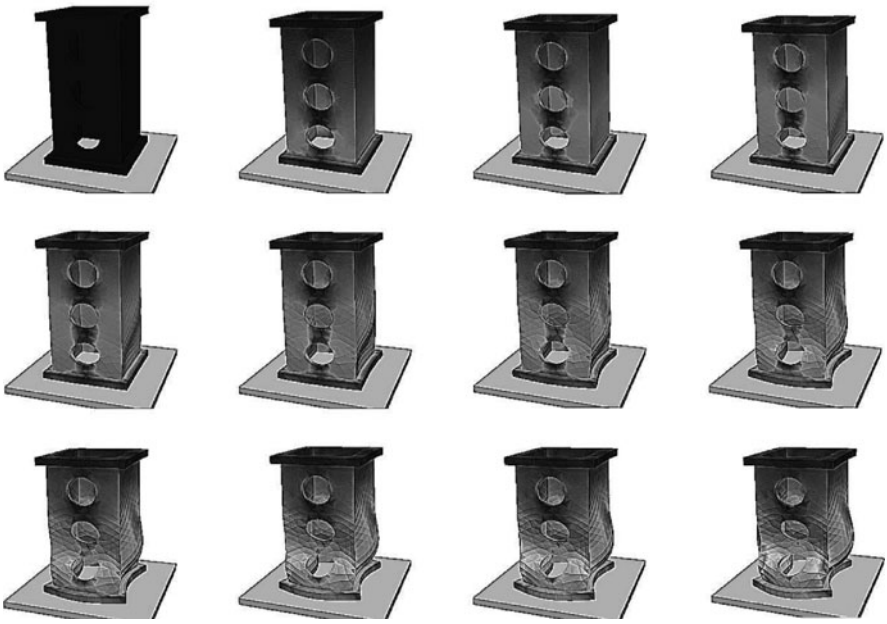


Fig. 1 Dynamic simulation of a can. Here, the solution of the problem of Sect. 4 is shown. As one can see, the can-like geometry moves in direction of the obstacle as indicated by the grey plane. Soon, the geometry and the obstacle stay in contact and the geometry's momentum yields the shown deformations. The last shown figure is the final configuration in this simulation. Different grey scales indicate the von-Mises stresses where light grey represents strong local stresses

employing the RMTR method [6] in combination with the APTS method [7] applied on the finest level.

Acknowledgements The research of Christian Groß and Mirjam Walloth was supported by the Bonn International Graduate School and the Hausdorff Center for Mathematics.

References

1. Bathe K. J. and Mirza, N. I. B. On a composite implicit time integration procedure for nonlinear dynamics, Computers and Structures, 83: 2513–2524 (2005)
2. Ciarlét, P. G.: Mathematical elasticity, volume I: Three-dimensional elasticity. Studies in Mathematics and its Applications, 20(186):715–716 (1988)
3. Deuflhard P., Krause R., and Ertel S.: A contact-stabilized Newmark method for dynamical contact problems. International Journal for Numerical Methods in Engineering, 73(9):1274–1290 (2008)
4. Ferris M. C. and Mangasarian O. L.: Parallel variable distribution. SIAM Journal on Optimization, 4(4):815–832 (1994)

5. Gratton S., Sartenaer A., and Toint P.L.: Recursive trust-region methods for multiscale nonlinear optimization. SIAM Journal on Optimization, 19(1):414–444 (2008)
6. Groß C. and Krause R.: On the convergence of recursive trust-region methods for multiscale non-linear optimization and applications to non-linear mechanics. SIAM Journal on Numerical Analysis, 47(4):3044–3069 (2009)
7. Groß C. and Krause R.: A new class of non-linear additively preconditioned Trust-Region strategies: Convergence results and applications to non-linear mechanics. INS preprint 904, Institute for Numerical Simulation, University of Bonn, 03 (2009)
8. Kane C., Repetto E., Ortiz M. and Marsden J.: Finite element analysis of nonsmooth contact. Computer Methods in Applied Mechanics and Engineering, 180:1–26 (1999)
9. Krause R. and Mohr C.: Level set based multi-scale methods for large deformation contact problems. ICS Preprint 2009-01, Institute of Computational Science, University of Lugano (2009). Submitted to Applied Numerical Mathematics
10. Krause R. and Walloth M.: Presentation and Comparison of Selected Algorithms for Dynamic Contact Based on the Newmark Scheme. ICS Preprint 2009-08, Institute of Computational Science, University of Lugano (2009)
11. Newmark N. M.: A method of computation for structural dynamics. Journal of Engineering Mechanics Division, 8 (1959)
12. Wen Z. and Goldfarb D.: Line search multigrid methods for large-scale non-convex optimization. Technical report, IEOR Columbia University (2008)

An Implementation Framework for Solving High-Dimensional PDEs on Massively Parallel Computers

Magnus Gustafsson and Sverker Holmgren

Abstract Accurate solution of time-dependent, high-dimensional PDEs requires massive-scale parallel computing. In this paper, we describe an implementation framework for block-decomposed structured grids and discuss techniques for optimization on clusters where the nodes have one or more multicore processors. We use a two-level parallelization scheme with message passing between nodes and multithreading within each node, and argue that this is the best compromise for memory efficiency. We present some examples where the time-dependent Schrödinger equation is solved.

1 Introduction

Numerical simulation of PDEs is a tool of great importance to better understand and predict the outcome of experiments, and to understand the world around us. For many realistic scenarios, such mathematical models lead to high-dimensional problem settings which must be solved using accurate numerical methods. To solve the corresponding computational problems, we need sophisticated parallel numerical algorithms, massive-scale parallel computers and optimized parallel implementations.

In this paper, we present an on-going effort towards developing a parallel implementation framework for adaptive solution of high-dimensional, time-dependent PDEs on large-scale distributed clusters where the nodes are built on multicore processors. Fundamental features of the framework is a block-decomposition of a structured spatial grid and the use of high-order finite difference stencils for the spatial discretization. A two-level parallelization scheme is implemented, where

M. Gustafsson (✉) and S. Holmgren
Department of Information Technology, Uppsala University Box 337, 751 05 Uppsala, Sweden
e-mail: magnus.gustafsson@it.uu.se, sverker.holmgren@it.uu.se

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_44,

message passing (MPI) is used for global communication across nodes and multi-threading (OpenMP) is used for work-sharing within each node. This is not the best alternative for run-time performance, but in the current implementation it leads to less explicit replication of memory and hence better memory utilization.

As a case study, we solve the time-dependent Schrödinger equation (TDSE) within the context of quantum chemistry, modelling the dynamics of molecules at the granularity of wave packets describing the atomic nuclei. Here, the dimensionality of a problem is given by $d = 3N-6$, where N is the number of nuclei involved in the interaction. Explicitly time-dependent events (e.g. interaction with laser pulses and particles colliding) combined with high demands on numerical accuracy make these problems extremely demanding to solve.

2 Parallel Computing Using Clusters of Multicore Processors

Up until a few years ago, microprocessor clock frequencies have constantly been increased. Due to this, programmers have been able to develop sequential programs that have executed faster and faster on new processor models. However, physical and technological limitations have put an end to this development. Instead, the architects of microprocessors seek to achieve increased efficiency by incorporating several processing cores that work in parallel on a single die. Multicore chips of today use a handful of cores but within the coming years, this number will increase significantly.

When clusters are composed of nodes with multicore processors, there will be a multi-level hierarchy among the components in the system. At the topmost level we have the distributed, parallel system of nodes. Each node in the cluster might have several processor sockets, and within each processor chip the cores may be arranged in a non-uniform fashion. The processors in a single node usually share memory logically, but the memory modules might be separate and distributed physically on the board, yielding a NUMA (Non-Uniform Memory Access) architecture. Furthermore, on-chip caches will be local to each chip and possibly organized in a hierarchichal structure within the chip as well.

Multicore processors introduce new challenges to programmers of parallel applications, partly due to the hierarchical system layout but also due to physical constraints in the chips. First of all, the bandwidth of the memory bus is limited and constitutes a major bottleneck for off-chip memory traffic. When the number of cores on a chip increases, this situation gets worse since the amount of bandwidth per core decreases correspondingly. Second, the physical space on a chip that can be dedicated to cache memory is limited and the effective amount of cache per core will decrease when the number of cores on a chip increases. In order to address these issues, we strive to find techniques that minimize communication to and from the chip, and maximize data reuse.

3 A PDE-Solver Framework

The target problems of our framework are d-dimensional, time-dependent linear PDEs

$$\frac{du}{dt} = P(x,t)u. \tag{1}$$

We use the *method of lines* approach and first discretize in space using a d-dimensional tensor-product grid, i.e., each grid point can be uniquely identified by a d-vector of integers $(j_1, \ldots, j_d)$. This transforms the PDE into a large system of ODEs, which we solve by integration in time using some suitable ODE solver.

The spatial derivatives are discretized using high-order finite difference stencils. Our implementation allows for any order of derivative to be discretized, using any combination of weights. For an arbitrary order of accuracy, the stencil coefficients can be computed by implementing the simple procedure presented by Fornberg [2]. A standard choice is to use $2p$-order centered stencils, where p can be adjusted according to the desired accuracy.

4 An Example: The Time-Dependent Schrödinger Equation

A general quantum dynamics problem can be described by the TDSE,

$$i\hbar\frac{\partial}{\partial t}\psi(\mathbf{r},t) = \hat{H}\psi(\mathbf{r},t), \tag{2}$$

where $\psi(\mathbf{r},t)$ is the wave function describing the probability distribution of the system and $\hat{H}$ is the linear Hamiltonian operator describing the total energy.

For a particle of mass m,

$$\hat{H} = \hat{T} + \hat{V} = -\frac{\hbar^2}{2m}\nabla^2 + V(\mathbf{r},t), \tag{3}$$

where $\hat{T}$ and $\hat{V}$ are the kinetic and potential energy operators respectively. Hence, the kinetic energy has constant coefficients, while the potential energy may contain coefficients that depend on both space and time.

By introducing the discretization in space described in Sect. 3, we obtain the following system of ODEs

$$i\hbar\frac{d}{dt}\Psi(t) = H(t)\Psi(t), \qquad \Psi(0) = \Psi^{(0)}, \tag{4}$$

where $H(t)$ is the Hamiltonian matrix of size $N \times N$ with

$$N = \prod_{L=1}^{d} N_L, \qquad L = 1, \ldots, d. \tag{5}$$

If the Hamiltonian is independent of time, the propagation of the system of ODEs in (4) is given by

$$\Psi^{(k+1)} = e^{-i\,\Delta t H}\,\Psi^{(k)}. \tag{6}$$

In our experiments, the Hamiltonian will be explicitly time-dependent and we cannot express the exact time-evolution of (4) in a simple manner. However, a second order accurate approximation is given by the exponential midpoint rule [6]

$$\Psi^{(k+1)} = e^{-i\,\Delta t H(t_k + \frac{\Delta t}{2})}\Psi^{(k)}. \tag{7}$$

Solving (7) involves exponentiating the Hamiltonian matrix, which is very large according to (5). Commonly used techniques for computing the matrix exponential are not feasible since these methods in general require $O(N^3)$ operations [7]. Instead, we use the Lanczos method, an iterative Krylov subspace method that computes approximations of only a few of the extremal eigenvalues.

5 Parallelization

Our implementation is parallelized on two levels; between nodes via message passing (MPI) and within each node using OpenMP for multi-threading. The spatial grid is decomposed into blocks and in the current pilot implementation, each computing node solves the local problem within a single block, exchanging information as necessary with other nodes.

Currently, the implementation framework applies a static, equidistant block-grid decomposition with equally sized blocks. Since the computational workload per grid element is constant, the workload among the nodes is well-balanced. Similarly, the work within a block is divided statically between the available threads in the node in equal chunks for low synchronization overhead between the threads. We use a static number of threads, which are spawned early on and maintained throughout the execution of the program. This minimizes the overhead from runtime management of the threads and reduces thread migration between cores.

Care must be taken when applying the finite difference stencils in many dimensions; the data access strides are very large and keeping all data in cache is not possible. We need a more elaborate implementation of the stencil operator than the straightforward, naïve approach of simply stepping through the grid dimension-by-dimension. Here, we use cache-tiling, where the data accesses form tiles in the lower dimensions that are streamed from memory into the caches. This has proven to be a useful approach on modern processors since it makes use of the prefetch mechanisms that are prevalent in new architectures [1].

For the Lanczos algorithm it is difficult to achieve massive scalability because of the three synchronization points in each iteration; two inner products of size N and one multiplication of the Hamiltonian with the wave vector. The inner product is a global operator involving all processes in the communication, whereas the multiplication with the Hamiltonian matrix only requires communication between nearest-neighbor processes.

Inner products: To improve scalability, we have adopted the approach in [5], where the Lanczos procedure has been restructured in order to bring the two inner products together so that they can be computed simultaneously. This effectively removes one synchronization point in the algorithm. In exact arithmetics the reformulated algorithm has been proven to yield the same results as the standard Lanczos method [5]. However, when used as an iterative algorithm for computing eigenvalues, run to convergence, this algorithm is known to sometimes exhibit stability problems. In our case, we execute a small predetermined number of iterations, and we do not encounter any problems with stability.

Multiplication with the Hamiltonian matrix: Close to the borders of a block, the finite difference stencils depend on function values located in neighboring blocks. These "remote" function values are stored in buffer regions, referred to as ghost cells, at the boundaries of the blocks. Thus, the data values stored in the ghost cells are replicated, containing the same data items as the corresponding entries in the neighboring blocks. Because of the high dimensionality of our target problems, the ghost cell blocks will be large and replicating them all will lead to a significant memory overhead. Our approach to this problem is to communicate ghost data in one dimension at a time and reuse the allocated buffers, which will lead to a d-fold reduction of the replicated memory. Furthermore, using multi-threading within each process reduces the number of ghostcell blocks, since this does not require explicit replication for communication.

We overlap communication delays with computation as follows. First, ghost data are sent (non-blocking) in the first dimension. While waiting for the data to arrive, each node computes all values possible without considering any ghost data. Once the first round of data has arrived, the next round is sent off (again, non-blocking) and while waiting, each node fills in the values that required the first set of data items. This is repeated until all dimensions are completed.

6 Results

In this paper we focus on the parallel performance of our implementation. However, for completeness, we present some results of the numerical accuracy of the spatial and temporal discretizations in Fig. 1. For a more detailed analysis of the numerical properties, see [4]).

We have run experiments on two clusters, *Grad* and *Isis*. The hardware details are listed in Table 1. The two clusters are similar in structure, with a number of distributed nodes interconnected by a standard switched Gigabit Ethernet network. Each node is configured with dual CPU-sockets and 2 GB of DRAM per core. The *Grad*-cluster is fitted with quad-core processors where the cores are configured in pairs. Each pair of cores share a 6 MB L2 cache which implies a hierarchy among the cores within a chip. L1 caches are private to each core, with separate 32 kB caches for the instruction- and data streams. The processors of the *Isis*-cluster are dual-core but each core has its private L2 cache of 1 MB, albeit significantly smaller than the 3 MB that is available per core on the *Grad*-cluster.

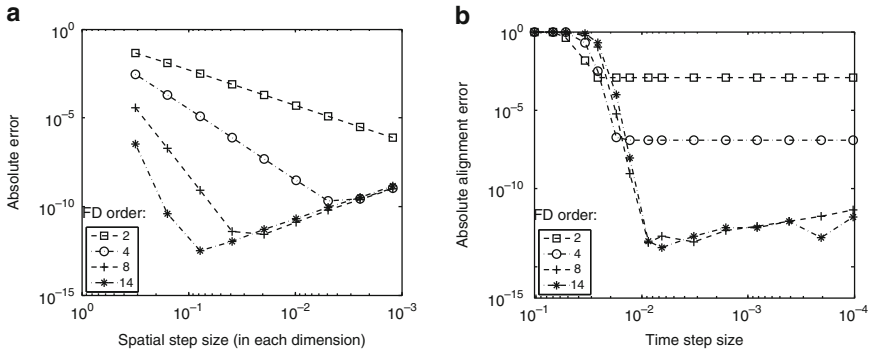


Fig. 1 Accuracy results for a 2D problem. (**a**) (*left*) The hamiltonian in (3) is applied once on a Gauss pulse with the potential (V) set to zero, and the result is compared to the analytic second derivative. (**b**) (*right*) One period of a simple harmonic oscillator is computed, using a grid of 240^2 elements and the Lanczos algorithm with 4 Lanczos-iterations. The final wave function is compared to the starting vector, taking the maximum absolute error

Table 1 Overview of the clusters that have been used for the computations

	Grad	Isis
No. nodes	64	200
No. CPUs/node	2	2
DRAM/node	16 GB	8 GB
Interconnect	Gb Ethernet	Gb Ethernet
CPU	Intel Xeon E5430	AMD Opteron 2220 SE
No. cores	4	2
Clock rate	2.66 GHz	2.8 GHz
L1-cache (I/D)	32 kB/32 kB	64 kB/64 kB
L2-cache	2 × 6 MB	2 × 1 MB

To analyze the performance of our implementation, we first look at how well it scales with problem size and the number of available cores. For these results we used the *Grad* cluster. The speed-up results from parallelizing a single block within a node with OpenMP are shown in Fig. 2a. We see that the speed-up depends heavily on the block size and that the best speed-up (about $6.5x$ on 8 cores) is obtained for blocks of 64^3 grid elements. This number is tightly coupled to the number of grid elements that fit in the cache; when the number of grid points is too large we get problems with data conflicts in the caches and when the number of grid points is too small there is not enough parallelism in the blocks. Thus, we gain performance by carefully choosing the size of the blocks so that they fit the caches well.

In Fig. 2b the performance of the hybrid MPI-OpenMP implementation is presented for three different problem sizes. Apparently, the larger the problem size, the better the speed-up, which is often the case with MPI-implementations since there is more computations to overlap the communication delays. Asymptotically, the speed-up seems to approach the optimal as the problem size grows. Important to

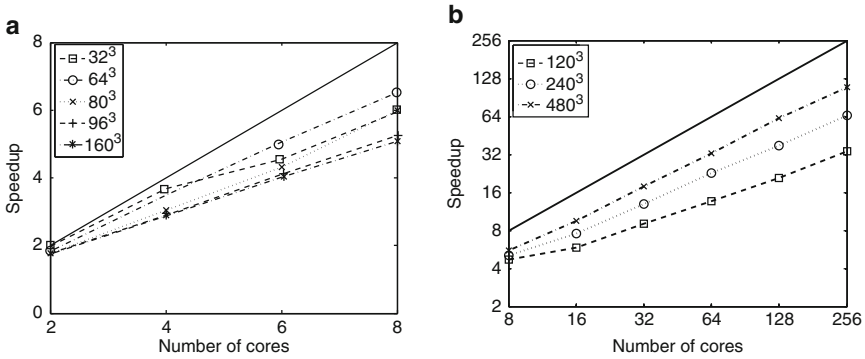


Fig. 2 Speedup of the implementation for fixed-size 3D problems. The *solid line* indicates ideal speedup. For these problems we used 10 Lanczos iterations. (**a**) (*left*) Single node speedup, using only OpenMP. (**b**) (*right*) Speedup using the hybrid parallelization scheme with MPI and OpenMP

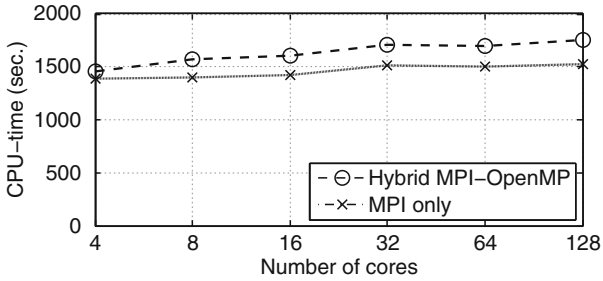


Fig. 3 Scaled speedup for the hybrid MPI-OpenMP implementation compared to using MPI only

note here is that the performance increase of the system as a whole is rather independent of the speed-up numbers we saw for the local parallelization. This is due to the fact that inter-node communication over the slow network is more penalizing for performance and hides the low utilization of each node.

In practice, when having more computing resources at hand we take advantage of this by solving larger problems rather than solving a fixed-size problem faster. A performance metric that better reflects this practical aspect is scaled speed-up [3]. In the final experiment, we first solve a problem on a single node, then double the size of the problem and solve it on two nodes and so forth. Ideally, for a perfectly linear increase in performance, we want the execution time to be constant for all problem sizes. Figure 3 shows the scaled speed-up when run on at most 32 nodes in the *Isis*-cluster. Each node solves a fixed size problem of 240^3 grid elements. We compare the results of the MPI-OpenMP implementation to using only MPI for parallelization; the MPI-only implementation has a slight performance advantage but using this type of implementation will also lead to more significant memory waste due to a larger fraction of ghost-cells. In the two-level parallelization scheme, inner products are computed in two steps, requiring first a local synchronization

point and then a global synchronization point. Thus, the synchronization overhead of computing the inner products might be slightly higher in the hybrid implementation.

7 Conclusion and Outlook

Numerical simulation of high-dimensional PDEs is computationally intensive and requires efficient techniques and massive-scale computing. In this paper we present our achievements so far in addressing this challenge. Among our future plans are:

s-step Lanczos: Kim and Chronopoulos [5] describe another optimization to the Lanczos algorithm, referred to as the s-step approach. Here, several Lanczos iterations are combined in a single step, dramatically reducing the number of global synchronization points. One iteration of the s-step Lanczos algorithm corresponds to s iterations of the standard algorithm and all the $2s$ inner products of these s iterations are computed at once. This seems promising for our purposes, but again the stability properties need to be carefully considered (cf. [8]).

Spatial adaptivity: For high-dimensional problems, the static, equidistant spatial discretization will lead to prohibitively large grids. Generally, the resolution that is required for a desired level of accuracy varies in different parts of the grid and having overly fine grid resolution is a waste of computational effort. We plan to implement a block-adaptive algorithm that operates on fixed-size blocks where the block-size is chosen to match the size of the caches at a suitable level. Apart from making the implementation more complex, this type of algorithm also makes load balancing much more demanding.

References

1. Datta, K., Kamil, S., Williams, S., Oliker, L., Shalf, J., Yelick, K.: Optimization and Performance Modeling of Stencil Computations on Modern Microprocessors, SIAM Rev., **51**, 129–159 (2009)
2. Fornberg, B.: Calculation of Weights in Finite Difference Formulas, SIAM Rev., **40**, 685–691 (1998)
3. Gustafson, J.L.: Reevaluating Amdahl's law, Commun. ACM, **31**, 532–533 (1988)
4. Gustafsson, M.: A PDE Solver Framework Optimized for Clusters of Multi-core Processors, Master's thesis UPTEC F09 004, School of Engineering, Uppsala University (2009)
5. Kim, S.K., Chronopoulos, A.T.: A Class of Lanczos-like Algorithms Implemented on Parallel Computers, Parallel Comput., **17**, 763–778 (1991)
6. Lubich, C.: Integrators for Quantum Dynamics: A Numerical Analyst's Brief Review, In: Grotendorst, J., Marx, D., Muramatsu, A. (eds.) Quantum Simulations of Complex Many-Body Systems: From Theory to Algorithms, vol 10, pp. 459–466. John von Neumann Institute for Computing, Jülich (2002)
7. Moler, C., van Loan, C.: Nineteen Dubious Ways to Compute the Exponential of a Matrix, Twenty-Five Years Later, SIAM Rev., **45**, 3–49 (2003)
8. van der Vorst, H.A.: Iterative Krylov Methods for Large Linear Systems, Cambridge University Press, London (2003)

Benchmarking FE-Methods for the Brinkman Problem

Antti Hannukainen, Mika Juntunen, and Rolf Stenberg

Abstract Various finite element families for the Brinkman flow (or Stokes–Darcy flow) are tested numerically. Particularly the effect of small permeability is studied. The tested finite elements are the MINI element, the Taylor–Hood element, and the stabilized equal order methods.

1 Introduction

The Brinkman equations are used in modeling porous media flow in the case of high porosity when shear effects of the fluid has to be taken into account, see e.g., [1–3, 16, 20].

In a recent paper [15] we have studied the finite element approximation of the model. We have proved both a priori and a posteriori estimates for some classes of methods that (in view of the analysis) are robust. The purpose here is to give numerical realizations of the theory in [15].

The Brinkman equations have been studied before e.g., in [17–19]. In [19] the mathematical setting, namely norms and solution spaces, are different. In [17, 18] the norms are similar to our work but we present also less abstract and thus computationally more appealing counterparts to the norms.

In next section we recall the scaled form of the Brinkman equations and the mathematical structure of the problem. Section 3 is devoted to the finite element approximations. We shortly recall the results of [15] and the methods presented therein. We also give the corresponding results for the so-called Taylor–Hood family. The main part of the paper is Sect. 4 in which we give the results of benchmark computations.

A. Hannukainen (✉), M. Juntunen, and R. Stenberg
Department of Mathematics and Systems Analysis, Helsinki University of Technology,
P.O. Box 1100, 02015 TKK, Finland
e-mail: antti.hannukainen@tkk.fi, mika.juntunen@tkk.fi, rolf.stenberg@tkk.fi

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_45,

2 The Brinkman Problem

The scaled version of the Brinkman equations is: Find $\boldsymbol{u}$ and p such that

$$-t^2 \boldsymbol{A}\boldsymbol{u} + \boldsymbol{u} + \nabla p = \boldsymbol{f} \text{ in } \Omega, \tag{1}$$

$$\mathbf{div}\boldsymbol{u} = g \text{ in } \Omega, \tag{2}$$

where the parameter $0 \leq t \leq C$. For $t = 0$ we have the Darcy equations, for which we consider the natural boundary condition

$$\boldsymbol{u} \cdot \boldsymbol{n}|_{\partial\Omega} = 0. \tag{3}$$

For $t > 0$ we have Dirichlet boundary conditions

$$\boldsymbol{u}|_{\partial\Omega} = \mathbf{0}. \tag{4}$$

For compatibility, assume $g \in L_0^2(\Omega)$ and to get a unique pressure assume also $p \in L_0^2(\Omega)$. When $t \approx 1$ the problem is a Stokes problem. For "small" t the problem is a singular perturbation of the Darcy equations.

The natural norm for the velocity is

$$\|\boldsymbol{v}\|_t^2 = t^2 \|\boldsymbol{\varepsilon}(\boldsymbol{v})\|_0^2 + \|\boldsymbol{v}\|_0^2. \tag{5}$$

Hence, for $t = 0$ the space for the velocity is $[L^2(\Omega)]^N$, and for $t > 0$ (by Korn's inequality) $[H_0^1(\Omega)]^N$. By defining

$$b(\boldsymbol{v}, q) = \begin{cases} -(\mathbf{div}\boldsymbol{v}, q) & \text{for } t > 0 \\ (\boldsymbol{v}, \nabla q) & \text{for } t = 0, \end{cases} \tag{6}$$

the norm for the pressure is

$$|\!|\!| q |\!|\!|_t = \sup_{v \in V} \frac{b(\boldsymbol{v}, q)}{\|\boldsymbol{v}\|_t}, \tag{7}$$

and the solution space is

$$Q = \{ q \in L_0^2(\Omega) \mid |\!|\!| q |\!|\!|_t < \infty \}. \tag{8}$$

Note that for $t = 0$ we have

$$|\!|\!| q |\!|\!|_t \equiv \|\nabla q\|_0, \tag{9}$$

whereas for $0 < t \leq C$ the Babuška–Brezzi inequality yields

$$C_1 \|q\|_0 \leq |\!|\!| q |\!|\!|_t \leq C_2 t^{-1} \|q\|_0. \tag{10}$$

Defining the bilinear forms

$$a(\boldsymbol{u},\boldsymbol{v}) = t^2\,(\boldsymbol{\varepsilon}(\boldsymbol{u}),\boldsymbol{\varepsilon}(\boldsymbol{v})) + (\boldsymbol{u},\boldsymbol{v})\,, \tag{11}$$

$$\mathscr{B}(\boldsymbol{u},\,p;\boldsymbol{v},q) = a(\boldsymbol{u},\boldsymbol{v}) + b(\boldsymbol{v},\,p) + b(\boldsymbol{u},q), \tag{12}$$

and the linear functional

$$\mathscr{L}(\boldsymbol{v},q) = (\boldsymbol{f}\,,\boldsymbol{v}) - (g,q)\,, \tag{13}$$

the weak formulation of the problem is: Find $(\boldsymbol{u},\,p) \in \boldsymbol{V}\times Q$ such that

$$\mathscr{B}(\boldsymbol{u},\,p;\boldsymbol{v},q) = \mathscr{L}(\boldsymbol{v},q) \quad \forall(\boldsymbol{v},q) \in \boldsymbol{V}\times Q. \tag{14}$$

This is a saddle point problem and Brezzi's conditions imply the stability

$$\sup_{(\boldsymbol{v},q)\in \boldsymbol{V}\times Q} \frac{\mathscr{B}(\boldsymbol{w},r;\boldsymbol{v},q)}{\|\boldsymbol{v}\|_t + |\!|\!|q|\!|\!|_t} \geq C\left(\|\boldsymbol{w}\|_t + |\!|\!|r|\!|\!|_t\right) \quad \forall(\boldsymbol{w},r) \in \boldsymbol{V}\times Q \tag{15}$$

by which the solution is unique.

3 Finite Elements and A Priori Error Estimates

The fact that the Brinkman model covers a whole range of problems, from Darcy to Stokes, has some consequences. For the Darcy problem a balanced method uses $P_k - P_{k-1}$ polynomials for the pressure and velocity, respectively. For the pure Stokes problem (with $t \approx 1$) it is the opposite, P_k for the velocity and P_{k-1} for the pressure. Hence, to obtain a method good for all values of t it seems natural to use equal order interpolation. Families of this kind are analyzed in our paper [15]. Here we recall the results and also show the results for the well-known Taylor–Hood family of Stokes element.

We assume a partitioning $\mathscr{C}_h$ of the domain Ω into simplices. With $K \in \mathscr{C}_h$ we denote an element of the partitioning, and the maximum size of $K \in \mathscr{C}_h$ is denoted by h. With Γ_h we denote the boundary edges of the partitioning.

In the following the discrete counterpart of the pressure norm (7) is utilized;

$$|\!|\!|q|\!|\!|_{t,h}^2 = \sum_{K\in\mathscr{C}_h} \frac{h_K^2}{t^2+h_K^2}\|\nabla q\|_{0,K}^2. \tag{16}$$

This norm has the advantage that it can be explicitly computed.

3.1 The Family Generalizing the MINI Element

For this family, generalizing the well-known MINI element of Arnold, Brezzi and Fortin [4]. The finite element spaces are

$$\boldsymbol{V}_h = \{\boldsymbol{v} \in [C(\Omega)]^N \cap \boldsymbol{V} \mid \boldsymbol{v}|_K \in [P_k(K) \cup B_{k+N}(K)]^N\}, \tag{17}$$

$$Q_h = \{q \in C(\Omega) \cap L_0^2(\Omega) \mid q|_K \in P_k(K)\}, \tag{18}$$

where $P_k(K)$ denotes the polynomials of degree k and $B_{k+N}(K) = P_{k+N}(K) \cap H_0^1(K)$ are the bubbles of degree $k+N$.

The finite element formulation is: Find $(\boldsymbol{u}_h, p_h) \in \boldsymbol{V}_h \times Q_h$ such that

$$\mathscr{B}(\boldsymbol{u}_h, p_h; \boldsymbol{v}, q) = \mathscr{L}(\boldsymbol{v}, q) \quad \forall (\boldsymbol{v}, q) \in \boldsymbol{V}_h \times Q_h. \tag{19}$$

The stability and a priori results are shown in [15]. In the case of a smooth solution and a quasiuniform mesh we get the estimate

$$\|\boldsymbol{u} - \boldsymbol{u}_h\|_t + |\!|\!| p - p_h |\!|\!|_{t,h} \leq C\left((t+h)h^k \|\boldsymbol{u}\|_{k+1} + (t+h)^{-1} h^{k+1} \|p\|_{k+1}\right). \tag{20}$$

Hence, we get a uniform convergence (with respect to t) of $\mathscr{O}(h^k)$.

3.2 Stabilized Methods

The linear stabilized method was introduced by Brezzi and Pitkäranta [10] and then generalized by Hughes and Franca [14]. In [15] we analyze the method using the techniques developed in [12, 13].

The method uses pure piecewise polynomials of equal degree:

$$\boldsymbol{V}_h = \{\boldsymbol{v} \in [C(\Omega)]^N \cap \boldsymbol{V} \mid \boldsymbol{v}|_K \in [P_k(K)]^N\}, \tag{21}$$

$$Q_h = \{q \in C(\Omega) \cap L_0^2(\Omega) \mid q|_K \in P_k(K)\}. \tag{22}$$

The stabilized method is then: Find $(\boldsymbol{u}_h, p_h) \in \boldsymbol{V}_h \times Q_h$ such that

$$\mathscr{B}_h(\boldsymbol{u}_h, p_h; \boldsymbol{v}, q) = \mathscr{L}_h(\boldsymbol{v}, q) \quad \forall (\boldsymbol{v}, q) \in \boldsymbol{V}_h \times Q_h, \tag{23}$$

with

$$\mathscr{B}_h(\boldsymbol{u}_h, p_h; \boldsymbol{v}, q) = \mathscr{B}(\boldsymbol{u}_h, p_h; \boldsymbol{v}, q) \tag{24}$$

$$- \alpha \sum_{K \in \mathscr{C}_h} \frac{h_K^2}{t^2 + h_K^2} \left(t^2 \boldsymbol{A}\boldsymbol{u}_h - \boldsymbol{u}_h - \nabla p_h, t^2 \boldsymbol{A}\boldsymbol{v} - \boldsymbol{v} - \nabla q\right)_K$$

and

$$\mathscr{L}_h(\boldsymbol{v}, q) = \mathscr{L}(\boldsymbol{v}, q) - \alpha \sum_{K \in \mathscr{C}_h} \frac{h_K^2}{t^2 + h_K^2} \left(\boldsymbol{f}, t^2 \boldsymbol{A}\boldsymbol{v} - \boldsymbol{v} - \nabla q \right)_K, \tag{25}$$

with a parameter $\alpha > 0$.

For consistency, assume

$$t^2 \boldsymbol{A}\boldsymbol{u} - \boldsymbol{u} - \nabla p = \boldsymbol{f} \in [L^2(\Omega)]^2. \tag{26}$$

Admissible values for the stability parameter α depend on the constant C_I of the following inverse inequality

$$h_K^2 \|\boldsymbol{A}\boldsymbol{w}\|_{0,K}^2 \leq C_I \|\nabla \boldsymbol{w}\|_{0,K}^2 \quad \forall \boldsymbol{w} \in [P_k(K)]^N. \tag{27}$$

In the range $0 < \alpha < \min\{1/(2C_I), 1/2\}$ the method is stable and for a smooth solution and a quasiuniform mesh we again get the uniform $\mathscr{O}(h^k)$ estimate (20).

3.3 The Taylor–Hood Family

The third method to be considered is the Taylor–Hood family with the finite element subspaces

$$\boldsymbol{V}_h = \{\boldsymbol{v} \in [C(\Omega)]^N \cap \boldsymbol{V} \mid \boldsymbol{v}|_K \in [P_{k+1}(K)]^N\}, \tag{28}$$

$$Q_h = \{q \in C(\Omega) \cap L_0^2(\Omega) \mid q|_K \in P_k(K)\}. \tag{29}$$

The finite element formulation is: Find $(\boldsymbol{u}_h, p_h) \in \boldsymbol{V}_h \times Q_h$ such that

$$\mathscr{B}(\boldsymbol{u}_h, p_h; \boldsymbol{v}, q) = \mathscr{L}(\boldsymbol{v}, q) \quad \forall (\boldsymbol{v}, q) \in \boldsymbol{V}_h \times Q_h. \tag{30}$$

For the Stokes problem ($t \approx 1$) this method has been proved to be optimal both in two and three space dimensions [6–9, 11, 21–23]. By established techniques the analysis can be carried over to the present case.

For this family the assumption of a quasiuniform mesh and a smooth solution gives the estimate

$$\|\boldsymbol{u} - \boldsymbol{u}_h\|_t + |\|p - p_h\||_{t,h} \leq C \left((t + h) h^{k+1} \|\boldsymbol{u}\|_{k+2} + (t+h)^{-1} h^{k+1} \|p\|_{k+1} \right). \tag{31}$$

From here we see that also for this method we have $\mathscr{O}(h^k)$ convergence rate uniformly with respect to t. Only for the Stokes limit with $t \approx 1$ we have a $\mathscr{O}(h^{k+1})$ convergence rate. In the Darcy limit $t = 0$ the two terms are not in balance; $\mathscr{O}(h^{k+2})$ for the velocity but only $\mathscr{O}(h^k)$ for the pressure.

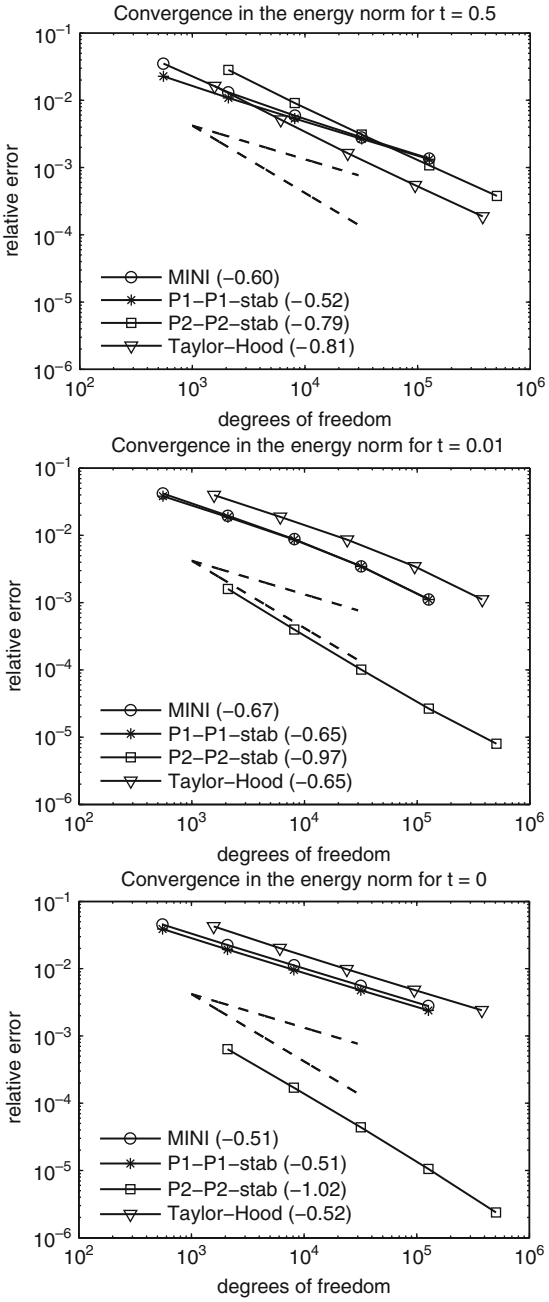


Fig. 1 Convergence of the finite element solutions in the energy norm using uniform refinement. The value in the brackets is the average rate of convergence; values -0.5 and -1.0 correspond to $\mathcal{O}(h)$ and $\mathcal{O}(h^2)$ rates of convergence. The *dashed lines* are reference slopes of $\mathcal{O}(h)$ and $\mathcal{O}(h^2)$ convergence. On the *top* problem is of the Stokes type and the two *below* are of the Darcy type

4 Numerical Examples

The finite element methods tested here are the lowest order MINI and Taylor–Hood elements, and the stabilized P_1-P_1 and P_2-P_2 methods. The residual stabilization parameter is $\alpha = 0.4$ for the P_1-P_1-stab and $\alpha = 0.01$ for the P_2-P_2-stab. We use Dirichlet conditions for the velocity on the whole boundary.

Our model problem is the unit square $\Omega = (0,1) \times (0,1)$ with the solution

$$p(x,y) = -\sin(x)\sinh(y) + C \quad \text{and}$$
$$\boldsymbol{u} = -\nabla p(x,y) = \begin{pmatrix} \cos(x)\sinh(y) \\ \sin(x)\cosh(y) \end{pmatrix},$$

with the constant C chosen so that $p \in L^2_0(\Omega)$. The pressure is harmonic, hence $g = \mathbf{div}\boldsymbol{u} = \Delta p = 0$. With similar reasoning $\boldsymbol{A}\boldsymbol{u} = \mathbf{0}$ which leads to $\boldsymbol{f} = \mathbf{0}$. Figure 1 shows the error as a function of the degrees of freedom for different values of the parameter t. The solution is smooth and all the methods perform as predicted by the theory. Notice the $\mathcal{O}(h)$ rate of convergence of the Taylor–Hood element in the Darcy type problem, that is, when the parameter t is small. This behavior is exactly as expected in the discussion following (31).

References

1. G. Allaire. Homogenization of the Navier-Stokes equations in open sets perforated with tiny holes. I. Abstract framework, a volume distribution of holes. *Arch. Rational Mech. Anal.*, 113(3):209–259, 1990
2. G. Allaire. Homogenization of the Navier-Stokes equations in open sets perforated with tiny holes. II. Noncritical sizes of the holes for a volume distribution and a surface distribution of holes. *Arch. Rational Mech. Anal.*, 113(3):261–298, 1990
3. T. Arbogast and H. L. Lehr. Homogenization of a Darcy-Stokes system modeling vuggy porous media. *Comput. Geosci.*, 10(3):291–302, 2006
4. D. N. Arnold, F. Brezzi, and M. Fortin. A stable finite element for the Stokes equations. *Calcolo*, 21(4):337–344 (1985), 1984
5. S. Badia and R. Codina. Unified stabilized finite element formulations for the Stokes and the Darcy problems. *SIAM J. Numer. Anal.*, 47(3):1971–2000, 2009
6. M. Bercovier and O. Pironneau. Error estimates for finite element method solution of the Stokes problem in the primitive variables. *Numer. Math.*, 33(2):211–224, 1979
7. D. Boffi. Stability of higher order triangular Hood–Taylor methods for the stationary Stokes equations. *Math. Models Methods Appl. Sci.*, 4(2):223–235, 1994
8. D. Boffi. Three-dimensional finite element methods for the Stokes problem. *SIAM J. Numer. Anal.*, 34(2):664–670, 1997
9. F. Brezzi and R. S. Falk. Stability of higher-order Hood–Taylor methods. *SIAM J. Numer. Anal.*, 28(3):581–590, 1991
10. F. Brezzi and J. Pitkäranta. On the stabilization of finite element approximations of the Stokes equations. In: *Efficient solutions of elliptic systems (Kiel, 1984)*, volume 10 of *Notes Numer. Fluid Mech.*, pages 11–19. Vieweg, Braunschweig, 1984
11. R. S. Falk. A Fortin operator for two-dimensional Taylor-Hood elements. *M2AN Math. Model. Numer. Anal.*, 42(3):411–424, 2008

12. L. P. Franca and R. Stenberg. Error analysis of Galerkin least squares methods for the elasticity equations. *SIAM J. Numer. Anal.*, 28(6):1680–1697, 1991
13. L. P. Franca, T. J. R. Hughes, and R. Stenberg. Stabilized finite element methods. In: M. Gunzburger and R.A. Nicolaides, editors, *Incompressible Computational Fluid Dynamics*, chapter 4, pages 87–107. Cambridge University Press, London, 1993
14. T. J. R. Hughes and L. P. Franca. A new finite element formulation for computational fluid dynamics. VII. The Stokes problem with various well-posed boundary conditions: symmetric formulations that converge for all velocity/pressure spaces. *Comput. Methods Appl. Mech. Engrg.*, 65(1):85–96, 1987
15. M. Juntunen and R. Stenberg. Analysis of finite element methods for the Brinkman problem. *Calcolo*, Online 2009, doi: 10.1007/s10092-009-0017-6
16. T. Lévy. Loi de Darcy ou loi de Brinkman? *C. R. Acad. Sci. Paris Sér. II Méc. Phys. Chim. Sci. Univers Sci. Terre*, 292(12):871–874, Erratum (17):1239, 1981
17. K. Mardal and R. Winther. Uniform preconditioners for the time dependent Stokes problem. *Numer. Math.*, 98:305–327, 2004
18. K. Mardal and R. Winther. Uniform preconditioners for the time dependent Stokes problem. *Numer. Math.*, 103:171–172, 2006
19. K. Mardal, X. Tai and R. Winther. A robust finite element method for Darcy-Stokes flow. *SIAM J. Numer. Anal.*, 40(5):1605–1631, 2002
20. K. R. Rajagopal. On a hierarchy of approximate models for flows of incompressible fluids through porous solids. *Math. Models Methods Appl. Sci.*, 17(2):215–252, 2007
21. L. R. Scott and M. Vogelius. Norm estimates for a maximal right inverse of the divergence operator in spaces of piecewise polynomials. *RAIRO Modél. Math. Anal. Numér.*, 19(1):111–143, 1985
22. R. Stenberg. On some three-dimensional finite elements for incompressible media. *Comput. Methods Appl. Mech. Engrg.*, 63(3):261–269, 1987
23. R. Stenberg. Error analysis of some finite element methods for the Stokes problem. *Math. Comp.*, 54:495–508, 1990

Finite Element Based Second Moment Analysis for Elliptic Problems in Stochastic Domains

H. Harbrecht

Abstract We present a finite element method for the numerical solution of elliptic boundary value problems on stochastic domains. The method computes the mean and the variance of the random solution with leading order in the amplitude of the stochastic boundary perturbation relative to an unperturbed, nominal domain. The variance is computed as the trace of the solution's two-point correlation which satisfies a deterministic boundary value problem on the tensor product of the nominal domain. This problem is discretized in the sparse tensor product space by a multilevel frame generated from standard finite elements. The computational complexity of the resulting approach stays essentially proportional to the number of finite elements required for the discretization of the nominal domain.

1 Introduction

Many problems in physics and engineering sciences lead to boundary value problems for an unknown function. In general, the numerical simulation is well understood provided that the input parameters are given exactly. Since, however, the input parameters are often not known exactly it is of growing interest to model such parameters stochastically.

A principal approach to solve boundary value problems with stochastic input parameters is the Monte Carlo approach, see e.g., [15] and the references therein. However, it is hard and extremely expensive to generate a large number of suitable samples and to solve a deterministic boundary value problem on each sample. Thus, we aim here at a direct, deterministic method to compute the stochastic solution.

Deterministic approaches to solve stochastic partial differential equations have been proposed in e.g., [1, 7–9, 14, 17]. Therein, loadings and coefficients have been

H. Harbrecht
Institute for Applied Analysis and Numerical Simulation, University of Stuttgart, Pfaffenwaldring 57, 70569 Stuttgart, Germany
e-mail: harbrecht@ians.uni-stuttgart.de

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_46,

considered as stochastic input parameters. Recently, in [6,12,16], also the underlying domain has been modeled as a stochastic input parameter $D(\omega)$. For example, this enables the consideration of tolerances in the shape of products fabricated by line production. Other applications arise from blurred interfaces like cell membranes or molecular surfaces.

The present paper is dedicated to elliptic boundary value problems on stochastic domains. We assume small stochastic perturbations around a nominal domain $\overline{D}$ with known statistics. Then, according to [12], we can linearize to derive deterministic equations for the random solution's expectation and two point-correlation

$$\mathrm{E}_u(\mathbf{x}) = \int_\Omega u(\mathbf{x},\omega)dP(\omega), \quad \mathrm{Cor}_u(\mathbf{x},\mathbf{y}) = \int_\Omega u(\mathbf{x},\omega)u(\mathbf{y},\omega)dP(\omega), \quad \mathbf{x},\mathbf{y}\in\overline{D}.$$

From these quantities the variance is derived by $\mathrm{Var}_u(\mathbf{x}) = \mathrm{Cor}_u(\mathbf{x},\mathbf{x}) - \mathrm{E}_u^2(\mathbf{x})$. Thus, applying the finite element based sparse tensor product discretization from [13], we are able to compute, to leading order in the amplitude of the random boundary perturbation, the solution's second order statistics. The complexity of our algorithm stays essentially proportional to the number of unknowns required to discretize the domain $\overline{D}$. In the present paper we describe the whole approach, specifying implementational details and related error estimates.

2 Elliptic Boundary Value Problems on Stochastic Domains

Let (Ω,Σ,P) be a suitable probability space. We consider the domain as the uncertain input parameter of an elliptic boundary value problem, i.e.,

$$\left.\begin{aligned} -\mathbf{div}\big[\mathbf{A}(\mathbf{x})\nabla u(\mathbf{x},\omega)\big] &= f(\mathbf{x}), \quad \mathbf{x}\in D(\omega) \\ u(\mathbf{x},\omega) &= g(\mathbf{x}), \quad \mathbf{x}\in\partial D(\omega) \end{aligned}\right\} \quad \omega\in\Omega. \tag{1}$$

To model the stochastic domain $D(\omega)$ let $\overline{D}$ denote a smooth reference domain and consider stochastic boundary variations in direction of the outer normal $\mathbf{U}(\mathbf{x},\omega) = \varepsilon\kappa(\mathbf{x},\omega)\mathbf{n}(\mathbf{x}) : \partial\overline{D}\to\mathbb{R}^n$ with $\kappa(\omega)\in L_P^2(\Omega, C^{2,1}(\partial\overline{D}))$ and $\|\kappa(\omega)\|_{C^{2,1}(\partial\overline{D})}\le 1$ almost surely. Then, the stochastic domain $D(\omega)$ is described via perturbation of identity

$$\partial D(\omega) = \big\{\big(\mathbf{I}+\varepsilon\mathbf{U}(\omega)\big)(\mathbf{x}) = \mathbf{x}+\varepsilon\kappa(\mathbf{x},\omega)\mathbf{n}(\mathbf{x}) : \mathbf{x}\in\partial\overline{D}\big\}.$$

For what follows we assume that the expectation E_κ and the two-point correlation Cor_κ of κ are given. Without loss of generality (otherwise we redefine $\overline{D}$ correspondingly) we assume that the perturbation field κ is centered, i.e., that $\mathrm{E}_\kappa\equiv 0$.

For small parameters $\varepsilon > 0$ one can linearize (1) by means of shape optimization:

Theorem 1 ([11, 12]). *Assume that the compact set $K \Subset \overline{D}$ satisfies $K \subset D(\omega)$ almost surely. Then, it holds that*

$$E_u(\mathbf{x}) = \overline{u}(\mathbf{x}) + \mathscr{O}(\varepsilon^2), \quad \mathrm{Cov}_u(\mathbf{x}, \mathbf{y}) = \varepsilon^2 \, \mathrm{Cor}_{du}(\mathbf{x}, \mathbf{y}) + \mathscr{O}(\varepsilon^3), \quad \mathbf{x}, \mathbf{y} \in K.$$

Herein, $\overline{u} \in H^1(\overline{D})$ and $\mathrm{Cor}_{du} \in H^{1,1}(\overline{D} \times \overline{D})$ satisfy the deterministic boundary value problems

$$\begin{aligned} -\boldsymbol{div}\big[\mathbf{A}(\mathbf{x})\nabla\overline{u}(\mathbf{x})\big] &= f(\mathbf{x}), \quad \mathbf{x} \in \overline{D}, \\ \overline{u}(\mathbf{x}) &= g(\mathbf{x}), \quad \mathbf{x} \in \partial\overline{D}, \end{aligned} \tag{2}$$

and

$$\begin{aligned} (\boldsymbol{div}_{\mathbf{x}} \otimes \boldsymbol{div}_{\mathbf{y}})\big[\big(\mathbf{A}(\mathbf{x}) \otimes \mathbf{A}(\mathbf{y})\big)(\nabla_{\mathbf{x}} \otimes \nabla_{\mathbf{y}})\,\mathrm{Cor}_{du}(\mathbf{x}, \mathbf{y})\big] &= 0, \quad \mathbf{x}, \mathbf{y} \in \overline{D}, \\ \boldsymbol{div}_{\mathbf{x}}\big[\mathbf{A}(\mathbf{x})\nabla_{\mathbf{x}}\,\mathrm{Cor}_u(\mathbf{x}, \mathbf{y})\big] &= 0, \quad \mathbf{x} \in \overline{D},\ \mathbf{y} \in \partial\overline{D}, \\ \boldsymbol{div}_{\mathbf{y}}\big[\mathbf{A}(\mathbf{y})\nabla_{\mathbf{y}}\,\mathrm{Cor}_u(\mathbf{x}, \mathbf{y})\big] &= 0, \quad \mathbf{x} \in \partial D,\ \mathbf{y} \in \overline{D}, \\ \mathrm{Cor}_{du}(\mathbf{x}, \mathbf{y}) = \mathrm{Cor}_{\kappa}(\mathbf{x}, \mathbf{y})\left[\frac{\partial(\overline{u} - g)}{\partial \mathbf{n}}(\mathbf{x}) \otimes \frac{\partial(\overline{u} - g)}{\partial \mathbf{n}}(\mathbf{y})\right], &\quad \mathbf{x}, \mathbf{y} \in \partial\overline{D}. \end{aligned} \tag{3}$$

3 Finite Element Discretization

3.1 Parametric Finite Elements

Starting point of the definition of the sparse multilevel frame is a nested sequence of finite dimensional trial spaces

$$V_0 \subset V_1 \subset \cdots \subset V_j \subset \cdots \subset H^1(\overline{D}). \tag{4}$$

In general, due to our smoothness assumptions on the domain, we have to deal with non-polygonal domains. To realize the multiresolution analysis (4) we will use parametric finite elements.

Let $\triangle$ denote the reference simplex in $\mathbb{R}^n$. We assume that the domain $\overline{D}$ is partitioned into a finite number of patches

$$\mathrm{clos}(\overline{D}) = \bigcup_k \tau_{0,k}, \quad \tau_{0,k} = \boldsymbol{\kappa}_k(\triangle), \quad k = 1, 2, \ldots, M,$$

where each $\boldsymbol{\kappa}_k : \triangle \to \tau_{0,k}$ defines a diffeomorphism of $\triangle$ onto $\tau_{0,k}$. The intersection $\tau_{0,k} \cap \tau_{0,k'}$, $k \neq k'$, of the patches $\tau_{0,k}$ and $\tau_{0,k'}$ is either $\emptyset$, or a lower dimensional face. The parametric representation is supposed to be globally continuous which means that the diffeomorphisms $\boldsymbol{\kappa}_i$ and $\boldsymbol{\kappa}_{i'}$ coincide at common patch

interfaces except for orientation. A mesh of level j on $\overline{D}$ is then induced by regular subdivisions of depth j of $\triangle$ into 2^{jn} simplices. This generates the $2^{jn}M$ curved elements $\{\tau_{j,k}\}$.

The ansatz functions $\Phi_j = \{\varphi_{j,k} : k \in \mathscr{I}_j\}$ are defined via parameterization, lifting continuous piecewise linear Lagrangian finite elements from $\triangle$ to the domain $\overline{D}$ by using the mappings $\boldsymbol{\kappa}_i$ and gluing across patch boundaries. Setting $V_j = \operatorname{span} \Phi_j$ yields (4), where $\dim V_j \sim 2^{jn}$.

To treat the non-homogeneous Dirichlet data in (2) and (3), we shall further distinguish between interior basis functions $\Phi_j^{\overline{D}} = \{\varphi_{j,k} : k \in \mathscr{I}_j^{\overline{D}}\}$ with $\varphi_{j,k}\big|_{\partial\overline{D}} \equiv 0$ and boundary functions $\Phi_j^{\partial\overline{D}} = \{\varphi_{j,k} : k \in \mathscr{I}_j^{\partial\overline{D}}\}$ with $\varphi_{j,k}\big|_{\partial\overline{D}} \not\equiv 0$.

The solution of the mean field equation (2) by multigrid accelerated finite element methods is straightforward and along the lines of standard literature, see e.g., [2,4]. Therefore, we will skip all the details here.

3.2 Multilevel Frames for Sparse Tensor Product Spaces

We will discretize (3) in the *sparse* tensor product space $\widehat{V}_J = \sum_{j+j' \le J} V_j \otimes V_{j'}$. Abbreviating $N_J := \dim V_J$ there holds $\widehat{N}_J := \dim \widehat{V}_J \sim N_J \log N_J$ which is substantially smaller than the dimension N_J^2 of the full tensor product space $V_J \otimes V_J = \sum_{j,j' \le J} V_j \otimes V_{j'}$. Nevertheless, the approximation power in $\widehat{V}_J$ is essentially the same as in the full tensor product space provided that there is extra regularity in terms of the anisotropic Sobolev spaces $H^{s,s}(\overline{D} \times \overline{D})$, see [5,17].

To discretize functions in $\widehat{V}_J$ one traditionally uses hierarchical bases like wavelet or multilevel bases, see for example [5]. In the present paper we use instead a multilevel frame as proposed in [13], i.e., we represent functions by the redundant but stable collection $\widehat{\Phi}_J := \{\varphi_{j,k} \otimes \varphi_{j',k'} : k \in \mathscr{I}_j,\ k' \in \mathscr{I}_{j'},\ j + j' \le J\}$. Thus, the structural and computational advantages of finite element methods are combined with the efficiency of sparse grid approximations.

It has been shown in [13] that $\operatorname{card}(\widehat{\Phi}_J) \sim \widehat{N}_J \sim N_J \log N_J$, i.e., this frame has still essentially optimal cardinality. Notice that the frame $\widehat{\Phi}_J$ is the restriction to $\widehat{V}_J$ of the two-fold tensor product of the frame that underlies the BPX-preconditioner [3].

3.3 Galerkin Discretization

We shall be concerned with Galerkin's method for solving the boundary value problem (3) in the sparse tensor product space. We abbreviate the mean's Neumann data by $\sigma := \partial(\overline{u} - g)/\partial\mathbf{n}$ and their approximate version by $\sigma_J := \langle \nabla(\overline{u}_J - g), \mathbf{n}\rangle$, with $\overline{u}_J \in V_J$ being the finite element solution of (2). Instead of the Dirichlet data of (3),

$$f := (\sigma \otimes \sigma)\,\mathrm{Cor}_\kappa \in H^{1/2,1/2}(\partial D \times \partial D), \tag{5}$$

we have only access to the approximation $f_J := (\sigma_J \otimes \sigma_J)\,\mathrm{Cor}_\kappa$ which lives on the full tensor product grid. Thus, we follow [12] and insert the L^2-orthoprojector $\widehat{\Pi}_J$ onto the sparse tensor product space $\widehat{V}_J|_{\partial D\times\partial D}$ according to

$$\widehat{f}_J := (\sigma_J \otimes \sigma_J)\widehat{\Pi}_J\,\mathrm{Cor}_\kappa\,. \tag{6}$$

We shall fix notation. Define for all $0 \le j, j' \le J$ the univariate stiffness matrices and, with respect to the traces of the ansatz functions, the mass matrices and the multiplication operators

$$\begin{aligned} \mathbf{A}^{\Theta}_{j,j'} &:= (\mathbf{A}\nabla\Phi^{\Theta}_{j'}, \nabla\Phi^{\overline{D}}_{j})_{L^2(\overline{D})}, \quad \Theta \in \{\overline{D}, \partial\overline{D}\}, \\ \mathbf{G}_{j,j'} &:= (\Phi^{\partial\overline{D}}_{j'}, \Phi^{\partial\overline{D}}_{j})_{L^2(\partial\overline{D})}, \quad \mathbf{M}_{j,j'} := (\sigma_J\Phi^{\partial\overline{D}}_{j'}, \Phi^{\partial\overline{D}}_{j})_{L^2(\partial\overline{D})}. \end{aligned} \tag{7}$$

Two-fold tensor products of these finite element matrices lead to the required matrices on the sparse tensor product space:

$$\widehat{\mathbf{A}}^{\Theta,\Xi}_J = \left[\mathbf{A}^{\Theta}_{j_1,j_2} \otimes \mathbf{A}^{\Xi}_{j_1',j_2'}\right]_{j_1+j_2,j_1'+j_2'\le J}, \quad \Theta,\Xi \in \{\overline{D}, \partial\overline{D}\},$$

$$\widehat{\mathbf{G}}_J = \left[\mathbf{G}_{j_1,j_2} \otimes \mathbf{G}_{j_1',j_2'}\right]_{j_1+j_2,j_1'+j_2'\le J}, \quad \widehat{\mathbf{M}}_J = \left[\mathbf{M}_{j_1,j_2} \otimes \mathbf{M}_{j_1',j_2'}\right]_{j_1+j_2,j_1'+j_2'\le J},$$

$$\left.\begin{aligned} \widehat{\mathbf{B}}^{\Theta}_J &= \left[\mathbf{A}^{\Theta}_{j_1,j_2} \otimes \mathbf{G}_{j_1',j_2'}\right]_{j_1+j_2,j_1'+j_2'\le J}, \\ \widehat{\mathbf{C}}^{\Theta}_J &= \left[\mathbf{G}_{j_1,j_2} \otimes \mathbf{A}^{\Theta}_{j_1',j_2'}\right]_{j_1+j_2,j_1'+j_2'\le J}, \end{aligned}\right\} \quad \Theta \in \{\overline{D}, \partial\overline{D}\}.$$

Finally, we need the data vector $\widehat{\mathbf{c}}_J = \left[(\mathrm{Cor}_\kappa, \Phi^{\partial\overline{D}}_j \otimes \Phi^{\partial\overline{D}}_{j'})_{L^2(\partial\overline{D}\times\partial\overline{D})}\right]_{j+j'\le J}$. Notice that (6) reads in the discrete form as $\widehat{\mathbf{f}}_J = \widehat{\mathbf{M}}_J\widehat{\mathbf{G}}_J^{-1}\widehat{\mathbf{c}}_J$.

In what follows we abbreviate Cor_{du} by v. To determine the approximate counterpart $\widehat{v}_J \in \widehat{V}_J$ we shall separate the degrees of freedom in order to solve the boundary value problem (3) successively: $\widehat{v}_J = \widehat{v}_J^{\overline{D},\overline{D}} + \widehat{v}_J^{\overline{D},\partial\overline{D}} + \widehat{v}_J^{\partial\overline{D},\overline{D}} + \widehat{v}_J^{\partial\overline{D},\partial\overline{D}}$, where

$$\widehat{v}_J^{\Theta,\Xi} := \sum_{j+j'\le J} (\Phi^{\Theta}_j \otimes \Phi^{\Xi}_{j'})\widehat{\mathbf{v}}^{\Theta,\Xi}_{j,j'}, \quad \Theta,\Xi \in \{\overline{D}, \partial\overline{D}\}.$$

Then we proceed as follows (see [11] for the details).

1. Determine $\widehat{v}_J^{\partial\overline{D},\partial\overline{D}}$ as the L^2-orthoprojection of the approximate Dirichlet data $\widehat{f}_J$ (6) onto the discrete trace space $\widehat{V}_J\big|_{\partial\overline{D}\times\partial\overline{D}}$ according to

$$\widehat{\mathbf{G}}_J\widehat{\mathbf{v}}_J^{\partial\overline{D},\partial\overline{D}} = \widehat{\mathbf{M}}_J\widehat{\mathbf{G}}_J^{-1}\widehat{\mathbf{c}}_J. \tag{8}$$

2. Compute $\widehat{v}_J^{\overline{D},\partial\overline{D}}$ such that $(\widehat{v}_J^{\partial\overline{D},\partial\overline{D}} + \widehat{v}_J^{\overline{D},\partial\overline{D}})\big|_{\overline{D}\times\partial\overline{D}} \in H^1(\overline{D}) \otimes H^{1/2}(\partial\overline{D})$ satisfies the homogeneous boundary condition on $\overline{D} \times \partial\overline{D}$. In complete analogy

determine $\widehat{v}_J^{\partial\overline{D},\overline{D}}$ which gives raise to

$$\widehat{\mathbf{B}}_J^{\overline{D}}\widehat{\mathbf{v}}_J^{\overline{D},\partial\overline{D}} = -\widehat{\mathbf{B}}_J^{\partial\overline{D}}\widehat{\mathbf{v}}_J^{\partial\overline{D},\partial\overline{D}}, \qquad \widehat{\mathbf{C}}_J^{\overline{D}}\widehat{\mathbf{v}}_J^{\partial\overline{D},\overline{D}} = -\widehat{\mathbf{C}}_J^{\partial\overline{D}}\widehat{\mathbf{v}}_J^{\partial\overline{D},\partial\overline{D}}. \tag{9}$$

3. Compute the function $\widehat{v}_J^{\overline{D},\overline{D}} \in H_0^{1,1}(\overline{D}\times\overline{D})$ inside the tensor product domain $\overline{D}\times\overline{D}$ according to

$$\widehat{\mathbf{A}}_J^{\overline{D},\overline{D}}\widehat{\mathbf{v}}_J^{\overline{D},\overline{D}} = -\widehat{\mathbf{A}}_J^{\partial\overline{D},\partial\overline{D}}\widehat{\mathbf{v}}_j^{\partial\overline{D},\partial\overline{D}} - \widehat{\mathbf{A}}_J^{\overline{D},\partial\overline{D}}\widehat{\mathbf{v}}_j^{\overline{D},\partial\overline{D}} - \widehat{\mathbf{A}}_J^{\partial\overline{D},\overline{D}}\widehat{\mathbf{v}}_j^{\partial\overline{D},\overline{D}}. \tag{10}$$

3.4 Error Estimates

Let $h_J := 2^{-J} \sim \max_k\{\operatorname{diam}(\tau_{J,k})\}$ denote the mesh size associated with the subspace V_J on $\overline{D}$. Then, from standard finite element theory for elliptic operators (e.g., [2,4]), we derive the following facts with respect to the approximate mean.

Proposition 1. *Equation* (2) *can be solved in linear complexity. The approximate mean* $\overline{u}_J$ *satisfies the error estimate* $\|\overline{u}-\overline{u}_J\|_{L^2(\overline{D})} \lesssim h_J^2\|\overline{u}\|_{H^2(\overline{D})}$ *provided that the given data are sufficiently smooth.*

In the Galerkin scheme we have to employ the perturbed Dirichlet data $\widehat{f}_J$ (6) instead of the original Dirichlet data f (5) to compute the approximate solution $\widehat{v}_J$ of (3). Therefore, we obtain only a reduced rate of convergence.

Theorem 2. *Assume that* $\overline{u}\in W^{2,\infty}(\overline{D})$, $g\in W^{1,\infty}(\overline{D})$, *and* $\operatorname{Cor}_\kappa \in H^{1,1}(\partial\overline{D}\times\partial\overline{D})$. *Then, the approximate solution* $\widehat{v}_J \in \widehat{V}_J$ *to* (3) *satisfies the error estimate*

$$\|v-\widehat{v}_J\|_{L^2(\overline{D}\times\overline{D})} \lesssim h_J\|\operatorname{Cor}_\kappa\|_{H^{1,1}(\partial\overline{D}\times\partial\overline{D})}\big\{\|\overline{u}\|_{W^{2,\infty}(\overline{D})} + \|g\|_{W^{1,\infty}(\overline{D})}\big\}^2.$$

Proof. The assertion follows immediately from [11] if we show that the consistency error of the right hand side satisfies

$$\big\|f-\widehat{f}_J\big\|_{L^2(\partial\overline{D}\times\partial\overline{D})} \lesssim h_J\|\operatorname{Cor}_\kappa\|_{H^{1,1}(\overline{D}\times\overline{D})}\big\{\|\overline{u}\|_{W^{2,\infty}(\overline{D})} + \|g\|_{W^{1,\infty}(\overline{D})}\big\}^2. \tag{11}$$

To show this estimate we proceed as follows:

$$\begin{aligned}
\big\|f-\widehat{f}_J\big\|_{L^2(\partial\overline{D}\times\partial\overline{D})} &= \big\|(\sigma\otimes\sigma)\operatorname{Cor}_\kappa - (\sigma_J\otimes\sigma_J)\widehat{\Pi}_J\operatorname{Cor}_\kappa\big\|_{L^2(\partial\overline{D}\times\partial\overline{D})}\\
&\le \big\|(\sigma\otimes\sigma-\sigma_J\otimes\sigma_J)\operatorname{Cor}_\kappa\big\|_{L^2(\partial\overline{D}\times\partial\overline{D})} + \big\|(\sigma_J\otimes\sigma_J)(I-\widehat{\Pi}_J)\operatorname{Cor}_\kappa\big\|_{L^2(\partial\overline{D}\times\partial\overline{D})}\\
&\le \|\sigma\otimes\sigma-\sigma_J\otimes\sigma_J\|_{L^\infty(\partial\overline{D}\times\partial\overline{D})}\|\operatorname{Cor}_\kappa\|_{L^2(\partial\overline{D}\times\partial\overline{D})}\\
&\qquad + \|\sigma_J\|^2_{L^\infty(\partial\overline{D})}\big\|(I-\widehat{\Pi}_J)\operatorname{Cor}_\kappa\big\|_{L^2(\partial\overline{D}\times\partial\overline{D})}.
\end{aligned} \tag{12}$$

We now estimate the two terms on the right hand side of this inequality separately. The L^2-orthoprojection onto the sparse grid space satisfies (cf. [5, 17])

$$\left\|(I-\widehat{\Pi}_J)\,\mathrm{Cor}_\kappa\right\|_{L^2(\partial\overline{D}\times\partial\overline{D})} \lesssim h_J \|\,\mathrm{Cor}_\kappa\,\|_{H^{1,1}(\partial\overline{D}\times\partial\overline{D})}. \tag{13}$$

Pointwise error estimates for piecewise linear finite elements (see e.g., [4]) imply

$$\|\sigma-\sigma_J\|_{L^\infty(\partial\overline{D})} = \|\langle\nabla(\overline{u}-\overline{u}_J),\mathbf{n}\rangle\|_{L^\infty(\partial\overline{D})} \le \|\nabla\overline{u}-\nabla\overline{u}_J\|_{L^\infty(\overline{D})} \lesssim h_J\|\overline{u}\|_{W^{2,\infty}(\overline{D})}.$$

This induces by standard tensor product arguments

$$\|\sigma\otimes\sigma-\sigma_J\otimes\sigma_J\|_{L^\infty(\partial\overline{D}\times\partial\overline{D})} \lesssim h_J\big(\|\sigma\|_{L^\infty(\partial\overline{D})}+\|\sigma_J\|_{L^\infty(\partial\overline{D})}\big)\|\overline{u}\|_{W^{2,\infty}(D)}. \tag{14}$$

Inserting (13), (14) and $\|\sigma_J\|_{L^\infty(\partial\overline{D})} \le \|\sigma-\sigma_J\|_{L^\infty(\partial\overline{D})} + \|\sigma\|_{L^\infty(\partial\overline{D})}$ into the estimate (12) yields the desired consistency result (11). □

3.5 *Fast Second Moment Computation*

The linear systems of equations arising from the sparse multilevel discretization can be assembled and solved in essentially linear complexity when using the following ingredients, developed in the papers [11–13].

1. Due to the non-uniqueness of the representation of functions in frame coordinates, all system matrices have a large kernel. Since the associated right hand side vectors lie in the related images, Krylov subspace methods converge without further modifications (see, e.g., [10, 13]). In practice, we apply the conjugate gradient method to solve (8)–(10).
2. The diagonally scaled system matrices are essentially well conditioned in the sense that all nonzero eigenvalues behave essentially like a fixed constant. Therefore, the conjugate gradient method converges with a rate that is essentially independent of the discretization level J (e.g., [10]).
3. Iterative solvers involve only matrix-vector multiplications. The fast matrix-vector multiplication developed in [11, 13] is of essentially linear complexity. Besides standard prolongations and restrictions, it involves only system matrices (7) with $0 \le j = j' \le J$, i.e., standard finite element matrices. Employing prolongations and restrictions, all coarse level matrices are successively derived from the finest grid matrices in linear time.
4. Numerical quadrature in the sparse tensor product space is performed as follows. We expand the two-point correlation into the hierarchical basis $\widehat{\Psi}_J := \{\varphi_{j,k}\otimes\varphi_{j',k'} : k\in\mathscr{I}_j\setminus\mathscr{I}_{j-1},\ k'\in\mathscr{I}_{j'}\setminus\mathscr{I}_{j'-1},\ j+j'\le J\} \subset \widehat{\Phi}_J$ of the sparse tensor product space $\widehat{V}_J$. For $\mathrm{Cor}_\kappa \in H^{s,s}(\partial\overline{D}\times\partial\overline{D})$ with $0\le s<2$ an approximation $\widehat{\mathrm{Cor}}_{\kappa,J}$ is obtained such that $\|\,\mathrm{Cor}_\kappa - \widehat{\mathrm{Cor}}_{\kappa,J}\|_{L^2(\partial\overline{D}\times\partial\overline{D})} = h_J^s\|\,\mathrm{Cor}_\kappa\,\|_{H^{s,s}(\partial\overline{D}\times\partial\overline{D})}$. In the case $s=2$ the factor $\sqrt{|\log h_J|}$ appears in addition, see [5] for the details.

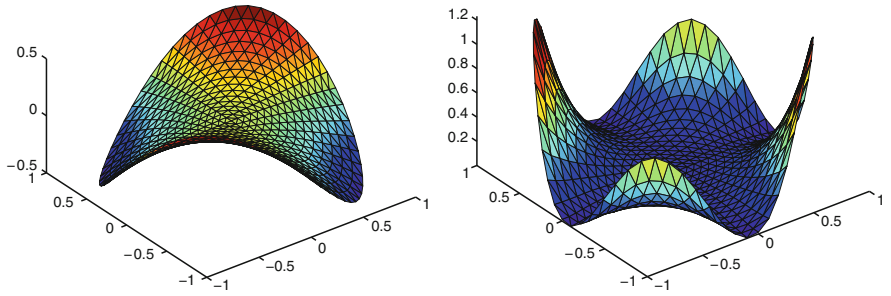


Fig. 1 Approximate mean (*left*) and variance (*right*) of u

Proposition 2. *By combining the ingredients* (1)–(4) *one arrives at an algorithm which computes the solution's second moment in a complexity being essentially proportional to the number of unknowns used to discretize the mean field equation* (2)*.*

4 Numerical Results

We consider the boundary value problem (1) with $\mathbf{A} \equiv \mathbf{I}$, $f \equiv 1/4$, $g(\mathbf{x}) = x \cdot y$, and $\overline{D}$ being the unit circle (i.e. $n = 2$). If we prescribe Gaussian correlation $\mathrm{Cor}_\kappa(\mathbf{x}, \mathbf{y}) = e^{-\|\mathbf{x}-\mathbf{y}\|^2}$ we get the solution's approximate mean and variance shown in Fig. 1. It turns out that the variance increases when approaching the boundary of the domain, i.e., the solution's sensitivity with respect to boundary perturbations is the larger the nearer the boundary is. This effect is stronger in regions where the modulus of the Dirichlet data g is large. The non-symmetry is induced by the present inhomogeneity $f \equiv 1/4$. Notice that the variance scales quadratically in the perturbation parameter ε and thus decreases correspondingly as $\varepsilon \to 0$. Further numerical results, especially a comparison with a Monte Carlo simulation, can be found in [11].

References

1. Babuška, I., Nobile, F., Tempone, R.: Worst case scenario analysis for elliptic problems with uncertainty. Numer. Math. **101**, 185–219 (2005)
2. Braess, D.: Finite elements. Theory, fast solvers, and applications in solid mechanics. Second edition. Cambridge University Press, Cambridge (2001)
3. Bramble, J., Pasciak, J., Xu, J.: Parallel multilevel preconditioners. Math. Comput. **55**, 1–22 (1990)
4. Brenner, S.C., Scott, L.R.: The mathematical theory of finite element methods. Texts in Applied Mathematics, 15. Springer, New York (1994)
5. Bungartz, H.J., Griebel, M.: Sparse Grids. Acta Numerica **13**, 147–269 (2004)

6. Canuto, C., Kozubek, T.: A fictitious domain approach to the numerical solution of PDEs in stochastic domains. Numer. Math. **107**, 257–293 (2007)
7. Deb, M.K., Babuška, I., Oden, J.T.: Solution of stochastic partial differential equations using Galerkin finite element techniques. Comput. Methods Appl. Mech. Eng. **190**, 6359–6372 (2001)
8. Frauenfelder, P., Schwab, C., Todor, R.A.: Finite elements for elliptic problems with stochastic coefficients. Comput. Meth. Appl. Mech. Eng. **194**, 205–228 (2004)
9. Ghanem, R.G., Spanos, P.D.: Stochastic finite elements: a spectral approach. Springer, New York (1991)
10. Griebel, G.: Multilevelmethoden als Iterationsverfahren über Erzeugendensystemen. Teubner Skripten zur Numerik. B.G. Teubner, Stuttgart (1994)
11. Harbrecht, H.: A finite element method for elliptic problems with stochastic input data. Appl. Numer. Math. **60**, 227–244 (2010)
12. Harbrecht, H., Schneider, R., Schwab, C.: Sparse second moment analysis for elliptic problems in stochastic domains. Numer. Math. **109**, 167–188 (2008)
13. Harbrecht, H., Schneider, R., Schwab, C.: Multilevel frames for sparse tensor product spaces. Numer. Math. **110**, 199–220 (2008)
14. Matthies, H.G., Keese, A.: Galerkin methods for linear and nonlinear elliptic stochastic partial differential equations. Comput. Methods Appl. Mech. Eng. **194**, 1295–1331 (2005)
15. Protter, P.: Stochastic integration and differential equations: a new approach. Springer, Berlin (1990)
16. Tartakovsky, D.M., Xiu, D.: Numerical methods for differential equations in random domains. SIAM J. Sci. Comput. **28**, 1167–1185 (2006)
17. von Petersdorff, T., Schwab, C.: Sparse wavelet methods for operator equations with stochastic data. Appl. Math **51**, 145–180 (2006)

On Robust Parallel Preconditioning for Incompressible Flow Problems

Timo Heister, Gert Lube, and Gerd Rapin

Abstract We consider time-dependent flow problems discretized with higher order finite element methods. Applying a fully implicit time discretization or an IMEX scheme leads to a saddle point system. This linear system is solved using a preconditioned Krylov method, which is fully parallelized on a distributed memory parallel computer.

We study a robust block-triangular preconditioner and besides numerical results of the parallel performance we explain and evaluate the main building blocks of the parallel implementation.

1 Introduction

The numerical simulation of time-dependent flow problems is an important task in research and industrial applications. The flow of Newtonian incompressible fluids is described by the system of the Navier–Stokes equations in a bounded domain $\Omega \subset \mathbb{R}^d$, $d = 2, 3$, where one has to find a velocity field $\mathbf{u} : [0, T] \times \Omega \to \mathbb{R}^d$ and a pressure field $p : [0, T] \times \Omega \to \mathbb{R}$ such that

$$\begin{aligned} \frac{\partial \mathbf{u}}{\partial t} - \nu \Delta \mathbf{u} + (\mathbf{u} \cdot \nabla)\, \mathbf{u} + \nabla p = \mathbf{f} \quad \text{in} \quad (0, T] \times \Omega, \\ \nabla \cdot \mathbf{u} = 0 \quad \text{in} \quad [0, T] \times \Omega. \end{aligned} \tag{1}$$

Here $\mathbf{f} : (0, T] \times \Omega \to \mathbb{R}^d$ is a given force field, and ν is the kinematic viscosity. For brevity initial and boundary conditions are omitted. One has to cope with some

T. Heister (✉) and G. Lube
Department of Mathematics, University of Göttingen, Germany
e-mail: heister@math.uni-goettingen.de

G. Rapin
VW, Interior Engineering, Wolfsburg, Germany

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_47, © Springer-Verlag Berlin Heidelberg 2010

modifications for turbulent flows, namely using $\nabla \cdot (2\nu S(\mathbf{u}))$ with $S(\mathbf{u}) := \frac{1}{2}(\nabla \mathbf{u} + \nabla \mathbf{u}^T)$ instead of $\nu \triangle \mathbf{u}$, variable and non-linear viscosity $\nu := \nu_{\text{const}} + \nu_t(\mathbf{u})$, and additional velocity terms from turbulence models.

In Sect. 2 this system of equations is discretized in space and time. The high spatial resolution needed for a typical domain $\Omega \subset \mathbb{R}^3$ leads to a number of unknowns in the order of millions of degrees of freedom. We need to calculate the solution at many time-steps, especially for optimization or inverse problems. This results in a demand for a robust and fast algorithm. We define such an algorithm in Sect. 3. The memory and performance requirements for the solution process can typically be met by a distributed memory cluster.

Let us state the requirements for the solver: *Flexibility*, to allow comparisons between different turbulence models, stabilization schemes, time discretizations, solvers, etc. *Parallelization*, ranging from multicore workstations to clusters with hundred or more CPUs. *Scalability*, with respect to the number of CPUs and problem size.

Combining these three requirements is a challenge. Research codes are usually *flexible*, but often lack the other requirements. On the other hand commercial codes usually work with lowest order discretization and are not flexible enough. For higher accuracy and flexibility we favor a coupled approach for the saddle point system instead of a splitting scheme.

We use the standard Multiple Instruction, Multiple Data streams (MIMD) parallel architecture model. The basis for the parallel implementation are parallel linear algebra routines running on top of MPI to allow parallel assembling and solving of the linear systems. The data matrices and vectors are split row-wise between the CPUs (Sect. 4). We conclude the paper with numerical results in Sect. 5.

2 Discretization

We start by semi-discretizing the continuous equation (1) in time. The solution $(\mathbf{u}, p)$ and the data $\mathbf{f}$ are expressed only at discrete time-steps $0 = t_0 < t_1 < \ldots < t_{max} = T$ of the time interval $[0, T]$, denoted by the superscript n, e.g., $\mathbf{u}^n$. We consider two different discretization schemes, the typical *implicit time discretization* and an implicit-explicit (short *IMEX*) scheme, c.f. [1]. The fully *implicit time discretization* leads to a sequence of non-linear stationary problems of the form

$$\begin{aligned} -\nu \triangle \mathbf{u}^n + c\mathbf{u}^n + (\mathbf{u}^n \cdot \nabla)\mathbf{u}^n + \nabla p^n &= \hat{\mathbf{f}}(\mathbf{u}^{n-1}, p^{n-1}), \\ \nabla \cdot \mathbf{u}^n &= 0, \end{aligned} \tag{2}$$

where $c \in \mathbb{R}$ is a reaction coefficient related to the inverse of the time-step size $\tau_n := t_{n+1} - t_n$ and $\hat{\mathbf{f}}$ is a modified right-hand side. Many time discretizations fit into this implicit scheme, for instance implicit Euler, BDF(2) or diagonal-implicit Runge–Kutta schemes. The non-linear system (2) is linearized by a fixed-point or Newton-type iteration. Hence, we have to solve a sequence of linear systems with a given divergence-free field $\mathbf{b}$ in the convective term $(\mathbf{b} \cdot \nabla)\mathbf{u}$.

The iteration for the non-linearity in (2) implies high computational costs. Explicit time-stepping is not desirable because of the strong restrictions on the time-step size. A remedy is to treat the non-linear term $(\mathbf{u}^n \cdot \nabla)\mathbf{u}^n$ in an explicit way, while the remainder of the equation is kept implicit. These methods are called *IMEX-schemes*. An elegant option is to combine an explicit Runge–Kutta scheme for the convection and an diagonal-implicit scheme, as used above, for the rest. With this method, the non-linearity disappears.

Thus, in both cases we end up with the solution of stationary Oseen problems:

$$\begin{aligned} -\nu\triangle\mathbf{u} + c\mathbf{u} + (\mathbf{b}\cdot\nabla)\mathbf{u} + \nabla p &= \mathbf{f}, \\ \nabla\cdot\mathbf{u} &= 0, \end{aligned} \tag{3}$$

which are discretized via Galerkin FEM on quadrilateral meshes with continuous, piece-wise (tensor-) polynomials Q_k of order $k > 0$. The inf-sup-stability is ensured using a Taylor-Hood pair Q_{k+1}/Q_k for velocity and pressure. This stable discretization leads to a finite-dimensional, linear saddle point system

$$\begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix} \begin{pmatrix} u \\ p \end{pmatrix} = \begin{pmatrix} f \\ g \end{pmatrix} \tag{4}$$

with finite element matrices A containing diffusion, reaction, and convection and B containing the pressure-velocity coupling.

3 The Solver

The system (4) is solved using the preconditioned Krylov subspace method FGMRES. This is a variant of the standard GMRES algorithm, see [12, 13]. FGMRES can cope with a changing preconditioner in each iteration. This is required because the preconditioner is not calculated explicitly as a matrix but is given as an implicit operator which uses iterative solvers internally. The usage of FGMRES in the context of flow problems is also described in detail in [9]. System (4) is preconditioned with an operator P^{-1} of block triangular type:

$$\begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix} P^{-1} \begin{pmatrix} v \\ q \end{pmatrix} = F \quad \text{with} \quad P^{-1} = \begin{pmatrix} \widetilde{A} & B^T \\ 0 & \widetilde{S} \end{pmatrix}^{-1}.$$

Here approximations $\widetilde{A}^{-1} \approx A^{-1}$ and $\widetilde{S}^{-1} \approx S^{-1}$ for the Schur complement $S := -BA^{-1}B^T$ are used. With exact evaluations of A and S the number of outer (F)GMRES steps is at most two, see [4]; this motivates the choice of the preconditioner. The inverse can be calculated by

$$P^{-1} = \begin{pmatrix} \widetilde{A}^{-1} & -\widetilde{A}^{-1} B^T \widetilde{S}^{-1} \\ 0 & \widetilde{S}^{-1} \end{pmatrix} = \begin{pmatrix} \widetilde{A}^{-1} & 0 \\ 0 & I \end{pmatrix} \begin{pmatrix} I & B^T \\ 0 & -I \end{pmatrix} \begin{pmatrix} I & 0 \\ 0 & -\widetilde{S}^{-1} \end{pmatrix}.$$

Each outer iteration requires the solution of two inner problems: the applications of $\widetilde{A}^{-1}$ and $\widetilde{S}^{-1}$, and there is one matrix-vector product with the matrix B^T.

There are several reasons for choosing a coupled approach. Using a projection method would introduce a CFL-like condition restricting the maximum time-step size. The main advantage of projection type methods (computational speed) can be simulated by only applying the preconditioner with a simple iteration method with a fixed number of steps (e.g., one). The result is comparable to a projection step method. Furthermore, the coupled approach fits better to higher order methods. Finally, this method has the advantage that the approximation quality of $\widetilde{A}^{-1}$ and $\widetilde{S}^{-1}$ is adjustable at will; the outer iteration converges either way.

The A-block forms a vector-valued convection–diffusion-reaction problem, which is a lot larger than the Schur complement. It is non-symmetric due to the convection part and the vector components may be coupled as a result of modifications for turbulent calculations, c.f. Sect. 1. An important part is the (strong) reaction term, which results in the low condition number of the matrix. Thus a BiCGStab solver with algebraic multi-grid preconditioning through BoomerAMG, [8], provides good results for $\widetilde{A}^{-1}$.

The approximation of the Schur complement $\widetilde{S}^{-1}$ is more difficult, because $S = -BA^{-1}B^T$ is dense and hence cannot be built explicitly as a matrix. Fortunately, the reaction-dominated A can be simplified with the mass matrix M_u:

$$S^{-1} \approx \left[B(cM_u)^{-1} B^T \right]^{-1} = c \left(BM_u^{-1} B^T \right)^{-1}.$$

We approximate $p = \widetilde{S}^{-1} q$ by a pressure Poisson problem:

$$-\frac{1}{c} \triangle p = q \tag{5}$$

and suitable boundary conditions, see [14]. The correct boundary conditions stem from $BM_u^{-1}B^T$, which cannot be applied directly. As an approximation there are Neumann boundary conditions applied in (5) where Dirichlet data is applied to the velocity in (1). Vice versa if Neumann boundary conditions are given in (1), homogeneous Dirichlet boundary conditions are applied in the Schur complement, (5). Periodic boundary conditions for the velocity can be treated with periodic boundary conditions in (5), which provide good results, c.f. Sect. 5.

The block triangular structure has been used for years, a good general overview is given in [4]. The form of the preconditioner is already described in [10], although we neclect the viscosity term in the Schur complement. A discussion of block preconditioners for flow problems can be found in [6] and [11]. Using FGMRES instead of e.g., GMRES proved to be a huge advantage not discussed there, but is motivated in [9].

4 Implementation Overview

The implementation of the solver described in this paper is built on top of a collection of known libraries. The basis is given by an MPI implementation for the parallel communication and the library PETSc, see [2], which supplies us with data structures and algorithms for scalable parallel calculations: matrices, vectors, and iterative solvers. The finite elements, mesh handling and assembling are performed by deal.II, see [3], which directly interfaces with the linear algebra objects from PETSc.

For the parallel calculations the rows in the system matrix have to be partitioned, such that each row is stored on exactly one CPU. This can be done with the following algorithm: first, create a graph, with cells as vertices and edges between two vertices if the corresponding cells are neighbors. This graph is partitioned into mostly equal-sized sets, such that each CPU "owns" a number of cells. The library METIS minimizes the number of cut edges. This reduces the amount of communication in parallel calculations. With the partition of the cells one can assign the owner for each degree of freedom. If two neighboring cells are owned by different CPUs, degrees of freedom on the shared face have to be assigned to one or the other CPU. By controlling this allocation one tries to balance the number of local rows per CPU. This improves the scalability of the solution process. The authors improved the way deal.II assigns these degrees of freedom, which decreases the imbalance of the number of degrees of freedom by up to 50%. This is done by making a (deterministic) pseudo-random choice.

The main loop is structured as follows: the outermost loop is the time stepping. For each time step the inner loop is repeated for each stage of the time discretization. For a fully implicit time discretization a fixed-point iteration surrounds the inner part. Finally the inner part consists of assembling and then solving the linear system with FGMRES. In each iteration the preconditioner is applied once. Finally, the preconditioner consists of the preconditioned inner solvers for A and S.

5 Numerical Results

We present the simulation of "Homogeneous Decaying Isotropic Turbulence" which is a widespread turbulence benchmark. The domain is given by a cube $[0, 2\pi]^3$ with periodic boundary conditions. A starting value (isotropic random velocity, see Fig. 1) from a given energy spectrum (calculated via Fourier transform) is prescribed. The problem has a Taylor-scale Reynolds number of Re_λ =150 and the viscosity is $\nu \approx$ 1.5e-5 (air). As a turbulence model we choose a standard LES Smagorinsky model with $\nu_t = (C_s \triangle_h)^2 |S(u)|$, $|M| := (2M \cdot M)^{\frac{1}{2}}$. The energy dissipation in time is compared to experimental data from [5], see Fig. 1, right. The calculations were done with Q_2/Q_1 elements on a mesh with 16^3 cells and the Smagorinsky constant $C_s = 0.17$. Here the filter-width $\triangle_h$ is given by h. This

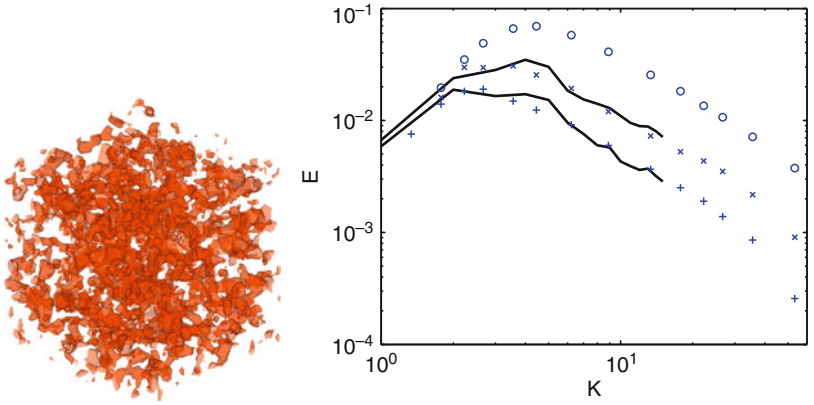


Fig. 1 *Left*: iso-surface of initial velocity spectrum; *right*: energy spectra at $t = 0.87$ and $t = 2.00$ (*upper* and *lower line*) and corresponding experimental data (*symbols*) with starting value

Table 1 *Left*: number of FMGRES iterations with respect to mesh size; *right*: speed-up and efficiency of assembling and solving

1/h	# DoFs	# It.
8	2,312	5
16	112,724	5
32	859,812	5
48	2,855,668	6
64	6,714,692	5

# CPUs	Speed-up assembling	Efficiency assembling	Speed-up solving	Efficiency solving
4	1.00	100%	1.00	100%
8	1.92	96%	1.72	86%
16	3.70	93%	2.91	73%
32	6.33	79%	4.69	59%
64	12.79	79%	7.39	46%

constant was not optimized but the results show good agreement to experimental data. For time discretization we apply a second order IMEX-scheme with a time-step size of 0.0087. The outer FGMRES residual is chosen as 1e-7 to the starting residual, whereas the inner residuals are set to 1e-5 (also relative). There are several important numerical results. The number of outer FGMRES iterations is independent of the number of CPUs, because there is no difference to the serial algorithm. The number of iterations is independent of the mesh size and lies between 5 and 6 iterations, see Table 1, left. This proves that the preconditioner design works well and the accuracy of $\widetilde{A}^{-1}$ and $\widetilde{S}^{-1}$ is sufficient. Now, we consider the so-called *strong scalability*, where the number of CPUs n is increased while the mesh size is kept fixed at $h = 1/32$, see Table 1, right. The scaling up to 64 processors is quite reasonable. The performance degrades slightly for larger number of processors, especially in the solution process. There are two reasons for this. First, the problem size is getting too small for the local calculations to dominate the communication costs. Second, the solver for the A-block, which takes the most time in the solution process, does not scale linearly. Table 2 shows the *weak scalability*, where the problem size increases together with the number of CPUs. This keeps the number of degrees

Table 2 Weak scalability of assembly- and solution-process w.r.t. increasing number of CPUs and number of degrees of freedom (time and efficiency)

# CPUs	1/h	# DoFs	Assembly		Solving	
6	24	368,572	20.07s	100%	44.86s	100%
16	32	859,812	18.21s	96%	49.14s	80%
54	48	2,855,668	19.16s	90%	49.02s	79%
128	64	6,714,692	19.98s	86%	54.64s	70%

of freedom per CPU nearly constant. The results are satisfying and efficiency only degrades slightly with a large number of CPUs.

6 Summary and Outlook

We recap our plans stated in the introduction and critically look what we accomplished in this paper. The development of a highly scalable parallel Navier–Stokes solver is underway. The parallel scalability is shown with the numerical results, but is constrained to a larger number of CPUs for several reasons. Some parts of deal.II are not yet parallellized, e.g., mesh handling and management of the degrees of freedom. The result is degraded performance and an increased demand on memory for a larger number of CPUs. This is visible starting at around 100 CPUs. With [7] we see good scaling up to thousands of CPUs with respect to computational costs and memory requirements. The goal of flexibility is solved in large parts. On the one hand we are able to compare different time discretizations, finite element orders and turbulence models. On the other hand we only look at instationary problems with small viscosities. Extending and testing the solver for a broad spectrum of test problems is still work in progress. The different regime of stationary and convection dominated flow poses challenges. The performance of the algebraic multi-grid for the A-Block needs to be verified there.

We present a flexible, parallel, and scalable solver framework for the solution of the incompressible Navier-Stokes equations. The numerical results prove that the design of the preconditioner is promising.

Acknowledgement T. Heister is partly supported by the DFG through GK 1023.

References

1. Ascher, U.M., Ruuth, S.J., Spiteri, R.J.: Implicit–explicit Runge–Kutta methods for time-dependent partial differential equations. Appl. Numer. Math. Trans. IMACS **25**(2–3), 151–167 (1997)
2. Balay, S., Buschelman, K., Gropp, W.D., Kaushik, D., Knepley, M.G., McInnes, L.C., Smith, B.F., Zhang, H.: PETSc Web page (2009). http://www.mcs.anl.gov/petsc

3. Bangerth, W., Hartmann, R., Kanschat, G.: deal.II – a general purpose object oriented finite element library. ACM Trans. Math. Softw. **33**(4), 27 (2007)
4. Benzi, M., Golub, G.H., Liesen, J.: Numerical solution of saddle point problems. Acta Numerica **14**, 1–137 (2005)
5. Comte-Bellot, G., Corrsin, S.: Simple eulerian time correlation of full- and narrow-band velocity signals in grid generated isotropic turbulence. J. Fluid Mech. **48**, 273–337 (1971)
6. Elman, H.C., Silvester, D.J., Wathen, A.J.: Finite Elements and Fast Iterative Solvers with Applications in Incompressible Fluid Dynamics. Numerical Mathematics and Scientific Computation. Oxford University Press, Oxford (2005)
7. Heister, T., Burstedde, C., Kronbichler, M., Bangerth, W.: Algorithms and Data Structures for Massively Parallel Generic Finite Element Codes
8. Henson, V.E., Yang, U.M.: Boomeramg: a parallel algebraic multigrid solver and preconditioner. Appl. Numer. Math. **41**(1), 155–177 (2002)
9. John, V.: Large Eddy Simulation of Turbulent Incompressible Flows: Analytical and Numerical Results for a Class of LES Models. *Lect. Notes Comput. Sci. Eng.*, vol. 34. Springer, Berlin (2004)
10. Loghin, D., Wathen, A.J.: Schur complement preconditioners for the Navier–Stokes equations. Int. J. Num. Meth. Fluids **40**(3–4), 403–412 (2002)
11. Olshanskii, M.A., Vassilevski, Y.V.: Pressure schur complement preconditioners for the discrete oseen problem. SIAM J. Sci. Comput. **29**(6), 2686–2704 (2007)
12. Saad, Y.: A Flexible Inner-Outer Preconditioned GMRES Algorithm. Tech. Rep. 91-279, Minnesota Supercomputer Institute, University of Minnesota (1991)
13. Saad, Y.: Iterative Methods for Sparse Linear Systems, second edn. Society for Industrial and Applied Mathematics, Philadelphia, PA (2003)
14. Turek, S.: Efficient Solvers for Incompressible Flow Problems: An Algorithmic and Computational Approach. Springer, Berlin (1999)

Hybrid Modeling of Plasmas

Mats Holmström

Abstract Space plasmas are often modeled as a magnetohydrodynamic (MHD) fluid. However, many observed phenomena cannot be captured by fluid models, e.g., non-Maxwellian velocity distributions and finite gyro radius effects. Therefore kinetic models are used, where also the velocity space is resolved. This leads to a six-dimensional problem, making the computational demands of velocity space grids prohibitive. Particle in cell (PIC) methods discretize velocity space by representing the charge distribution as discrete particles, and the electromagnetic fields are stored on a spatial grid. For the study of global problems in space physics, such as the interaction of a planet with the solar wind, it is difficult to resolve the electron spatial and temporal scales. Often a hybrid model is then used, where ions are represented as particles, and electrons are modeled as a fluid. Then the ion motions govern the spatial and temporal scales of the model. Here we present the mathematical and numerical details of a general hybrid model for plasmas. All grid quantities are stored at cell centers on the grid. The most common discretization of the fields in PIC solvers is to have the electric and magnetic fields staggered, introduced by Yee [IEEE Transactions on Antennas and Propagation 14:302–307 1966]. This automatically ensures that $\nabla \cdot \mathbf{B} = 0$, down to round-off errors. Here we instead present a cell centered discretization of the magnetic field. That the standard cell centered second order stencil for $\nabla \times \mathbf{E}$ in Faraday's law will preserve $\nabla \cdot \mathbf{B} = 0$ was noted by Tóth [Journal of Computational Physics 161:605–652 2000]. The advantage of a cell centered discretization is ease of implementation, and the possibility to use available solvers that only handle cell centered variables. We also show that the proposed method has very good energy conservation for a simple test problem in one-, two-, and three dimensions, when compared to a commonly used algorithm.

M. Holmström
Swedish Institute of Space Physics, P.O. Box 812, SE-98128 Kiruna, Sweden
e-mail: matsh@irf.se

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_48,

1 Introduction

Space plasmas are often modeled as a magnetohydrodynamic (MHD) fluid. However, many observed phenomena cannot be captured by fluid models, e.g., non-Maxwellian velocity distributions and finite gyro radius effects. Therefore kinetic models are used, where also the velocity space is resolved. This leads to a six-dimensional problem, making the computational demands of velocity space grids prohibitive. Particle in cell (PIC) methods discretize velocity space by representing the charge distribution as discrete particles, and the electromagnetic fields are stored on a spatial grid. For the study of global problems in space physics, such as the interaction of a planet with the solar wind, it is difficult to resolve the electron spatial and temporal scales. Often a hybrid model is then used, where ions are represented as particles, and electrons are modeled as a fluid. Then the ion motions govern the spatial and temporal scales of the model. Here we present the mathematical and numerical details of a general hybrid model for plasmas. All grid quantities are stored at cell centers on the grid. The most common discretization of the fields in PIC solvers is to have the electric and magnetic fields staggered, introduced by Yee [17]. This automatically ensures that $\nabla \cdot \mathbf{B} = 0$, down to round-off errors. Here we instead present a cell centered discretization of the magnetic field. That the standard cell centered second order stencil for $\nabla \times \mathbf{E}$ in Faraday's law will preserve $\nabla \cdot \mathbf{B} = 0$ was noted by [14]. The advantage of a cell centered discretization is ease of implementation, and the possibility to use available solvers that only provide for cell centered variables. We also show that the proposed method has very good energy conservation for a simple test problem in one-, two-, and three dimensions, when compared to a commonly used algorithm.

2 Definitions

We have N_I ions at positions $\mathbf{r}_i(t)$ [m] with velocities $\mathbf{v}_i(t)$ [m/s], mass m_i [kg] and charge q_i [C], $i = 1, \ldots, N_I$. By spatial averaging,[1] we can define the charge density $\rho_I(\mathbf{r}, t)$ [Cm^{-3}] of the ions, their average velocity $\mathbf{u}_I(\mathbf{r}, t)$ [m/s], and the corresponding current density $\mathbf{J}_I(\mathbf{r}, t) = \rho_I \mathbf{u}_I$ [$\mathrm{Cm}^{-2}\mathrm{s}^{-1}$]. Electrons are modelled as a fluid with charge density $\rho_e(\mathbf{r}, t)$, average velocity $\mathbf{u}_e(\mathbf{r}, t)$, and current density $\mathbf{J}_e(\mathbf{r}, t) = \rho_e \mathbf{u}_e$. The electron number density is $n_e = -\rho_e/e$, where e is the elementary charge. If we assume that the electrons are an ideal gas, then $p_e = n_e k T_e$, so the pressure is directly related to temperature (k is Boltzmann's constant).

The trajectories of the ions are computed from the Lorentz force,

$$\frac{d\mathbf{r}_i}{dt} = \mathbf{v}_i, \quad \frac{d\mathbf{v}_i}{dt} = \frac{q_i}{m_i}\left(\mathbf{E} + \mathbf{v}_i \times \mathbf{B}\right), \quad i = 1, \ldots, N_I$$

where $\mathbf{E} = \mathbf{E}(\mathbf{r}, t)$ is the electric field, and $\mathbf{B} = \mathbf{B}(\mathbf{r}, t)$ is the magnetic field.[2]

[1] Usually, charge and current densities are deposited on a grid, using shape functions [4].

[2] [1] modifies the electric field in the Lorentz force by a term proportional to $\mathscr{C}$ and $\nabla \times \mathbf{B}$ to preserve momentum.

2.1 Hybrid Approximations

A brief overview of hybrid codes can be found in [16]. A more complete survey can be found in [10]. Most hybrid solvers for global simulations have the following assumptions in common.

1. Quasi-neutrality, $\rho_I + \rho_e = 0$, so that given the ion charge density, the electron charge density is specified by $\rho_e = -\rho_I$.
2. Ampere's law whithout the transverse displacement current (also called the Darwin approximation, or the nonradiative limit) provides the total current, given $\mathbf{B}$, by
$$\mathbf{J} = \mu_0^{-1} \nabla \times \mathbf{B},$$
where $\mu_0 = 4\pi \cdot 10^{-7}$ [Hm^{-1}] is the magnetic constant ($\epsilon_0 \mu_0 c^2 = 1$), and from the total current we get the electron current, $\mathbf{J}_e = \mathbf{J} - \mathbf{J}_I$, and thus the electron velocity, since the quasi-neutrality implies that $\mathbf{u}_e = \mathbf{J}_e / \rho_e = (\mathbf{J}_I - \mathbf{J})/\rho_I$.
3. Massless electrons, $m_e = 0$, lead to the electron momentum equation
$$n_e m_e \frac{d\mathbf{u}_e}{dt} = \mathbf{0} = \rho_e \mathbf{E} + \mathbf{J}_e \times \mathbf{B} - \nabla p_e + \mathscr{C}$$
where the force terms $\mathscr{C}$ can be due to collisions, such as electron-ion collisions, electron-neutral [13] collisions, or anomalous, i.e., representing electron-wave interactions [1]. In our numerical experiments we have assumed that $\mathscr{C} = 0$. This provides an equation of state (Ohm's law) for the electric field
$$\mathbf{E} = \frac{1}{\rho_I} \left[(\mathbf{J} - \mathbf{J}_I) \times \mathbf{B} - \nabla p_e + \mathscr{C} \right],$$
with $\mathbf{J}$ from Ampere's law. So the electric field is not an unknown. Whenever it is needed, it can be computed.
4. Faraday's law is used to advance the magnetic field in time,
$$\frac{\partial \mathbf{B}}{\partial t} = -\nabla \times \mathbf{E}.$$
5. The electron pressure is isotropic (p_e is a scalar, not a tensor).

For the electrons, the remaining degree of freedom is the pressure, p_e. Note that p_e only affects the ion motions through the electric field. The evolution of the magnetic field is not affected since we have $\nabla \times \nabla p_e = 0$ in Faraday's law. There are several ways to handle the electron pressure [15, p. 8790],

1. Assume p_e is constant, or zero [5].
2. Assume p_e is adiabatic (small collision frequency). Then the electron pressure is related to the electron charge density by $p_e \propto |\rho_e|^\gamma$, where γ is the adiabatic index. Commonly used values are $\gamma = 5/3$ [1, 8], and $\gamma = 2$ [2, 12].

3. Solve the massless fluid energy equation [8, 11],

$$\frac{\partial p_e}{\partial t} + \mathbf{u}_e \cdot \nabla p_e + \gamma p_e \nabla \cdot \mathbf{u}_e = (\gamma - 1)\,\eta |\mathbf{J}|^2,$$

Here we assume that p_e is adiabatic. Then the relative change in electron pressure is related to the relative change in electron density by

$$\frac{p_e}{p_{e0}} = \left(\frac{n_e}{n_{e0}}\right)^{\gamma},$$

where the zero subscript denote reference values. From charge neutrality and $p_e = n_e k T_e$ we have that

$$p_e = A\rho_I^{\gamma} \text{ with } A = \frac{k}{e}\rho_I^{1-\gamma} T_e$$

a constant that is evaluated using reference values of ρ_I and T_e, e.g., solar wind values. Note that $\gamma = 1$ corresponds to assuming that T_e is constant, and $\gamma = 0$ gives a constant p_e.

2.1.1 Hybrid Equations

If we store the magnetic field on a discrete grid $\mathbf{B}_j$, the unknowns are $\mathbf{r}_i$, $\mathbf{v}_i$, and $\mathbf{B}_j$ (supplemented by p_e on a grid, if we include the electron energy equation). The time advance of the unknowns can then be written as the ODE

$$\frac{d}{dt}\begin{pmatrix} \mathbf{r}_i \\ \mathbf{v}_i \\ \mathbf{B}_j \end{pmatrix} = \begin{pmatrix} \mathbf{v}_i \\ \frac{q_i}{m_i}\left(\mathbf{E} + \mathbf{v}_i \times \mathbf{B}\right) \\ -\nabla_j \times \mathbf{E} \end{pmatrix} \tag{1}$$

where $\nabla_j \times$ is a discrete rotation operator, and the electric field is

$$\mathbf{E}_j = \frac{1}{\rho_I}\left(-\mathbf{J}_I \times \mathbf{B}_j + \mu_0^{-1}\left(\nabla_j \times \mathbf{B}_j\right) \times \mathbf{B}_j\right) - \nabla p_e + \mathscr{C}.$$

3 Discretization

An overview of different discretizations of the above equations can be found in [6, Appendix A]. [1, Sect. 3.1] provides a consise description of the CAM-CL algorithm introduced by [9]. All our grid variables will be cell centered: $\mathbf{B}_j$, $\mathbf{J}_j$, and $\boldsymbol{\rho}_j$ (here $\mathbf{J}_j$ and $\boldsymbol{\rho}_j$ are the ionic current and the ionic charge density at cell centers – from now we omit the subscript I for simplicity). We follow the Current Advance Method and Cyclic Leapfrog (CAM-CL) algorithm [9], but omit the CAM part. The Current

Advance Method is used to avoid multiple iterations over the particles, but does not conserve energy well, as we will see for a test problem.

For the particles, we have to solve for $\mathbf{r}_i$ and $\mathbf{v}_i$; and for the grid cells we have to solve for $\mathbf{B}_j$, $\mathbf{J}_j$ and $\boldsymbol{\rho}_j$. We denote time level $t = n\Delta t$ by superscript n. Given $\mathbf{B}_j^{n-1/2}$, $\mathbf{r}_i^{n-1/2}$ and $\mathbf{v}_i^n$, we do the following steps.

$$\mathbf{r}_i^{n+1/2} \leftarrow \mathbf{r}_i^{n-1/2} + \Delta t\, \mathbf{v}_i^n,$$

$$\mathbf{r}_i^n \leftarrow \frac{1}{2}\left(\mathbf{r}_i^{n+1/2} + \mathbf{r}_i^{n-1/2}\right).$$

At $\mathbf{r}_i^n$, deposit particle charges and currents

$$\rho_i \rightarrow \rho_j^n, \qquad \rho_i \mathbf{v}_i^n \rightarrow \mathbf{J}_j^n,$$

$$\mathbf{B}_j^{n+1/2} \leftarrow \mathbf{B}_j^{n-1/2}, \rho_j^n, \mathbf{J}_j^n \qquad \text{according to CL.} \tag{2}$$

At $\mathbf{r}_i^{n+1/2}$, deposit particle charge

$$\rho_i \rightarrow \rho_j^{n+1/2}.$$

Estimate electric field at $n + 1/2$ using the currents at n

$$\mathbf{E}_j^* \leftarrow \mathbf{B}_j^{n+1/2}, \rho_j^{n+1/2}, \mathbf{J}_j^n, p_e,$$

$$\mathbf{v}_i^{n+1/2} \leftarrow \mathbf{v}_i^n + \frac{\Delta t}{2}\frac{q_i}{m_i}\left(\mathbf{E}_j^* + \mathbf{v}_i^n \times \mathbf{B}_j^{n+1/2}\right).$$

At $\mathbf{r}_i^{n+1/2}$, deposit particle current

$$\rho_i \mathbf{v}_i^{n+1/2} \rightarrow \mathbf{J}_j^{n+1/2},$$

$$\mathbf{E}_j^{n+1/2} \leftarrow \mathbf{B}_j^{n+1/2}, \rho_j^{n+1/2}, \mathbf{J}_j^{n+1/2}, p_e$$

$$\mathbf{v}_i^{n+1} \leftarrow \mathbf{v}_i^n + \Delta t \frac{q_i}{m_i}\left(\mathbf{E}_j^{n+1/2} + \mathbf{v}_i^{n+1/2} \times \mathbf{B}_j^{n+1/2}\right).$$

Now we have $\mathbf{B}_j^{n+1/2}$, $\mathbf{r}_i^{n+1/2}$ and $\mathbf{v}_i^{n+1}$. Set $n \leftarrow n + 1$ and start over again.

For each particle we need a temporary vector. First $\mathbf{r}_i^{n+1/2}$ is temporarily saved during the deposit at $\mathbf{r}_i^n$. Then $\mathbf{v}_i^n$ is temporarily saved until the final velocity update. We also need to store the current corresponding to each particle, $\rho_i \mathbf{v}_i^*$, in preparation of the deposit operations.

The update of the magnetic field in (2) using cyclic leapfrog (CL) is done in m sub-time steps of length $h = \Delta t/m$. With the notation $\mathbf{B}_j^p \equiv \mathbf{B}_j((n + 1/2)\Delta t + ph)$ we have the iteration

$$\begin{cases} \mathbf{B}_j^1 \leftarrow \mathbf{B}_j^0 - h\nabla \times \mathbf{E}_j^0, \\ \mathbf{B}_j^{p+1} \leftarrow \mathbf{B}_j^{p-1} - 2h\nabla \times \mathbf{E}_j^p, \qquad p = 1, 2, \ldots, m-1, \\ \widetilde{\mathbf{B}}_j^m \leftarrow \mathbf{B}_j^{m-1} - h\nabla \times \mathbf{E}_j^m, \quad \text{and} \\ \mathbf{B}_j^{n+1/2} \leftarrow \frac{1}{2}\left(\mathbf{B}_j^m + \widetilde{\mathbf{B}}_j^m\right). \end{cases}$$

Since the magnetic field is leapfrogged in time, we need one temporary grid cell vector.

3.1 Non-Periodic Boundary Conditions

To be able to model the interaction of objects with the solar wind, non-periodic boundary conditions in the x-direction have been implemented. At $x_{\min}$ we have an inflow boundary, and at $x_{\max}$ an outflow boundary. The other boundaries are still periodic. The computation of $\nabla \times \mathbf{E}$ in the interior of the simulation domain requires $\mathbf{E}$ in one extra layer of cells in the x-directions. Also, computing $\mathbf{E}$ in the interior of the simulation domain involves $\nabla \times \mathbf{B}$, thus also requiring $\mathbf{B}$ in one outer layer of cells. At the inflow boundary we specify solar wind values of $\mathbf{B}$ and $\mathbf{E} = -\mathbf{u}_I \times \mathbf{B}$. At the outflow boundary we extrapolate $\mathbf{E}$ and $\mathbf{B}$ from the interior of the simulation domain to one external cell layer (a simple copy of the values from the upstream cells).

3.2 Spatial and Temporal Scales

If we want solutions of the discrete equations to be accurate approximations of the solutions to the continuous equations, a necessary condition is that the discretisation resolves all relevant spatial and temporal scales. The smallest spatial scale for the hybrid equations is the ion inertial length (the ion skin depth) $\delta_i = c/\omega_{pi}$, where c is the speed of light and ω_{pi} is the ion plasma frequency, $\omega_{pi}^2 = n_i q_i^2/(\epsilon_0 m_i)$, n_i the ion number density, q_i the ion charge, m_i the ion mass, and $\epsilon_0 \approx 8.854 \cdot 10^{-12}$ [Fm^{-1}] the vacuum permittivity. The ion inertial length is associated with the $\mathbf{J} \times \mathbf{B}$ term in Ohm's law (the Hall term) that describes whistler dynamics. The fastest temporal scale is also associated with whistler dynamics. The whistler wave spectrum is cutoff at the electron cyclotron frequency, but due to the assumption of massless electrons it is unbounded for the hybrid equations, and the frequency scales like $\omega/\Omega_i = (kc/\omega_{pi})$ for large k [10]. Here $\Omega_i = q_i B/m_i$ is the ion gyrofrequency. This gives the CFL constraint

$$\Delta t < \frac{\Omega_i^{-1}}{\sqrt{n}\pi}\left(\frac{\Delta x}{\delta_i}\right)^2$$

where n is the spatial dimension.

Table 1 Energy errors (total energy) for quiet plasma runs at times T. Numbers in parentheses indicate that the parameter was not stated in the reference

Reference	Dim.	Particles per cell	Δx δ_i	Δt Ω_i^{-1}	T Ω_i^{-1}	Error Ref.	Error here
[9]	1	16	0.5	0.1	100	9%	0.9%
					300	47%	3%
	2	32	0.5	0.1	100	2.6%	0.9%
					300	14%	3%
[3]	3	4	(1.54)	0.0056	112	<1%	0.25%

4 A Quiet Plasma Test Problem

A uniform, or quiet, plasma is a first test of any simulation code. The solution should only show small statistical fluctuations, and energy should be preserved for long simulation times. Matthews [9] describes one- and two-dimensional quiet plasma runs, and Brecht [3] present three-dimensional results.

The number of cells used here is 16, 64^2, and 32^3. All boundary conditions are periodic. Ion and electron temperatures are given by, $\beta_i = 1$, and $\beta_e = 0$. Brecht [3] uses a transport equation for the electron temperature. The number of magnetic field sub cycles is 4 in [9], 3 here, and [3] does not use sub cycling.

Total energy, the sum of the energy stored in the electric and magnetic fields and the kinetic energy of the particles, should be conserved. In Table 1 we compare the relative errors in total energy with the published values in one-, two-, and three dimensions.

5 Conclusions

The hybrid method stores the magnetic field on a grid. Here we have presented a cell centered algorithm as an alternative to the staggered grid commonly used. The cell centered method preserves $\nabla \cdot \mathbf{B} = 0$ down to round-off errors. In Table 1 it is evident that the proposed method conserves energy well when compared to the commonly used CAM-CL method [9]. That the CAM-CL method does not conserve energy well has been noted before [3, 7].

Acknowledgements This research was conducted using the resources of the High Performance Computing Center North (HPC2N), Umeå University, Sweden, and the Center for Scientific and Technical Computing (LUNARC), Lund University, Sweden. The software used in this work was in part developed by the DOE-supported ASC / Alliance Center for Astrophysical Thermonuclear Flashes at the University of Chicago.

References

1. Bagdonat, T., Motschmann, U.: 3D hybrid simulation code using curvilinear coordinates. Journal of Computational Physics **183**, 470–485 (2002). doi:10.1006/jcph.2002.7203
2. Bößwetter, A., Bagdonat, T., Motschmann, U., Sauer, K.: Plasma boundaries at Mars: A 3-D simulation study. Annales Geophysicae **22**, 4363–4379 (2004)
3. Brecht, S.H., Ledvina, S.A.: The solar wind interaction with the martian ionosphere/atmosphere. Space Science Reviews **126**, 15–38 (2006)
4. Hockney, R., Eastwood, J.: Computer Simulation Using Particles. Adam Hilger, Bristol (1989)
5. Kallio, E., Janhunen, P.: Modelling the solar wind interaction with Mercury by a quasi-neutral hybrid model. Annales Geophysicae **21**, 2133–2145 (2003)
6. Karimabadi, H., Krauss-Varban, D., Huba, J., Vu, H.: On magnetic reconnection regimes and associated three-dimensional asymmetries: Hybrid, hall-less hybrid, and hall-mhd simulations. Journal of Geophysical Research **109**, A09, 205 (2004). doi:10.1029/2004JA010478
7. Krauss-Varban, D.: From theoretical foundation to invaluable research tool: Modern hybrid simulations. In: Proceedings of the 7th International Symposium for Space Simulations (ISSS-7), pp. 15–18. Kyoto (2005)
8. Lipatov, A.S.: The Hybrid Multiscale Simulation Technology. Springer, Berlin (2002). p. 33
9. Matthews, A.P.: Current advance method and cyclic leapfrog for 2D multispecies hybrid plasma simulations. Journal of Computational Physics **112**, 102–116 (1994)
10. Pritchett, P.L.: Particle-in-cell simulations of magnetosphere electrodynamics. IEEE Transactions on Plasma Science **28**(6), 1976–1990 (2000)
11. Richardson, A., Chapman, S.C.: Self consistent one-dimensional hybrid code simulations of a relaxing field reversal. Journal of Geophysical Research **99**(A9), 17391 (1994)
12. Roussos, E., Müller, J., Simon, S., Bößwetter, A., Motschmann, U., Krupp, N., Fränz, M., Woch, J., Khurana, K.K., Dougherty, M.K.: Plasma and fields in the wake of Rhea: 3-D hybrid simulation and comparison with Cassini data. Annales Geophysicae **26**, 619–637 (2008)
13. Terada, N., Machida, S., Shinagawa, H.: Global hybrid simulation of the Kelvin-Helmholtz instability at the Venus ionopause. Journal of Geophysical Research **107**(A12), 1471 (2002). doi:10.1029/2001JA009224
14. Tóth, G.: The $\nabla \cdot B = 0$ constraint in shock-capturing magnetohydrodynamics codes. Journal of Computational Physics **161**, 605–652 (2000). doi:10.1006/jcph.2000.6519
15. Winske, D., Quest, K.: Electromagnetic ion beam instabilities: Comparison of one- and two-dimensional simulations. Journal of Geophysical Research **91**(A8), 8789–8797 (1986)
16. Winske, D., Yin, L.: Hybrid codes: Past, present and future. In: Proceedings of ISSS-6, pp. 1–4 (2001)
17. Yee, K.: Numerical solution of inital boundary value problems involving Maxwell's equations in isotropic media. IEEE Transactions on Antennas and Propagation **14**, 302–307 (1966). doi:10.1109/TAP.1966.1138693

A Priori Error Estimates for DGFEM Applied to Nonstationary Nonlinear Convection–Diffusion Equation

J. Hozman and V. Dolejší

Abstract We deal with a numerical solution of a scalar nonstationary convection–diffusion equation with nonlinear convective as well as diffusive terms which represents a model problem for the solution of the system of the compressible Navier–Stokes equations describing a motion of viscous compressible fluids. We present a discretization of this model equation by the discontinuous Galerkin finite element method (DGFEM) with several variants of the interior penalty, namely nonsymmetric (NIPG), symmetric (SIPG) and incomplete (IIPG) types of stabilizations of diffusion terms. Moreover, under some assumptions on the nonlinear terms, domain partitions and the regularity of the exact solution, we recall a priori hp error estimates in the L^2-norm and in the H^1-seminorm. A set of numerical experiments evaluating the experimental orders of convergence in the dependence on the polynomial degree of approximation and used type of stabilization is presented.

1 Introduction

Our goal is to develop a sufficiently robust, accurate and efficient numerical method for the solution of the system of the compressible Navier–Stokes equations describing a motion of viscous compressible fluids. Since the relevant mathematical theory is missing for the Navier–Stokes equations, it is convenient to study a model scalar equation at first, for which the theoretical base is established. The studied model scalar equation should correspond to the system of the Navier–Stokes equations in some simplification. Therefore, we introduce the concept of the *scalar nonstationary nonlinear convection-diffusion equation* with nonlinear convection as well as diffusion.

J. Hozman (✉)and V. Dolejší
Department of Numerical Mathematics, Faculty of Mathematics and Physics, Charles University Prague, Sokolovská 83, Prague, 186 75, Czech Republic
e-mail: jhozmi@volny.cz, dolejsi@karlin.mff.cuni.cz

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_49,

Among a wide class of numerical methods, the *discontinuous Galerkin finite element method* (DGFEM) seems to be a promising technique for the solution of convection-diffusion problems. DGFEM is based on a piecewise polynomial but discontinuous approximation, for a survey, see e.g. [3,4]. Within this paper we deal with the space semidiscretization of the model problem with the aid of three variants of DGFEM, namely nonsymmetric (NIPG), symmetric (SIPG) and incomplete interior penalty Galerkin (IIPG) techniques, see [2].

The analysis of a nonstationary convection-diffusion equation with a linear diffusion and a nonlinear convection was presented, e.g. in [6–9]. Further, the analysis of a scalar quasi-linear convection-diffusion equation can be found in [1]. A discontinuous Galerkin method applied to nonlinear parabolic equations with diffusion term $-\mathrm{div}(a(u)\nabla(u))$, $a : \mathbb{R} \to \mathbb{R}$ was analysed in [14]. Moreover, the solution of quasi-linear elliptic problems with a more general type of diffusion was studied in [11]. Finally, let us cite works [10, 13], where the nonstationary convection–diffusion equation with a diffusion of type from [14] was analysed.

In this paper, we considered a more general type of diffusion term, namely $-\mathrm{div}(A(u)\nabla(u))$, where $A : \mathbb{R} \to \mathbb{R}^{d,d}$ is a generally nonsymmetric matrix. We present a priori *hp* error estimates and a set of numerical examples verifying the theoretical results is included.

2 Scalar Nonstationary Nonlinear Convection–Diffusion Equation

Let $\Omega \subset \mathbb{R}^d$, $d = 2, 3$, be a bounded open polygonal (if $d = 2$) or polyhedral (if $d = 3$) domain with Lipschitz-continuous boundary $\partial\Omega = \partial\Omega_D \cup \partial\Omega_N$, $\partial\Omega_D \cap \partial\Omega_N = \emptyset$, and $T > 0$. Let us assume that the $(d-1)$ measure of $\partial\Omega_D$ is positive. We consider the following convection–diffusion problem: Find $u : Q_T = \Omega \times (0,T) \to \mathbb{R}$ such that

$$\frac{\partial u}{\partial t} + \nabla \cdot \mathbf{f}(u) = \mathrm{div}(\mathbf{K}(u)\,\nabla u) + g \quad \text{in } Q_T, \tag{1}$$

$$u\big|_{\partial\Omega_D \times (0,T)} = u_D, \tag{2}$$

$$\mathbf{K}(u)\nabla u \cdot \mathbf{n}\big|_{\partial\Omega_N \times (0,T)} = g_N, \tag{3}$$

$$u(x,0) = u^0(x), \quad x \in \Omega, \tag{4}$$

where $g : Q_T \to \mathbb{R}$, $u_D : \partial\Omega_D \times (0,T) \to \mathbb{R}$, $g_N : \partial\Omega_N \times (0,T) \to \mathbb{R}$, $u^0 : \Omega \to \mathbb{R}$ are given functions, $\mathbf{n} = (n_1, \ldots, n_d)$ is a unit outer normal to $\partial\Omega$, $\mathbf{f} = (f_1, \ldots, f_d) : \mathbb{R} \to \mathbb{R}^d$ represents convective terms and the regular matrix $\mathbf{K}(u) \in \mathbb{R}^{d,d}$ plays a role of nonlinear anisotropic diffusive coefficients. Moreover, if $\mathbf{K}u = \varepsilon\mathbf{I}$, where ε is a positive constant and $\mathbf{I} \in \mathbb{R}^{d,d}$ the unit matrix, then the problem (1)–(4) reduces to the equation with linear diffusion.

3 Discretization

Let $\{\mathcal{T}_h\}$ $(h > 0)$, be a family of partitions of the closure $\overline{\Omega}$ of the domain $\Omega \subset \mathbb{R}^d$ into a finite number of closed d-dimensional simplexes and/or parallelograms K with mutually disjoint interiors. We do not require the conformity of the mesh, i.e. the so-called hanging nodes are allowed. The symbols h_K and ρ_K stand for the diameter of K and radius of the largest d-dimensional ball inscribed into K, respectively. By $\mathcal{F}_h$ we denote the smallest possible set of all open $(d-1)$-dimensional faces (open edges when $d = 2$ or open faces when $d = 3$) of all elements $K \in \mathcal{T}_h$. Further, we label by $\mathcal{F}_h^I$ the set of all $\Gamma \in \mathcal{F}_h$ that are contained in Ω (inner faces), by $\mathcal{F}_h^D$ the set of all $\Gamma \in \mathcal{F}_h$ that $\Gamma \subset \partial\Omega_D$ (Dirichlet faces) and by $\mathcal{F}_h^N$ the set of all $\Gamma \in \mathcal{F}_h$ that $\Gamma \subset \partial\Omega_N$ (Neumann faces). Obviously, $\mathcal{F}_h = \mathcal{F}_h^I \cup \mathcal{F}_h^D \cup \mathcal{F}_h^N$. For a shorter notation we put $\mathcal{F}_h^{ID} \equiv \mathcal{F}_h^I \cup \mathcal{F}_h^D$. Finally, for each $\Gamma \in \mathcal{F}_h$, we define a unit normal vector $\mathbf{n}_\Gamma$. We assume that $\mathbf{n}_\Gamma$, $\Gamma \subset \partial\Omega$ has the same orientation as the outer normal of $\partial\Omega$. For $\mathbf{n}_\Gamma$, $\Gamma \in \mathcal{F}_h^I$ the orientation is arbitrary but fixed for each edge.

To each $K \in \mathcal{T}_h$, we assign a *local Sobolev index* $s_K \in I\!N$ and *local polynomial degree* $p_K \in I\!N$. Then we set the vectors $\mathsf{s} \equiv \{s_K, K \in \mathcal{T}_h\}$ and $\mathsf{p} \equiv \{p_K, K \in \mathcal{T}_h\}$. Over the triangulation $\mathcal{T}_h$ we define the so-called *broken Sobolev space* corresponding to the vector s

$$H^{\mathsf{s}}(\Omega, \mathcal{T}_h) \equiv \{v; v|_K \in H^{s_K}(K) \ \forall\, K \in \mathcal{T}_h\} \tag{5}$$

with the seminorm $|v|_{H^{\mathsf{s}}(\Omega,\mathcal{T}_h)} \equiv \left(\sum_{K\in\mathcal{T}_h} |v|^2_{H^{s_K}(K)}\right)^{1/2}$, where $|\cdot|_{H^{s_K}(K)}$ denotes the standard seminorm on the Sobolev space $H^{s_K}(K)$, $K \in \mathcal{T}_h$. Moreover, the approximate solution is sought in a space of discontinuous piecewise polynomial functions associated with the vector p

$$S_{h\mathsf{p}} \equiv S_{h\mathsf{p}}(\Omega, \mathcal{T}_h) \equiv \{v;\ v \in L^2(\Omega),\ v|_K \in P_{p_K}(K)\ \forall K \in \mathcal{T}_h\}, \tag{6}$$

where $P_{p_K}(K)$ denotes the space of all polynomials on K of degree $\le p_K$, $K \in \mathcal{T}_h$. In order to derive a priori hp error estimates we additionally assume that there exists a constant $C_P \ge 1$ such that $p_K/p_{K'} \le C_P\ \forall K, K' \in \mathcal{T}_h$ sharing a common face.

For each $\Gamma \in \mathcal{F}_h^I$ there exist two elements K_L, $K_R \in \mathcal{T}_h$ such that $\Gamma \subset \overline{K_L} \cap \overline{K_R}$. We use a convention that K_R lies in the direction of $\mathbf{n}_\Gamma$ and K_L in the opposite direction of $\mathbf{n}_\Gamma$. For $v \in S_{h\mathsf{p}}$, by $v|_\Gamma^{(L)}$ = trace of $v|_{K_L}$ on Γ, $v|_\Gamma^{(R)}$ = trace of $v|_{K_R}$ on Γ we denote the *traces* of v on edge Γ, which are different in general. Additionally, $[v]_\Gamma = v|_\Gamma^{(L)} - v|_\Gamma^{(R)}$ and $\langle v\rangle_\Gamma = \frac{1}{2}\left(v|_\Gamma^{(L)} + v|_\Gamma^{(R)}\right)$ denotes the *jump* and the *mean value* of function v over the edge Γ, respectively. For $\Gamma \in \partial\Omega$ there exists an element $K_L \in \mathcal{T}_h$ such that $\Gamma \subset \overline{K_L} \cap \partial\Omega$. Then for $v \in S_{h\mathsf{p}}$, we put: $v|_\Gamma^{(L)}$ = trace of $v|_{K_L}$ on Γ, $\langle v\rangle_\Gamma = [v]_\Gamma = v|_\Gamma^{(L)}$. In case that $[\cdot]_\Gamma$ and

$\langle\cdot\rangle_\Gamma$ are arguments of $\int_\Gamma \ldots \,\mathrm{d}S$, $\Gamma \in \mathscr{F}_h$ we omit the subscript Γ and write simply $[\cdot]$ and $\langle\cdot\rangle$, respectively.

Similarly as in [6], it is possible to derive the space semi-discretization of (1)–(4). A particular attention should be paid to the nonlinear diffusive term. In order to replace the inter-element continuity, we add some *stabilization* and *penalty* terms into formulation of the discrete problem. The convective term is approximated with the aid of a *numerical flux*, known from the finite volume method.

Therefore, we say that $u_h \in C^1(0,T;S_{h\mathsf{p}})$ is the *semi-discrete solution* of (1)–(4) if $(u_h(0), v_h) = (u^0, v_h)\ \forall v_h \in S_{h\mathsf{p}}$ and

$$\begin{aligned}\left(\frac{\partial u_h(t)}{\partial t}, v_h\right) + b_h(u_h(t), v_h) + a_h^\Theta(u_h(t), v_h) + \alpha J_h^\sigma(u_h(t), v_h) \qquad (7)\\ = l_h^\Theta(u_h(t), v_h)\,(t) \qquad\qquad \forall\, v_h \in S_{h\mathsf{p}},\ \forall\, t \in (0,T),\end{aligned}$$

where

$$\begin{aligned}a_h^\Theta(u,v) = &\sum_{K\in\mathscr{T}_h}\int_K \mathbf{K}(u)\,\nabla u\cdot\nabla v\,\mathrm{d}x - \sum_{\Gamma\in\mathscr{F}_h}\int_\Gamma \langle \mathbf{K}(u)\,\nabla u\cdot\mathbf{n}\rangle [v]\,\mathrm{d}S\\ &+\Theta\sum_{\Gamma\in\mathscr{F}_h^{ID}}\int_\Gamma \langle \mathbf{K}(u)^T\,\nabla v\cdot\mathbf{n}\rangle [u]\,\mathrm{d}S, \qquad (8)\end{aligned}$$

$$b_h(u,v) = -\sum_{K\in\mathscr{T}_h}\int_K \mathbf{f}(u)\cdot\nabla v\,\mathrm{d}x + \sum_{\Gamma\in\mathscr{F}_h}\int_\Gamma H(u|_\Gamma^{(L)}, u|_\Gamma^{(R)}, \mathbf{n}_\Gamma)\,[v]\mathrm{d}S, \qquad (9)$$

$$J_h^\sigma(u,v) = \sum_{\Gamma\in\mathscr{F}_h^{ID}}\int_\Gamma \sigma [u]\,[v]\,\mathrm{d}S, \qquad (10)$$

$$\begin{aligned}l_h^\Theta(u,v)(t) = &\int_\Omega g(t)\,v\,\mathrm{d}x + \sum_{\Gamma\in\mathscr{F}_h^N}\int_\Gamma g_N(t)\,v\,\mathrm{d}S\\ &+\sum_{\Gamma\in\mathscr{F}_h^D}\int_\Gamma \Big(\Theta\,\mathbf{K}(u)^T\,\nabla v\cdot\mathbf{n}\,u_D(t) + \sigma\,u_D(t)\,v\Big)\,\mathrm{d}S. \qquad (11)\end{aligned}$$

According to value of parameter Θ, we speak of *symmetric* (SIPG, $\Theta = -1$), *incomplete* (IIPG, $\Theta = 0$) or *nonsymmetric* (NIPG, $\Theta = 1$) variants of stabilization of DGFEM, i.e. we generally consider three variants of the diffusion form a_h^Θ and right-hand side form l_h^Θ. Penalty terms are represented by J_h^σ and the penalty parameter function σ in (10) is defined by

$$\sigma|_\Gamma = \frac{C_W}{d(\Gamma)} \quad \text{with} \quad d(\Gamma) = \begin{cases} \min\left(h_{K_p}/p_{K_p}^2, h_{K_n}/p_{K_n}^2\right), & \Gamma \in \mathscr{F}_h^I,\\ h_{K_p}/p_{K_p}^2, & \Gamma \in \mathscr{F}_h^D, \end{cases} \qquad (12)$$

where $C_W > 0$ is a suitable constant depending on the used variant of scheme and on the degree of polynomial approximation. The value of multiplicative constant α before the penalty form J_h^σ depends on the properties of matrix $\mathbf{K}$ and will be specified in Sect. 4, assumption (14). As for the convective form b_h we treat boundary terms similarly as in the finite volume method, i.e. with the aid of a numerical flux $H(u, v, \mathbf{n})$.

We shall assume that the numerical flux H is Lipschitz continuous (i.e. $|H(u, v, \mathbf{n}) - H(u^*, v^*, \mathbf{n})| \le C(|u-u^*| + |v-v^*|)\ \forall u, u^*, v, v^* \in \mathbb{R}\ \forall \mathbf{n} \in \mathbb{R}^d$), consistent with the convective fluxes $\mathbf{f}$ (i.e. $H(u, u, \mathbf{n}) = \mathbf{f}(u) \cdot \mathbf{n}\, \forall u \in \mathbb{R}\ \forall \mathbf{n} \in \mathbb{R}^d$) and conservative (i.e. $H(u, v, \mathbf{n}) = -H(v, u, -\mathbf{n})\ \forall u, v \in \mathbb{R}\ \forall \mathbf{n} \in \mathbb{R}^d$). Then we find that the sufficiently regular solution u of (1)–(4) satisfies

$$\left(\frac{\partial u(t)}{\partial t}, v_h\right) + b_h(u(t), v_h) + a_h^\Theta(u(t), v_h) + \alpha J_h^\sigma(u(t), v_h) = l_h^\Theta(u(t), v_h)\,(t) \quad (13)$$
$$\forall\, v_h \in S_{hp}\ \forall\, t \in (0, T),$$

which implies the *Galerkin orthogonality property* of the error.

4 A Priori Error Analysis

The basic framework refers to [5, 10, 13] with some generalization for considered problem. In order to carry out the error analysis we need to specify the additional assumptions on mesh, nonlinear diffusion term and regularity of the solution u of the continuous problem (1)–(4). Therefore, we assume that

(A1) The triangulations $\mathcal{T}_h$, $h \in (0, h_0)$, $h_0 > 0$, are *locally quasi-uniform*, i.e. there exists a constant $C_Q > 0$ such that $h_{K_L} \le C_Q h_{K_R}\ \forall K_L, K_R \in \mathcal{T}_h$ sharing face $\Gamma \in \mathcal{F}_h^I$ and *shape-regular*, i.e. there exists a constant $C_S > 0$ such that $h_K \le C_S \rho_K\ \forall K \in \mathcal{T}_h$.

(A2) The matrix $\mathbf{K}(v) = \{k_{ij}(v)\}_{i,j=1}^d$, $k_{ij}(v) : \mathbb{R} \to \mathbb{R}$, appearing in the diffusion terms satisfies

$$\begin{aligned} &\text{(a)}\ \|\mathbf{K}(v)\|_\infty \le C_U \text{ and } \|\mathbf{K}(v)^T\|_\infty \le C_U\ \forall v \in \mathbb{R}, \\ &\text{(b)}\ \|\mathbf{K}(v_1) - \mathbf{K}(v_2)\|_\infty \le C_L |v_1 - v_2|\ \forall v_1, v_2 \in \mathbb{R}, \\ &\text{(c)}\ \mathbf{z}^T \mathbf{K}(v) \mathbf{z} \ge \alpha \|\mathbf{z}\|^2,\ \alpha > 0,\ \forall v \in \mathbb{R},\ \forall \mathbf{z} \in \mathbb{R}^d, \end{aligned} \quad (14)$$

where $\|\cdot\|_\infty$ represents the l^∞-matrix norm, i.e. $\|\mathbf{K}\|_\infty = \max\limits_{1 \le i \le n} \sum_{j=1}^n |k_{ij}|$.

(A3) The weak solution u is sufficiently regular, namely

$$\begin{aligned} &\text{(a)}\ u \in L^2(0, T; H^{\bar{s}}(\Omega)),\ \frac{\partial u}{\partial t} \in L^2(0, T; H^{\bar{s}}(\Omega)),\ u \in L^\infty(0, T; H^{\bar{s}}(\Omega)), \\ &\text{(b)}\ \|\nabla u(t)\|_{L^\infty(\Omega)} \le C_D \quad \text{for a.a. } t \in (0, T), \end{aligned} \quad (15)$$

where $\bar{s} = \max\{s_K,\ s_K \in \mathsf{S}\} \ge 2$.

Now, we can proceed to formulation of the main result of this paper.

Theorem 1. *Let the numerical flux H from (9) be consistent, conservative and Lipschitz continuous, let $\mathcal{T}_h$, $h \in (0, h_0)$ be a family of triangulations satisfying (A1) and let assumptions (A2) be satisfied. Let u be the exact solution of the continuous problem satisfying (A3) and $u_h \in S_{h\mathsf{p}}$ the solution of the discrete problem given by (7), where the penalty parameter σ satisfies (12). Then the discretization error $e_h = u_h - u$ satisfies the estimate*

$$\max_{t\in[0,T]} \|e_h(t)\|_{L^2(\Omega)} + \frac{\alpha}{2}\int_0^T |||e_h(\vartheta)|||^2 \,\mathrm{d}\vartheta \le Q(T) \sum_{K\in\mathcal{T}_h} \frac{h_K^{2\mu_K-2}}{p_K^{2s_K-3}} \|u\|_K^2, \quad (16)$$

where $\mu_K = \min(p_K + 1, s_K)$, $K \in \mathcal{T}_h$, $Q(T)$ is a function depending on T and constants from assumptions but independent of h_K, p_K, s_K, $K \in \mathcal{T}_h$ and

$$|||v|||^2 \equiv |v|^2_{H^1(\Omega,\mathcal{T}_h)} + J_h^\sigma(v,v), \quad (17)$$

$$\|u\|_K^2 \equiv \|u\|^2_{L^2(0,T;H^{s_K}(K))} + \|\partial u/\partial t\|^2_{L^2(0,T;H^{s_K}(K))} + \|u\|^2_{L^\infty(0,T;H^{s_K}(K))}. \quad (18)$$

Proof. The main framework is based on the application of the *multiplicative trace inequality*, *inverse inequality* and *approximation properties of the space* $S_{h\mathsf{p}}$, for more details see [8, Lemmas 4.2–4.4]. The whole proof can be found in [12, Theorem 4.3.2]. □

Remark 1. The estimate (16) cannot be used for $\alpha \to 0+$, because the term $Q(t)$ blows up exponentially with respect to $1/\alpha$. The case $\alpha \to 0+$ corresponds to a vanishing diffusion term, see (14)c. The divergence of the estimate is a consequence of the application of Young's inequality and Gronwall's lemma, necessary for overcoming the nonlinearity of the convective and diffusive terms.

Remark 2. Let $p_k = p$ and $s_K = s \;\forall\, K \in \mathcal{T}_h$. If u is sufficiently regular exact solution, we observe that the error estimate (16) is

(a) h-suboptimal in the $L^\infty(0,T,L^2(\Omega))$-norm, namely $O(h^p)$,
(b) h-optimal in the $L^2(0,T,H^1(\Omega))$-seminorm, namely $O(h^p)$.

If u is not sufficiently regular exact solution, one can see that the error estimate (16) is impressed only with given regularity, namely $O(h^{s-1})$.

Moreover, in both cases, the estimate (16) is

(c) p-suboptimal in the $L^\infty(0,T,L^2(\Omega))$-norm and the $L^2(0,T,H^1(\Omega))$-seminorm, namely $O(p^{-(s-3/2)})$.

This suboptimality of the p-estimate is caused by an application of the multiplicative trace inequality.

5 Numerical Example

In this section we verify the a priori error estimates (16). We consider the 2D viscous Burgers equation

$$\frac{\partial u}{\partial t} + \sum_{s=1}^{2} u \frac{\partial u}{\partial x_s} = \operatorname{div}(\mathbf{K}(u)\,\nabla u) + g \quad \text{in } \Omega \times (0,T), \tag{19}$$

with the nonsymmetric matrix $K(w)$ in the following form

$$\mathbf{K}(w) = \varepsilon \begin{pmatrix} \frac{1}{2}(3 + \arctan(w)) & \frac{1}{3}(2 - \arctan^2(w)) \\ 0 & 4 + \arctan(w) \end{pmatrix}. \tag{20}$$

We set $\varepsilon = 0.02$, $\Omega = (0,1)^2$, $T = 10$ and define the function g and the initial and boundary conditions in such a way that the exact solution has the steady-state form

$$u(x_1, x_2, t) = 2\left(1 - \mathrm{e}^{-10t}\right)(x_1^2 + x_2^2)x_1x_2(1 - x_1)(1 - x_2). \tag{21}$$

In the presented numerical experiments we use piecewise linear (P^1), quadratic (P^2) and cubic (P^3) polynomial approximations on a sequence of six triangular

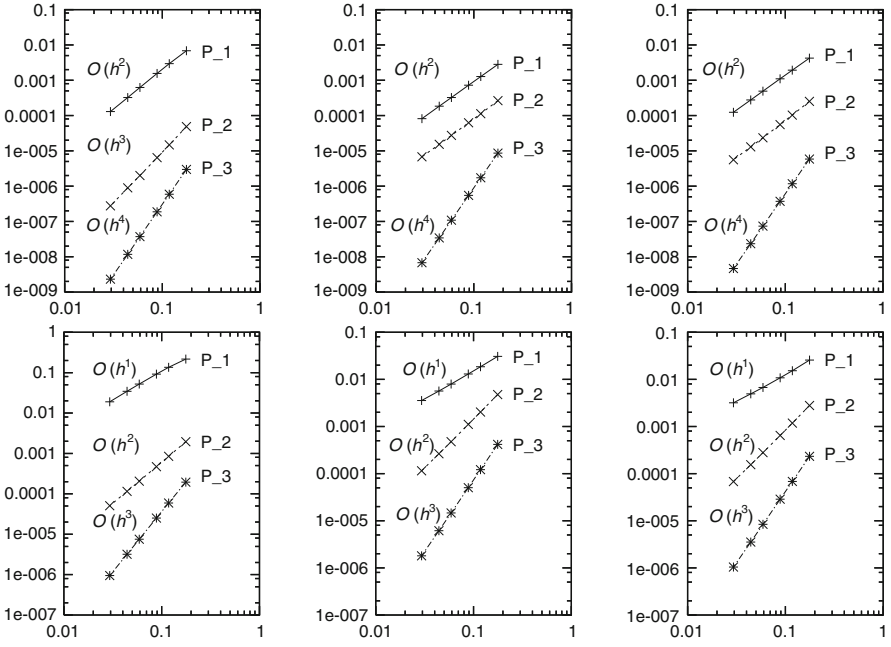


Fig. 1 Computational errors and the corresponding EOC in the L^2-norm (*top*) and H^1-seminorm (*bottom*) for SIPG (*left*), NIPG (*middle*) and IIPG (*right*) variant

meshes having 128, 288, 512, 1152, 2048 and 4608 elements for SIPG, NIPG and IIPG variants of DGFEM.

Figure 1 depicts computational errors at time $t = T = 10$ and experimental orders of convergence (EOC) for each IPG variant in the corresponding L^2-norm and in H^1-seminorm, respectively. Since $u(\cdot, \cdot, t)$ is sufficiently regular solution over Ω it follows from Remark 2 that the theoretical errors estimates are of order $O(h^p)$. On the other hand, we observe that

- ***L^2-norm:*** The obtained numerical results indicate a better behaviour of EOC, which is expected to be asymptotically $O(h^{p+1})$ for p odd and $O(h^p)$ for p even in the case of NIPG and/or IIPG variant. Moreover, in the case of SIPG variant we observe optimal EOC for p arbitrary.
- ***H^1-seminorm:*** EOC is in agreement with theoretical results, in other words, all IPG techniques produce optimal order of convergence $O(h^p)$.

Acknowledgements This work is a part of the research project MSM 0021620839 financed by the Ministry of Education of the Czech Republic and it was partly supported by the Grant No. 10209/B-MAT/MFF of the Grant Agency of the Charles University Prague.

References

1. Arnold, D.N.: An interior penalty finite element method with discontinuous elements. SIAM J. Numer. Anal. **19**(4), 742–760 (1982)
2. 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 (2002)
3. Cockburn, B.: Discontinuous Galerkin methods for convection dominated problems. In: Barth, T.J., Deconinck, H. (eds.) High-Order Methods for Computational Physics, Lecture Notes in Computational Science and Engineering 9, pp. 69–224. Springer, Berlin (1999)
4. Cockburn, B., Karniadakis, G.E., Shu, C.-W. (eds.): Discontinuous Galerkin Methods. Springer, Berlin (2000)
5. Dolejší, V.: Analysis and application of IIPG method to quasilinear nonstationary convection-diffusion problems. J. Comp. Appl. Math. (2007). doi: 10.1016/j.cam.2007.10.055
6. Dolejší, V., Feistauer, M.: Error estimates of the discontinuous Galerkin method for nonlinear nonstationary convection-diffusion problems. Numer. Funct. Anal. Optim. **26**(25–26), 2709–2733 (2005)
7. Dolejší, V., Feistauer, M., Kučera, V., Sobotíková, V.: An optimal $L^\infty(L^2)$-error estimate of the discontinuous Galerkin method for a nonlinear nonstationary convection-diffusion problem. IMA J. Numer. Anal. **28**(3), 496–521 (2008)
8. Dolejší, V., Feistauer, M., Sobotíková, V.: A discontinuous Galerkin method for nonlinear convection-diffusion problems. Comput. Meth. Appl. Mech. Eng. **194**, 2709–2733 (2005)
9. Feistauer, M., Dolejší, V., Kučera, V., Sobotíková, V.: $L^\infty(L^2)$-error estimates for the DGFEM applied to convection-diffusion problems on nonconforming meshes. J. Numer. Math. **17**, 45–65 (2009)
10. Feistauer, M., Kučera, V.: Analysis of the DGFEM for nonlinear convection-diffusion problem. ETNA **32**, 33–48 (2008)
11. Houston, P., Robson, J., Süli, E.: Discontinuous Galerkin finite element approximation of quasilinear elliptic boundary value problems I: The scalar case. IMA J. Numer. Anal. **25**, 726–749 (2005)

12. Hozman, J.: Discontinuous Galerkin method for convection-diffusion problems. PhD thesis, Charles University Prague, Faculty of Mathematics and Physics (2009)
13. Kučera, V.: Higher order methods for the solution of compressible flows. PhD thesis, Charles University Prague, Faculty of Mathematics and Physics (2008)
14. Rivière, B., Wheeler, M.F.: A discontinuous Galerkin method applied to nonlinear parabolic equations. In: Cockburn, B., Karniadakis, G.E., Schu, C.-W. (eds.) Discontinuous Galerkin methods. Theory, computation and applications, volume 11 of Lect. Notes Comput. Sci. Eng., pp. 231–244. Springer, Berlin (2000)

Stable Crank–Nicolson Discretisation for Incompressible Miscible Displacement Problems of Low Regularity

Max Jensen and Rüdiger Müller

Abstract In this article we study the numerical approximation of incompressible miscible displacement problems with a linearised Crank–Nicolson time discretisation, combined with a mixed finite element and discontinuous Galerkin method. At the heart of the analysis is the proof of convergence under low regularity requirements. Numerical experiments demonstrate that the proposed method exhibits second-order convergence for smooth and robustness for rough problems.

1 Introduction and Initial Boundary Value Problem

Mathematical models which describe the miscible displacement of fluids are of particular economical relevance in the recovery of oil in underground reservoirs by fluids which mix with oil. They also play a significant role in CO_2 stratification.

This publication extends the analysis of [1], which studies the discretisation of miscible displacement under low regularity. Unlike to [1] which is based on a first-order implicit Euler time-step (leading to a nonlinear system of equations in each time step), here we examine the discretisation in time by a linearised second-order Crank–Nicolson scheme. Crucially, the new, more efficient method inherits stability under low regularity. Like in [1], the concentration equation is approximated with a discontinuous Galerkin method, while Darcy's law and the incompressibility condition is formulated as a mixed method. High-order time-stepping for miscible displacement under low regularity has recently also been addressed in [4], however, with a continuous Galerkin discretisation in space and discontinuous Galerkin in time. We refer for an outline of the general literature to [1–4].

M. Jensen (✉)
Mathematical Sciences, University of Durham, England
e-mail: m.p.j.jensen@durham.ac.uk

R. Müller
WIAS, Berlin, Germany
e-mail: mueller@wias-berlin.de

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_50,

Definition 1 (Weak Formulation). A triple (u, p, c) in

$$L^\infty(0,T;H_N(\mathrm{div};\Omega)) \times L^\infty(0,T;L^2_0(\Omega)) \times \big(L^2(0,T;H^1(\Omega)) \cap H^1(0,T;H^2(\Omega)^*)\big)$$

is called weak solution of the incompressible miscible flow problem if

(W1) for $t \in (0,T)$, $v \in H_N(\mathrm{div};\Omega)$ and $q \in L^2_0(\Omega)$

$$\begin{aligned}\big(\mu(c)\,\mathbf{K}^{-1}u, v\big) - \big(p, \mathbf{div} v\big) &= \big(\rho(c)\,g, v\big)\\ \big(q, \mathbf{div} u\big) &= \big(q^I - q^P, q\big).\end{aligned}$$

(W2) for all $w \in \mathscr{D}(0,T;H^2(\Omega))$

$$\int_0^T -\big(\phi\, c, \partial_t w\big) + \big(\mathbb{D}(u)\nabla c, \nabla w\big) + \big(u\cdot\nabla c, w\big) + \big(q^I c, w\big) - \big(\hat{c} q^I, w\big)\mathrm{d}t = 0.$$

(W3) $c(0,\cdot) = c_0$ in $H^2(\Omega)^*$.

Here $H_N(\mathrm{div};\Omega)$ denotes the functions in $H(\mathrm{div};\Omega)$ whose trace vanishes in normal direction. Equation (W2) implements homogenous Neumann boundary conditions. For the data qualification we refer to condition (A1)–(A8) in [1] and for the physical interpretation of the system to [1–3]. We point out that $\mathbb{D}$ growths proportionally with u:

$$d_\circ(1+|u|)|\xi|^2 \le \xi^{\mathsf{T}}\mathbb{D}(u,x)\xi \le d^\circ(1+|u|)|\xi|^2, \qquad u,\xi \in \mathbb{R}^d,\ x \in \Omega.$$

Thus $\mathbb{D}$ is in general unbounded on Lipschitz domains Ω and in the presence of discontinuous coefficients, which are permitted in this paper.

2 The Finite Element Method

We compactly recall the definition of the finite element spaces from [1]. Let $0 = t_0 < t_1 < \ldots < t_M = T$ be a partition of the time interval $[0,T]$. Let $k_j := t_j - t_{j-1}$ and $d_t a^j := k_j^{-1}\big(a^j - a^{j-1}\big)$. We consider meshes $\mathscr{T}$ of Ω with elements K and set $h_K := \mathrm{diam}(K)$. We denote by $\mathscr{S}^s(\mathscr{T})$ the space of elementwise polynomial functions of total or partial degree s. For $w_h \in \mathscr{S}^s(\mathscr{T})$ the function $\nabla_h w_h$ is defined through $(\nabla_h w_h)|_K = \nabla(w_h|_K)$. The sets of interior and boundary faces are $\mathscr{E}_\Omega(\mathscr{T})$ and $\mathscr{E}_{\partial\Omega}(\mathscr{T})$. We set $\mathscr{E}(\mathscr{T}) = \mathscr{E}_\Omega(\mathscr{T}) \cup \mathscr{E}_{\partial\Omega}(\mathscr{T})$ and assign to each $E \in \mathscr{E}(\mathscr{T})$ its diameter h_E. We denote jump and the average operators by $[\cdot]$ and $\{\cdot\}$. The concentration c is discretised at time j on the mesh $\mathscr{T}_c^j$ or simply by $\mathscr{T}^j$. The approximation space for the variable c at time step j is denoted by $\mathscr{S}_c^j$. Often we abbreviate $\mathscr{E}^j := \mathscr{E}(\mathscr{T}_c^j)$, $\mathscr{E}_\Omega^j := \mathscr{E}_\Omega(\mathscr{T}_c^j)$, $\mathscr{E}_{\partial\Omega}^j := \mathscr{E}_{\partial\Omega}(\mathscr{T}_c^j)$. We denote the Raviart–Thomas space of order ℓ by $\mathrm{RT}^\ell(\mathscr{T}_u^j)$. The approximation spaces of u and p are $\mathscr{S}_u^j := \mathrm{RT}^\ell(\mathscr{T}_u^j) \cap H_N(\mathrm{div};\Omega)$ and

$\mathscr{S}_p^j := \mathscr{S}^\ell(\mathscr{T}_u^j) \cap L_0^2(\Omega)$. We frequently use the global mesh size and time step $h^j := \max_{K\in\mathscr{T}_c^j\cup\mathscr{T}_u^j} h_K$, $\tilde{h} := \max_{0\le j\le M} h^j$, $\tilde{k} := \max_{0\le j\le M} k^j$ as well as to $\mathscr{S}_u = \prod_{j=1}^M \mathscr{S}_u^j, \mathscr{S}_p = \prod_{j=1}^M \mathscr{S}_p^j, \mathscr{S}_c = \prod_{j=0}^M \mathscr{S}_c^j$. In addition we impose conditions (M1)–(M5) of [1] which are on shape-regularity, boundedness of the polynomial degree, control $\|v_h\|_{L^4} \lesssim \|v_h\|_{H^1}$ and the structure of hanging nodes.

To deal with discontinuous coefficients and the time derivative, we substitute $\mathbb{D}$ by $\mathbb{D}_h : L^2(\Omega)^d \to \mathscr{S}^s(\mathscr{T}_c, \mathbb{R}^{d\times d}),\ v \mapsto \Pi_{\mathscr{T}} \circ \mathbb{D}(v,\cdot)$ where the $\Pi_{\mathscr{T}}$ are projections such that $\|\Pi_{\mathscr{T}} D\|_K \lesssim \|D\|_K$. Given quantities a^j, a^{j-1} and a^{j-2} at times t_j, t_{j-1}, t_{j-2}, we denote $\overline{a}^j = \frac{1}{2}a^j + \frac{1}{2}a^{j-1}$ and $\breve{a} = \frac{3}{2}a^{j-1} - \frac{1}{2}a^{j-2}$.

The diffusion term of the concentration equation is discretised by the symmetric interior penalty discontinuous Galerkin method: Given $c_h, w_h \in \mathscr{S}_c^j$, $u_h \in \mathscr{S}_u^j$, we set

$$\begin{aligned} B_d(c_h, w_h; u_h) := {} & \big(\mathbb{D}_h^j(u_h)\nabla_h c_h, \nabla_h w_h\big) - \big([c_h], \{\mathbb{D}_h^j(u_h)\,\nabla_h w_h\}\big)_{\mathscr{E}_\Omega^j} \\ & - \big([w_h], \{\mathbb{D}_h^j(u_h)\,\nabla_h c_h\}\big)_{\mathscr{E}_\Omega^j} + \big(\sigma^2[c_h], [w_h]\big)_{\mathscr{E}_\Omega^j} \end{aligned}$$

where σ is chosen sufficiently large to ensure coercivity of B_d, cf. [1]. The convection, injection and production terms are represented by

$$\begin{aligned} & B_{cq}(c_h, w_h; u_h) \\ & := {}^1\!/_2 \Big(\big(u_h\nabla_h c_h, w_h\big) - \big(u_h c_h, \nabla_h w_h\big) + \big((\overline{q}^I + \overline{q}^P)c_h, w_h\big) \\ & \quad + \sum_{K\in\mathscr{T}^j} \big(c_h^+, (u_h\cdot n_K)_+ [w_h]_K\big)_{\partial K\setminus\partial\Omega} - \big((u_h\cdot n_K)_- [c_h]_K, w_h^+\big)_{\partial K\setminus\partial\Omega} \Big), \end{aligned} \tag{1}$$

where $(u_h\cdot n)_+ := \max\{u_h\cdot n, 0\}$ and $(u_h\cdot n)_- := \min\{u_h\cdot n, 0\}$. We set $B = B_d + B_{cq}$.

ALGORITHM (A^{dG}). *Choose $c_h^j \in \mathscr{S}_c^j$ for $j = 0, 1$. Given c_h^j, find $(u_h^j, p_h^j) \in \mathscr{S}_u^j \times \mathscr{S}_p^j$ such that*

$$\begin{aligned} \big(\mu(c_h^j)\,\mathbf{K}^{-1}u_h, v_h\big) - \big(p_h, \mathbf{div} v_h\big) &= \big(\rho(c_h^j)\,g, v_h\big), \\ \big(q_h, \mathbf{div} u_h\big) &= \big((q^I - q^P)^j, q_h\big). \end{aligned} \tag{2}$$

For $2 \le j \le M$ find $c_h^j \in \mathscr{S}_c^j$ such that, for all $w_h \in \mathscr{S}_c^j$,

$$\big(\phi\, d_t c_h^j, w_h\big) + B(\overline{c_h}^j, w_h; \breve{u}_h^j) = \big(\overline{\hat{c}}^j \overline{q}^{I\,j}, w_h\big) \tag{3}$$

and solve (2) to obtain $(u_h^j, p_h^j) \in \mathscr{S}_u^j \times \mathscr{S}_p^j$.

The algorithm only requires the solution of a *linear* system in each time step. The iterate c_h^1 can be computed with an implicit Euler method and fine time steps. The use of extrapolated values such as $\breve{u}_h^j$ is classical, e.g. see [5, p. 218].

3 Unconditional Well-Posedness, Boundedness and Convergence

Given c_h^{j-1} and c_h^{j-2}, there exists a solution $c_h^j \in \mathscr{S}_u^j$ of (3) because the bilinear form B is positive definite. For $t \in [t_{j-1}, t_j]$, let $\tilde{c}_h(t,\cdot) := \frac{t-t_{j-1}}{k_j} c_h^j + \frac{t_j - t}{k_j} c_h^{j-1}$. Then $\partial_t \tilde{c}_h(t,\cdot) = d_t c_h^j(\cdot)$. We interpret elements of $\mathscr{S}_u$, $\mathscr{S}_p$ and $\mathscr{S}_c$ as time-dependent functions with stepwise constant values. Let

$$|c_h|^2_{\breve{u}_h} := \big(\mathbb{D}_h(\breve{u}_h)\nabla_h c_h, \nabla_h c_h\big) + \big(\sigma^2[c_h],[c_h]\big)_{\mathscr{E}_\Omega^j} + \big(|\breve{u}_h \cdot n_{\mathscr{E}^j}|\,[c_h],[c_h]\big)_{\mathscr{E}_\Omega^j}.$$

Theorem 1. *Let $\rho^\circ = \|\rho\|_\infty$. There exists a constant $C > 0$ such that*

$$\|\breve{u}_h^j\| + \|\boldsymbol{div}\breve{u}_h^j\| + \|\breve{p}_h^j\| \lesssim \big(\|\rho^\circ g\| + \|\breve{q}^I - \breve{q}^P\|\big) \tag{4}$$

holds for all $j = 2, 3 \ldots, M$. Equally we have

$$\|\phi^{1/2} c_h^j\|^2 + \int_{t_1}^{t_j} |\overline{c_h}|^2_{\breve{u}_h^j}\,\mathrm{d}t \le \|\phi^{1/2} c_h^1\|^2 + \int_{t_1}^{t_j} \|\big(\overline{q}^{I\,i}\big)^{1/2}\,\overline{\hat{c}}^i\|^2\,\mathrm{d}t \tag{5}$$

for all $j = 2, 3 \ldots, M$.

Proof. The stability of u^{j-1}, u^{j-2}, p^{j-1}, p^{j-2} follows from a classical inf-sup argument. This implies stability of $\breve{u}^j$ and $\breve{p}^j$. We choose $w_h = \overline{c_h}^i$ in (3) to verify that

$$\begin{aligned} d_t\|\phi^{1/2} c_h^i\|^2 + |\overline{c_h}^i|^2_{\breve{u}_h^i} + \|(\overline{q}^I + \overline{q}^P)^{1/2}\overline{c_h}^i\|^2 \\ \le 2\big(\phi\, d_t c_h^i, \overline{c_h}^i\big) + 2B(\overline{c_h}^i, \overline{c_h}^i; \breve{u}_h^i) = 2\big(\overline{\hat{c}}^i\,\overline{q}^{I\,i}, \overline{c_h}^i\big). \end{aligned}$$

The Cauchy–Schwarz inequality, multiplication by k_i and summation over i give

$$\|\phi^{1/2} c_h^j\|^2 + \sum_{i=2}^{j} k_i |\overline{c_h}^i|^2_{\breve{u}_h^i} \le \|\phi^{1/2} c_h^1\|^2 + \sum_{i=2}^{j} k_i \|\big(\overline{q}^{I\,i}\big)^{1/2}\,\overline{\hat{c}}^i\|^2$$

for all $j = 2, 3, \ldots, M$. □

For simplicity the next theorem is stated assuming meshes are not adapted in time. For the extension to changing meshes consult [1]. However, observe that that the discretisation with the implicit Euler method gives additional stability in $k_i\|\phi^{1/2} d_t c_h^i\|^2$, which allows to change meshes more rapidly.

Theorem 2. *The time derivative $\partial_t \tilde{c}_h$ belongs to $L^2(t_1, T; H^2(\Omega)^*)$ and*

$$\|\partial_t \tilde{c}_h\|_{L^2(t_1,T;H^2(\Omega)^*)} = \|d_t c_h\|_{L^2(t_1,T;H^2(\Omega)^*)} \lesssim 1,$$

independently of the mesh size and time step.

Proof. Let $w_h \in \mathscr{S}_c^j$. We recall from [1]

$$B_d(c_h^j, w_h; \breve{u}_h^j) \lesssim (1+\|\breve{u}_h^j\|^{1/2})|c_h^j|_{\mathscr{T}^j}(\|\nabla_h w_h\|_{L^4(\Omega)}+\|w_h\|_{L^4(\Omega)}+\|\sigma[w_h]\|_{\mathscr{E}_\Omega^j}),$$
$$B_{cq}(c_h^j, w_h; \breve{u}_h^j) \lesssim (1+\|\breve{u}_h^j\|^{1/2})|c_h^j|_{\mathscr{T}^j}(\|\nabla_h w_h\| + \|w_h\|_{L^4(\Omega)} + \|\sigma[w_h]\|_{\mathscr{E}_\Omega^j}),$$
$$\|\sigma[w_h]\|^2_{\mathscr{E}_\Omega^j} \lesssim (1+\|\breve{u}_h^j\|)\tilde{h}^{1/2}\|w\|^2_{H^2(\Omega)}.$$

With L^2-orthogonality and

$$\begin{aligned}\int_{t_1}^T \left(\phi\, d_t c_h^j, w\right) \mathrm{d}t &= \int_{t_1}^T \left(\bar{\hat{c}}^j \overline{q}^{I\,j}, w_h\right) - B(\overline{c_h}^j, w_h; \breve{u}_h^j)\,\mathrm{d}t \\ &\lesssim \int_0^T (1+\|\breve{u}_h^j\|)(1+\|\breve{u}_h^j\|^{1/2}_{H(\mathrm{div};\Omega)})\,|c_h^j|_{u_h^j}\,\|w\|_{H^2(\Omega)}\,\mathrm{d}t \\ &\lesssim \|w\|_{L^2(0,T;H^2(\Omega))}\end{aligned}$$

one completes the proof. □

Theorem 3. *Let $(u_i, p_i, c_i)_{i\in\mathbb{N}}$ be a sequence of numerical solutions with $(\tilde{h}_i, \tilde{k}_i) \to 0$ as $i \to \infty$. Then there exists $c \in L^2(0,T;H^1(\Omega)) \cap H^1(0,T;H^2(\Omega)^*)$ such that, after passing to a subsequence, $c_i \to c$ in $L^2(\Omega_T)$, $\partial_t \tilde{c}_i \rightharpoonup \partial_t c$ in $L^2(0,T;H^2(\Omega)^*)$ and $\nabla c_i \rightharpoonup \nabla c$ in $L^2(0,T;H^{-1}(\Omega))$. If $c_i^0, c_i^1 \to c_0$ in $H^2(\Omega)^*$ then c satisfies* (W3).

The proof is, up to the treatment of the initial conditions, exactly as in [1]. It is based on the Aubin-Lions theorem and the embedding

$$\mathscr{S}^s(\mathscr{T}_i) \hookrightarrow [\mathrm{BV}(\Omega) \cap L^4(\Omega), L^4(\Omega)]_{1/2} \hookrightarrow L^2(\Omega),$$

where $[\cdot,\cdot]_\theta$ denotes the complex method of interpolation.

Theorem 4. *Let $(u_i, p_i, c_i)_{i\in\mathbb{N}}$ be numerical solutions with $(\tilde{h}_i, \tilde{k}_i) \to 0$ and $c_i \to c$ in $L^2(\Omega_T)$ as $i \to \infty$. There exists $u \in L^\infty(0,T;H_N(\mathrm{div};\Omega))$ and $p \in L^\infty(0,T;L^2_0(\Omega))$ such that, after passing to a subsequence, $u_i \to u$ in $H_N(\mathrm{div};\Omega)$ and $p_i \to p$ in $L^2_0(\Omega)$ as $(\tilde{h}_i, \tilde{k}_i) \to 0$. Furthermore, (u,p,c) satisfies* (W1).

Proof. Use Strang's lemma, for details see [1]. □

We interpret $\breve{u}_i$ as piecewise constant function in time, attaining in $(t_{j-1}, t_j]$ the value $\frac{3}{2}u(t^{j-1}) - \frac{1}{2}u(t^{j-2})$.

Theorem 5. *Let $(u_i, p_i, c_i)_{i\in\mathbb{N}}$ be numerical solutions with $(\tilde{h}_i, \tilde{k}_i) \to 0$ as $i \to \infty$ and let $u \in L^\infty(0,T;H_N(\boldsymbol{div};\Omega))$ and $c \in L^2(0,T;H^1(\Omega)) \cap H^1(0,T;H^2(\Omega)^*)$ be a limit of $(u_i, c_i)_i$ in the sense of Theorems 3 and 4. Then (u,c) satisfies* (W2).

Proof. Let $v \in \mathscr{D}(0,T;\mathscr{C}^\infty(\Omega))$ and $v_i(t) \in \mathscr{S}_c^j$ an approximation to $v(t)$ in $(t_{j-1},t_j]$. Using the strong convergence of $(\nabla_h v_i)_i$ in $L^\infty(\Omega_T)^d$ and the weak convergence of the lifted gradient of c_i in $L^2(\Omega_T)^d$, we find

$$\int_{t_1}^T \big(\nabla c, \mathbb{D}(u)\nabla v\big)\,\mathrm{d}t = \lim_{i\to\infty} \int_{t_1}^T \big(\nabla_h c_i, \mathbb{D}_h(\breve{u}_i)\nabla_h v_i\big) - \big([c_i], \{\mathbb{D}_h(\breve{u}_i)\nabla_h v_i\}\big)_{\mathscr{E}_\Omega}\,\mathrm{d}t.$$

As in [1] it follows that $B_d(c_i, v_i; \breve{u}_i)$ coincides in the limit with $\big(\nabla c, \mathbb{D}(u)\nabla v\big)$. One can also conclude by adapting [1] that

$$\int_{t_1}^T \big(u\cdot\nabla c, v\big) + \big(q^I c, v\big)\,\mathrm{d}t = \lim_{i\to\infty} \int_{t_1}^T B_{cq}(c_i, v_i; \breve{u}_i)\,\mathrm{d}t.$$

One arrives at

$$\begin{aligned}&\int_{t_1}^T -\big(\phi\, c, \partial_t v\big) + \big(\mathbb{D}(u)\nabla c, \nabla v\big) + \big(u\cdot\nabla c, v\big) + \big(q^I c, v\big) - \big(\hat{c} q^I, v\big)\,\mathrm{d}t\\ &= \lim_{i\to\infty} \int_{t_1}^T \big(\phi\, d_t c_h^j, w_h\big) + B(\overline{c_h}^j, w_h; \breve{u}_h^j) - \big(\overline{\hat{c}}^j \overline{q}^{I\,j}, w_h\big)\,\mathrm{d}t = 0.\end{aligned}$$

Hence (W2) is satisfied for $v \in \mathscr{D}(0,T;\mathscr{C}^\infty(\Omega))$. The extension to $\mathscr{D}(0,T;H^2(\Omega))$ follows from boundedness and density of smooth functions. □

4 Numerical Experiments

The numerical experiments are carried out in two space dimensions with the lowest-order method on a mesh which consists of shape-regular triangles without hanging nodes and which is not changed over time. The diffusion–dispersion tensor takes the form

$$\mathbb{D}(u,x) = \phi(x)\,(d_m \mathrm{Id} + |u|\, d_\ell\, E(u) + |u|\, d_t\, (\mathrm{Id} - E(u))) \tag{6}$$

where $E(u)$ denotes the orthogonal projection onto the span of u.

Numerical Example 1 (Singular velocities) To examine the effect of a singular velocity field caused by a discontinuous permeability distribution and a re-entrant corner we employ the L-shaped domain Ω and $\mathbf{K}$ with $k_1 = 0.1$ and $k_2 = 10^{-6}$ as depicted in Fig. 1. The injection and production wells are located at $(1,1)$ and $(0,0)$, respectively. The porous medium is almost impenetrable in the upper left quarter, forcing a high fluid velocity at the reentrant corner where the nearly impenetrable barrier is thinnest. This leads to a singularity $|u| \sim r^{-\alpha}$, where r is the distance to the reentrant corner and $\alpha \approx 1$, cf. [1]. Figure 2 shows the concentration when the front passes the corner and at a later time. The solution c_h contains steep fronts but shows only the localised oscillations that are characteristic for dG methods.

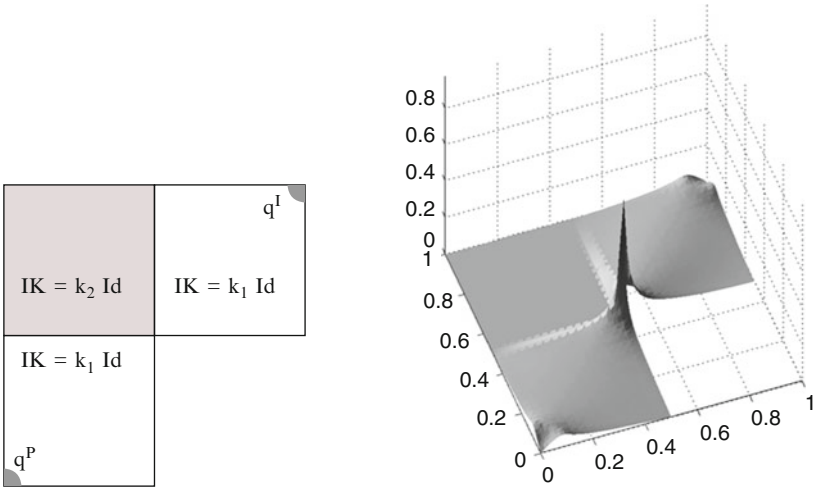


Fig. 1 *Example 1: Left*: computational domain; *right*: absolute value $|u_h|$ of the Darcy velocity at $t = 1.0$ before any interaction between the concentration front and the corner singularity

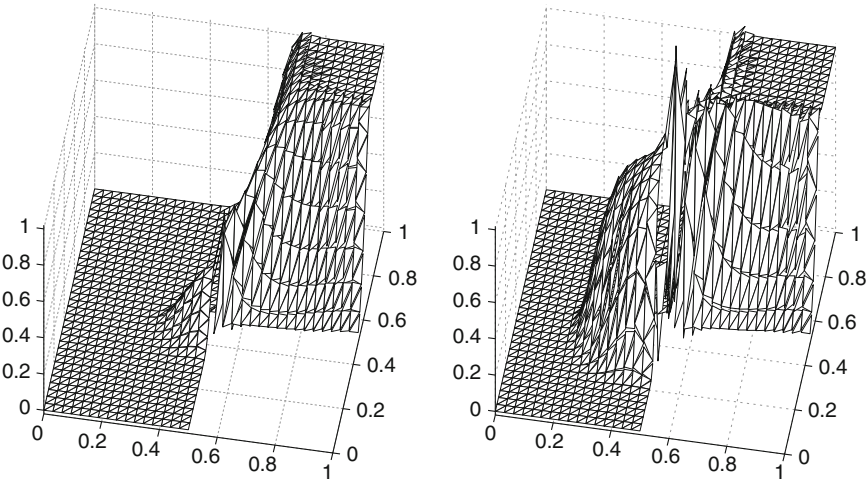


Fig. 2 Snapshots of c_h at $t = 1.5$ and 2.0, computed with the Crank–Nicolson scheme

Numerical Example 2 (Convergence rates) Convergence rates are determined by comparing the numerical solution c_h to a reference solution c_{ref} that is computed with high accuracy on a one dimensional grid. More precisely, we set $\phi = 1, \hat{c} = 1$, $\mathbf{K} = 1$ and $g = 0$ and choose Ω to be the ball $B(0, 1) \subset \mathbb{R}^2$. Using polar coordinates (r, φ), we choose $q^I = 4(1 - r)^6$ and $q^P = \frac{4}{7}r^6$. Then the Darcy velocity only

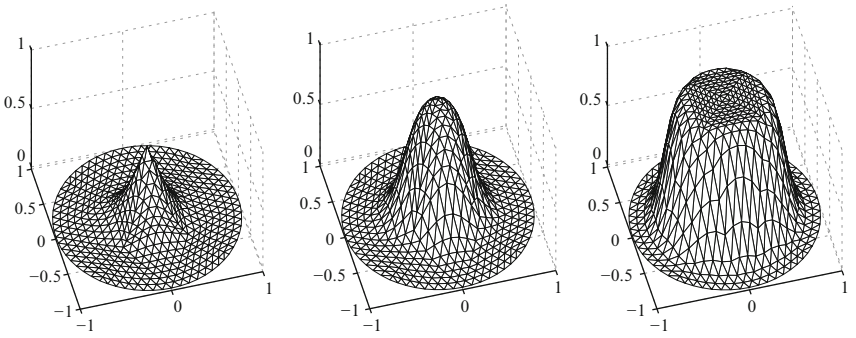


Fig. 3 *Example 2:* Snapshots of the concentration c_{ref} at $t = 0.25$, 1.0 and 3.0

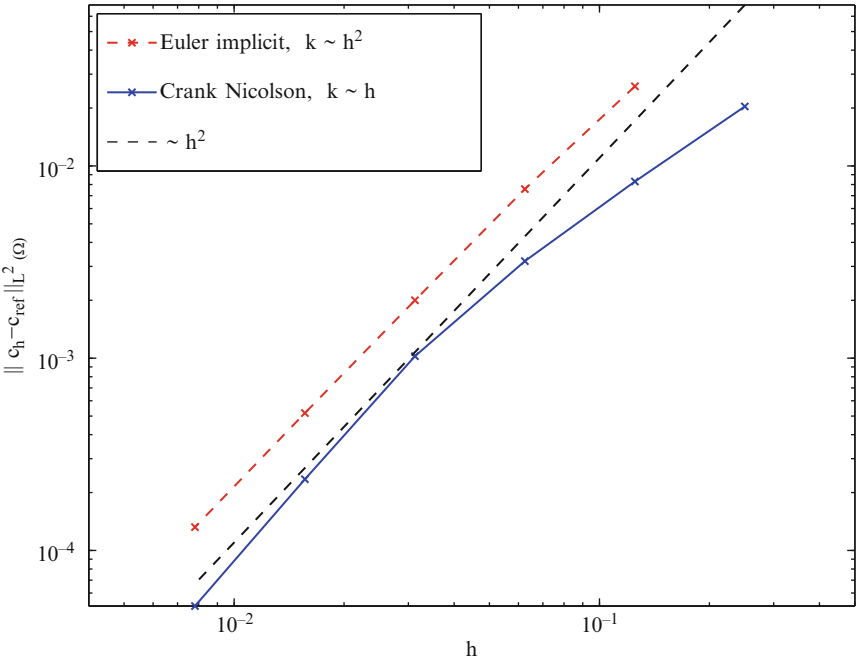


Fig. 4 Error $\|c_h - c_{\text{ref}}\|_{L^2(\Omega)}$ of the implicit Euler method the Crank–Nicolson method at time $t = 1$

changes in the radial direction and is determined by an ODE, which has the nonnegative exact solution $u(r) = \frac{r}{7}(3\,r^6 - 24\,r^5 + 70\,r^4 - 112\,r^3 + 105\,r^2 - 56\,r + 14)$. Consequently, the concentration equation reduces to a linear parabolic equation in one space dimension. Figure 3 shows snapshots of the solution c_{ref} with $d_m = 1.0 \times 10^{-5}$, $d_\ell = 4.0 \times 10^{-4}$ and Fig. 4 shows that L^2 error of implicit Euler method is of order $O(h^2 + k)$ whereas the Crank–Nicolson reaches the order $O(h^2 + k^2)$.

References

1. S. Bartels, M. Jensen, R. Müller, Discontinuous Galerkin finite element convergence for incompressible miscible displacement problems of low regularity, SIAM J. Numer. Anal. 47(5):3720–3743, 2009
2. Z. Chen, Reservoir simulation (Mathematical techniques in oil recovery), SIAM, 2007
3. X. Feng, Recent developments on modeling and analysis of flow of miscible fluids in porous media, Fluid flow and transport in porous media, Contemp. Math. 295:229–224, 2002
4. B. Rivière, N. Walkington, Convergence of a Discontinuous Galerkin Method for the Miscible Displacement Equations Under Minimal Regularity, 2009, http://www.math.cmu.edu/~noelw/Noelw/Papers/RiWa09.pdf
5. V. Thomée, Galerkin finite element methods for parabolic problems, Springer Series in Computational Mathematics 25, 1997

Simulations of 3D/4D Precipitation Processes in a Turbulent Flow Field

Volker John and Michael Roland

Abstract Precipitation processes are modeled by population balance systems. An expensive part of their simulation is the solution of the equation for the particle size distribution (PSD) since this equation is defined in a higher-dimensional domain than the other equations in the system. This paper studies different approaches for the solution of this equation: two finite difference upwind schemes and a linear finite element flux-corrected transport method. It is shown that the different schemes lead to qualitatively different solutions for an output of interest.

1 Introduction

Precipitation processes are very important in the chemical industry. Already a decade ago, over 50% of the products in chemical engineering were produced in particulate form [18]. Since that time, the importance of particulate products has been even increased. Nowadays, the main focus is on the production of particles with prescribed characteristics, such as size, shape or chemical properties. The numerical simulation of precipitation processes will make an essential contribution to the optimization of production processes.

Isothermal precipitation processes are modeled by a coupled system of the Navier–Stokes equations to describe the flow field, of convection–diffusion-reaction equations to describe the transport and the reaction of the chemical species, and of a transport equation for the particle size distribution (PSD).

V. John (✉)
Weierstrass Institute for Applied Analysis and Stochastics (WIAS), Mohrenstr. 39, 10117 Berlin, Germany
and
Free University of Berlin, Department of Mathematics and Computer Science, Arnimallee 6, 14195 Berlin, Germany
e-mail: john@wias-berlin.de

M. Roland
FR 6.1 - Mathematics, Saarland University, Postfach 15 11 50, 66041 Saarbrücken, Germany
e-mail: roland@math.uni-sb.de

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_51,

The flow field in applications is often turbulent. The numerical simulation of turbulent flows is by itself an active field of research [1, 6, 17]. In the numerical simulations presented in this paper, a finite element variational multiscale (VMS) method will be used [7, 8]. VMS methods are a rather new approach for turbulence modeling, which were derived from general principles for simulating multiscale phenomena [3, 4].

A chemical reaction happens in the flow field which is modeled by a nonlinear system of convection–diffusion-reaction equations. These equations are convection- and reaction-dominated. Also the numerical simulation of this type of equations is by itself an active field of research [16].

The main feature of precipitation processes is the nucleation of particles if the concentration of a species exceeds a saturation concentration. In applications, not the behavior of the individual particles is of interest, but the PSD. The PSD depends not only on time and space, but also on properties of the particles, so-called internal coordinates. Thus, the transport equation for the PSD is defined in a higher-dimensional domain than the other equations of the coupled system.

The simulations presented in this paper will consider the flow of a dilute solution. Hence, the effect of the particles on the flow field are negligible. Nucleation and growth of particles, which are the most important chemical mechanisms in a precipitation process, are included into the used model. Breakage and agglomeration of particles will not be considered, because they are of much less importance. The growth process of the particles in this model is realized by layering [15]. The PSD has one internal coordinate, namely the diameter of the particles.

The simulation of complex coupled systems is generally time-consuming. In simulations of precipitation processes, a very expensive part might be the solution of the higher-dimensional PSD equation. This paper studies different schemes for discretizing this equation: on the one hand rather inexpensive but also inaccurate schemes and on the other hand a more expensive but also more accurate scheme. It will be demonstrated that the use of the different schemes leads to qualitatively different results for an output of interest.

2 The Model of the Precipitation Process

For shortness of presentation, we will give here only the non-dimensionalized model, see [10, 13] for its derivation.

Let Ω be the flow domain and T a final time. We will consider a dilute fluid, i.e., the number of particles and their size are sufficiently small such that their influence on the flow field can be neglected. Then, the Navier–Stokes equations are given by

$$\frac{\partial \mathbf{u}}{\partial t} - \frac{1}{Re}\Delta \mathbf{u} + (\mathbf{u}\cdot\nabla)\mathbf{u} + \nabla p = \mathbf{0} \text{ in } (0,T]\times\Omega, \tag{1}$$

$$\nabla\cdot\mathbf{u} = 0 \text{ in } [0,T]\times\Omega, \tag{2}$$

where $\mathbf{u}$ is the velocity, p is the pressure and $Re = u_\infty l_\infty / \nu$ is the Reynolds number with l_∞ being a characteristic length scale and u_∞ a characteristic velocity scale of the problem. Let c_A and c_B be the concentrations of the reactants A and B, then their reaction is described by the following equations

$$\frac{\partial c_i}{\partial t} - \frac{D_i}{u_\infty l_\infty} \Delta c_i + \mathbf{u} \cdot \nabla c_i + k_R \frac{l_\infty c_\infty}{u_\infty} c_A c_B = 0 \text{ in } (0, T] \times \Omega, \tag{3}$$

$i \in \{A, B\}$, where D_i is a diffusion coefficient, k_R is the reaction rate constant and c_∞ is a characteristic concentration scale for the reactants. The equation for the concentration of the dissolved product C is given by

$$\begin{aligned} &\frac{\partial c_C}{\partial t} - \frac{D_C}{u_\infty l_\infty} \Delta c_C + \mathbf{u} \cdot \nabla c_C - \Lambda_{\text{chem}} c_A c_B + \Lambda_{\text{nuc}} \max\left\{0, (c_C - 1)^5\right\} \\ &\quad + \left(c_C - \frac{c_{C,\infty}^{\text{sat}}}{c_{C,\infty}}\right) \int_{d_{\text{p,min}}}^{1} d_{\text{p}}^2 f \, \mathrm{d}(d_{\text{p}}) = 0 \text{ in } (0, T] \times \Omega. \end{aligned} \tag{4}$$

Here, D_C is a diffusion constant, d_{p} describes the size of the particles, f denotes the PSD, $c_{C,\infty}^{\text{sat}}$ is the saturation concentration of the dissolved product C, and $c_{C,\infty} = c_{C,\infty}^{\text{sat}} \exp(C_2 / \tilde{d}_{\text{p},0})$ is a characteristic concentration scale for C with C_2 being a model constant and $\tilde{d}_{\text{p},0}$ ($\tilde{d}_{\text{p,max}}$) being the smallest (largest) possible particle diameter. The parameters in (4) are

$$\Lambda_{\text{chem}} = k_R \frac{c_\infty^2 l_\infty}{c_{C,\infty} u_\infty}, \quad d_{\text{p,min}} = \frac{\tilde{d}_{\text{p},0}}{d_{\text{p},\infty}}, \quad \Lambda_{\text{nuc}} = C_{\text{nuc}} d_{\text{p,min}}^3 d_{\text{p},\infty}^3 k_{\text{nuc}} \frac{l_\infty c_{C,\infty}^4}{u_\infty},$$

with $d_{\text{p},\infty}$ being an upper bound for the largest possible particle diameter and C_{nuc} and k_{nuc} are constants in the model for the nucleation process. To obtain the last term on the left hand side of (4) in the presented form, the characteristic scale of the PSD $f_\infty = u_\infty / (C_G k_G d_{\text{p},\infty}^3 l_\infty)$ was used, where C_G is a constant to model the growth of the particles and k_G is a growth rate constant. The last equation describes the PSD

$$\frac{\partial f}{\partial t} + \mathbf{u} \cdot \nabla f + G(c_C) \frac{l_\infty}{u_\infty d_{\text{p},\infty}} \frac{\partial f}{\partial d_{\text{p}}} = 0 \text{ in } (0, T] \times \Omega \times \left(d_{\text{p,min}}, \frac{\tilde{d}_{\text{p,max}}}{d_{\text{p},\infty}}\right) \tag{5}$$

with the growth rate $G(c_C) = k_G c_{C,\infty} (c_C - c_{C,\infty}^{\text{sat}} / c_{C,\infty})$.

In summary, the system of (1)–(3) for c_A, (3) for c_B, (4) and (5) has to be solved. All equations have to be equipped with initial and boundary conditions.

3 The Applied Numerical Methods

The Crank–Nicolson scheme with an equidistant time step is applied as temporal discretization for (1)–(4).

In the considered system, the velocity $\mathbf{u}$ is needed in all other equations but the other quantities do not influence the Navier–Stokes equations. For this reason, a straightforward approach consists in solving at each discrete time first (1) and (2). The velocity is approximated with the Q_2 finite element and the pressure with the P_1^{disc} finite element, i.e., with discontinuous piecewise linears. The simulations will study a turbulent flow, hence a turbulence model has to be applied. We will use the projection-based finite element variational multiscale method from [7, 8] with a piecewise constant large scale space.

With the obtained velocity field, the system for the concentrations c_A and c_B can be solved. This is done by a fixed point iteration. The linearized equations are discretized with the Q_1 finite element. They are strongly convection-dominated such that a stabilization has to be applied. Comparative studies of stabilized finite element schemes [11,12] have shown that for the Q_1 finite element FEM-FCT (flux-corrected transport) schemes outperform more standard approaches like SUPG. In the simulations presented in Sect. 4, a linear FEM-FCT scheme from [14] is used.

After having computed c_A and c_B, a coupled system for c_C and f remains. This system is decoupled and linearized in our approach by treating (4) in a semi-implicit way, namely by using c_C and the PSD f from the previous discrete time in the last two terms on the left hand side of (4). Thus (4) becomes a linear equation in each discrete time, which is solved also with a linear Q_1-FEM-FCT scheme.

The emphasis of the numerical studies is on the schemes for the PSD equation (5). This equation is given in a 4D domain and its solution might be rather expensive. For this reason, one can think about using comparatively cheap but also rather inaccurate schemes for (5). The first scheme of this kind which we apply is the forward Euler simple upwind finite difference scheme. The second scheme, the backward Euler simple upwind finite difference scheme, is only somewhat more expensive. The results and the costs of these schemes will be compared with the much more expensive linear Q_1-FEM-FCT scheme.

4 Numerical Studies

The calcium carbonate precipitation $Na_2CO_3 + CaCl_2 \longrightarrow CaCO_3 \downarrow + 2NaCl$ is considered in the numerical studies. The parameters of this process are given by

- $\nu = 10^{-6}\ \mathrm{m^2/s}$
- $\rho = 1\ \mathrm{kg/m^3}$
- $k_G = 10^{-7}\ \mathrm{m^4/kmol\ s}$
- $k_{\text{nuc}} = 10^{24}\ (1/\mathrm{m^3\ s})/(\mathrm{kmol/m^3})^5$
- $k_R = 10^{-2}\ \mathrm{m^3/kmol\ s}$
- $c_{C,\infty}^{\text{sat}} = 1.37\ 10^{-4}\ \mathrm{kmol/m^3}$
- $C_2 = 7.2\ 10^{-9}\ \mathrm{m}$
- $C_G = 45.98\ \mathrm{kmol/m^3}$

- $C_{\text{nuc}} = 15.33\ {}^{\text{kmol}}/_{\text{m}^3}$
- $D_A = D_B = D_C = 1.5\ 10^{-9}\ {}^{\text{m}^2}/_{\text{s}}$
- $\tilde{d}_{\text{p},0} = 10^{-9}$ m
- $\tilde{d}_{\text{p,max}} = 10^{-4}$ m.

The following reference quantities have been used in the dimensionless equations:

- $l_\infty = 1$ m
- $u_\infty = 10^{-2}\ {}^{\text{m}}/_{\text{s}}$
- $t_\infty = 10^2$ s
- $c_\infty = 1\ {}^{\text{kmol}}/_{\text{m}^3}$
- $c_{C,\infty} = 0.183502\ {}^{\text{kmol}}/_{\text{m}^3}$
- $d_{\text{p},\infty} = 10^{-4}$ m
- $f_\infty = 2.17486\ 10^{15}\ {}^{1}/_{\text{m}^4}$.

Concerning the flow, a situation similar to a driven cavity problem is considered. The flow domain is $(0,1)^3$. There are opposite inlets at $\{0\} \times (0.4375, 0.5625) \times (0.4375, 0.5625)$ and $\{1\} \times (0.4375, 0.5625) \times (0.4375, 0.5625)$ and an outlet at $(0.4375, 0.5625) \times (0.4375, 0.5625) \times \{0\}$. A situation like this is sometimes called T-mixer. The inflows are given by a profile which was precomputed by solving the Poisson equation with right hand side equal to the constant 1 on the inlets and with homogeneous Dirichlet boundary conditions. On the top of the cavity, the velocity $(1,0,0)^T$ is prescribed, outflow boundary conditions are given at the outlet and no slip boundary conditions on the remaining boundaries. Initially, the fluid was considered to be at rest and an impulsive start was performed. The Reynolds number of the flow is $Re = 10{,}000$. Even the driven cavity problem without inlets and outlet is a turbulent flow at this Reynolds number [2,5].

All concentrations inside the domain were zero at the initial time. On the boundary, the concentrations of the reactants A at the left inlet and B at the right inlet were set to 1 for all times. Homogeneous Neumann boundary conditions were used on all other parts of the boundary. For the substance C, homogeneous Neumann boundary conditions were applied on the whole boundary. The boundary condition for the PSD with respect to the internal coordinate was

$$
\begin{aligned}
f(t, x_1, x_2, x_3, d_{\text{p,min}}) &= \frac{B_{\text{nuc}}(c_C)}{f_\infty G(c_C)} && \text{if } G(c_C(t, x_1, x_2, x_3)) > 0, \\
f(t, x_1, x_2, x_3, d_{\text{p,min}}) &= 0 && \text{if } G(c_C(t, x_1, x_2, x_3)) = 0, \\
f(t, x_1, x_2, x_3, d_{\text{p,max}}) &= 0 && \text{if } G(c_C(t, x_1, x_2, x_3)) < 0,
\end{aligned}
$$

with the nucleation rate $B_{\text{nuc}}(c_C) = k_{\text{nuc}} c_{C,\infty}^5 \max\{0, (c_C - 1)^5\}$. With respect to the spatial coordinates, the PSD was set to be zero at the closure of the fluid flow inlets (no particles enter through the inlets).

The length of the time step was set to be $\Delta t = 0.001$. In space, a $16 \times 16 \times 16$ uniform mesh consisting of cubes was used which leads to 107,811 velocity d.o.f., 16,384 pressure d.o.f. and 4,913 d.o.f. for the concentrations. The internal coordinate was discretized with 16, 32 or 64 levels (83,521, 162,129 or 319,345 d.o.f.), were the mesh was finer for small diameters. Figure 1 shows the flow field and isosurfaces of the concentrations at around the starting time of the precipitation process. The flow field and the concentrations c_A, c_B, are always the same for all discretizations of the PSD equation. Until the start of the precipitation, i.e., until

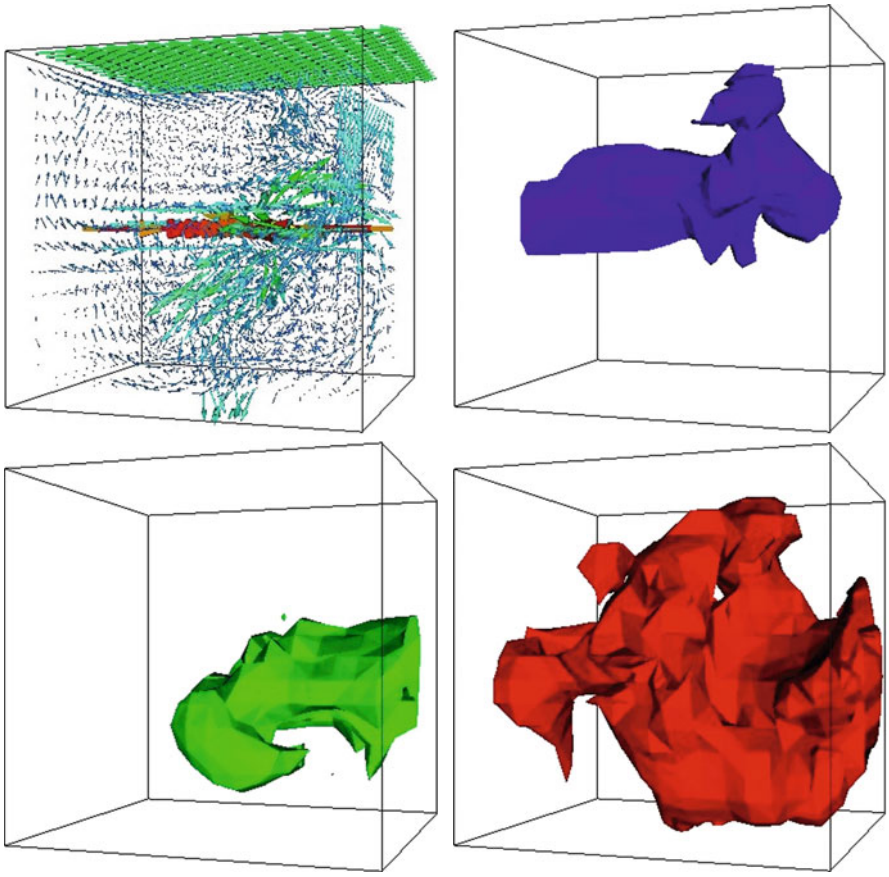


Fig. 1 Velocity field, isosurfaces for $c_A = 0.25$, $c_B = 0.25$, $c_C = 0.7$ at $t = 3{,}200$, *left* to *right*, *top* to *bottom*

the back coupling of f onto c_C starts, the concentration c_C is also identical for all discretizations of (5).

An output of interest is the median of the volume fraction at the center of the outlet. The volume fraction and the cumulative volume fraction are given by

$$q_3(\tilde{t}, \tilde{d}_\mathrm{p}) := \frac{\tilde{d}_\mathrm{p}^3 \tilde{f}(\tilde{t}, 0.5, 0.5, 0, \tilde{d}_\mathrm{p})}{\int_{\tilde{d}_{\mathrm{p},0}}^{\tilde{d}_{\mathrm{p,max}}} \tilde{d}_\mathrm{p}^3 \tilde{f}(\tilde{t}, 0.5, 0.5, 0, \tilde{d}_\mathrm{p}) \, \mathrm{d}(\tilde{d}_\mathrm{p})}, \quad Q_3(\tilde{t}, \tilde{d}_\mathrm{p}) := \int_{\tilde{d}_{\mathrm{p},0}}^{\tilde{d}_\mathrm{p}} q_3(\tilde{t}, \tilde{d}) \, \mathrm{d}\tilde{d}.$$

Then, the median of the volume fraction is defined to be the particle size for which $Q_3(\tilde{t}, \tilde{d}_\mathrm{p})$ takes the value 0.5: $\tilde{d}_{\mathrm{p},50}(\tilde{t}) := \{\tilde{d}_\mathrm{p} \, : \, Q_3(\tilde{t}, \tilde{d}_\mathrm{p}) = 0.5\}$.

The temporal developments of $\tilde{d}_{\mathrm{p},50}(\tilde{t})$ for the different schemes of discretizing the PSD equation (5) are presented in Fig. 2. It can be seen that the first particles

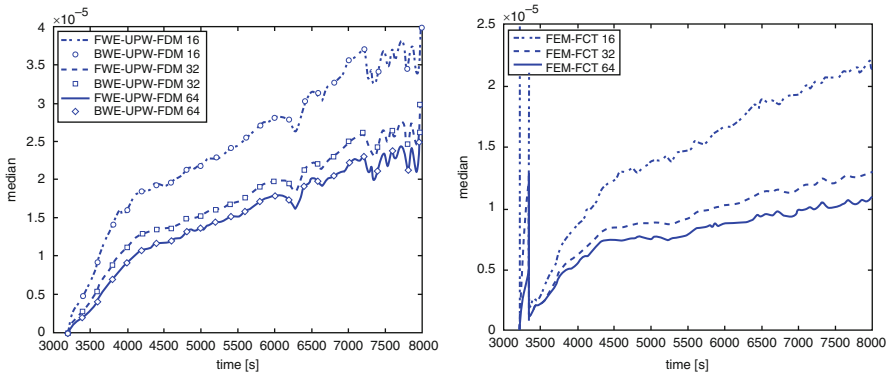


Fig. 2 Median of the volume fraction at the center of the outlet, the number of levels for discretizing the internal coordinate is given behind the schemes

Table 1 Average computing times per time step in s, different numbers of levels for discretizing the internal coordinate, computer with Intel Xeon CPU, 2.4 GHz

Scheme for solving (5)	16 levels	32 levels	64 levels
Forward Euler upwind FDM	19	19	21
Backward Euler upwind FDM	19	19	23
Crank–Nicolson FEM-FCT	26	33	64

reach the center of the outlet at around 3,200 s. The main observation is that the different numerical schemes lead to qualitatively different results. The predicted median of the volume fraction at 8,000 s is around twice as large for the finite difference Euler schemes compared with the FEM-FCT scheme. For all schemes, the median decreases on finer meshes for the internal coordinate.

In a numerical study at a coupled 2D/3D problem with prescribed solution in [10], it has been shown that the FEM-FCT scheme leads to considerably more accurate results than the finite difference upwind schemes. Based on this experience, it can be expected that the results with the FEM-FCT scheme for solving the PSD equation are more reliable.

The numerical studies were performed with the code `MooNMD` [9]. This is a flexible research code which is not tailored for solving population balance systems. The average computing times for a time step are given in Table 1. It can be observed that the solution of the PSD equation is much more time-consuming for the linear FEM-FCT scheme, in particular on fine grids for the PSD equation.

Some numerical studies for 2D/3D population balance systems with a structured laminar flow field were performed in [10]. It was shown that in this case the results obtained with the different discretization schemes for the PSD equation were rather similar. This suggests that the reason for the qualitatively different results is the presence of the turbulent flow field, compare also with some other studies in [10].

5 Summary

The paper studied different discretization schemes for the higher-dimensional transport equation in a 3D/4D population balance system with a turbulent flow field. It was demonstrated that the usage of inexpensive but inaccurate schemes on the one hand and a more expensive but also more accurate scheme on the other hand leads to qualitatively different results for an output of interest. This is similar to the observations in simulations of 2D/3D population balance systems with a highly time-dependent flow field in [10].

The presented results demonstrate that outputs of interest in the simulation of complex processes might highly depend on the applied numerical schemes. They emphasize the need of using accurate schemes and the necessity of implementing them such that they work efficiently. Our future work will focus on these topics.

References

1. L.C. Berselli, T. Iliescu, and W.J. Layton. *Mathematics of Large Eddy Simulation of Turbulent Flows*. Springer, Berlin, 2006
2. V. Gravemeier, W.A. Wall, and E. Ramm. Large eddy simulation of turbulent incompressible flows by a three–level finite element method. *Int. J. Numer. Meth. Fluids*, 48:1067–1099, 2005
3. J.-L. Guermond. Stabilization of Galerkin approximations of transport equations by subgrid modeling. *M2AN*, 33:1293–1316, 1999
4. T.J.R. Hughes. Multiscale phenomena: Green's functions, the Dirichlet-to-Neumann formulation, subgrid-scale models, bubbles and the origin of stabilized methods. *Comp. Meth. Appl. Mech. Eng.*, 127:387–401, 1995
5. T. Iliescu, V. John, W.J. Layton, G. Matthies, and L. Tobiska. A numerical study of a class of LES models. *Int. J. Comput. Fluid Dyn.*, 17:75–85, 2003
6. V. John. *Large Eddy Simulation of Turbulent Incompressible Flows. Analytical and Numerical Results for a Class of LES Models*, volume 34 of *Lecture Notes in Computational Science and Engineering*. Springer, Berlin, 2004
7. V. John and S. Kaya. A finite element variational multiscale method for the Navier-Stokes equations. *SIAM J. Sci. Comp.*, 26:1485–1503, 2005
8. V. John and A. Kindl. A variational multiscale method for turbulent flow simulation with adaptive large scale space. *J. Comput. Phys.*, 229:301–312, 2010
9. V. John and G. Matthies. MooNMD – a program package based on mapped finite element methods. *Comput. Visual. Sci.*, 6:163–170, 2004
10. V. John, M. Roland. On the impact of the scheme for solving the higher-dimensional equation in coupled population balance systems. *Int. J. Numer. Methods Engrg.*, 82:1450–1474, 2010
11. V. John and E. Schmeyer. Stabilized finite element methods for time–dependent convection–diffusion-reaction equations. *Comput. Meth. Appl. Mech. Eng.*, 198:475–494, 2008
12. V. John and E. Schmeyer. On finite element methods for 3d time-dependent convection–diffusion-reaction equations with small diffusion. In *BAIL 2008 – Boundary and Interior Layers*, volume 69 of *Lecture Notes in Computational Science and Engineering*, pages 173–182. Springer, Berlin, 2009
13. V. John, T. Mitkova, M. Roland, K. Sundmacher, L. Tobiska, and A. Voigt. Simulations of population balance systems with one internal coordinate using finite element methods. *Chem. Eng. Sci.*, 64:733–741, 2009
14. D. Kuzmin. Explicit and implicit FEM–FCT algorithms with flux linearization. *J. Comput. Phys.*, 228:2517–2534, 2009

15. K.C. Link and E.-U. Schlünder. Wirbelschicht–Sprühgranulation – Untersuchung des Coatingvorganges am frei schwebenden Einzelpartikel. *Chemie Ingenieur Technik*, 68:1139, 1996
16. H.-G. Roos, M. Stynes, and L. Tobiska. *Robust Numerical Methods for Singularly Perturbed Differential Equations*, volume 24 of *Springer Series in Computational Mathematics*, 2nd edition. Springer, Berlin, 2008
17. P. Sagaut. *Large Eddy Simulation for Incompressible Flows*, 3rd edition. Springer, Berlin, 2006
18. K. Wintermantel. Process and product engineering – achievements, present and future challenges. *Chem. Eng. Sci.*, 54:1601–1620, 1999

2D Finite Volume Lagrangian Scheme in Hyperelasticity and Finite Plasticity

Gilles Kluth and Bruno Després

Abstract System of conservation laws develop discontinous solutions, which can be captured by conservative and consistent Finite Volume schemes. In Lagrangian schemes, the mesh is moving; therefore material interfaces are well simulated. Cell-centered Lagrangian Finite Volume schemes have been recently developed in compressible hydrodynamic [J. Comput. Phys. 228:5160–5183, 2009, Comptes Rendus Académie des Sciences 331:327–372, 2003, Siam J. Sci. Comp. 29, 2007]. This paper shows how to extend these schemes to material strength. Moreover, we show that with an appropriate equation of state, this extension allows to simulate some plastic phenomenons.

In the first section, we present recent cell-centered Lagrangian Finite Volume schemes in compressible hydrodynamic [1, 3, 9]. These schemes are conservative and entropic. Our presentation is based on [1]. In the second section, we extend this scheme to material strength. We discretize the hyperelastic system, preserving the qualities of the previous hydrocode. A major issue is the discretization of the Jacobian matrix which has to be compatible with the usual discretization of its determinant. In the third section, we present an equation of state [7] which is an easy way to constrain the stress in a hyperelastic scheme, as it is the case in plasticity. Finally, numerical results validate the hyperelastic scheme, and shows plastic phenomenons, introduced by the previous equation of state.

G. Kluth (✉)
CEA, DAM, DIF, F-91297 Arpajon, France
e-mail: gkluth@gmail.com

B. Després
Laboratoire Jacques-Louis Lions, Université Pierre et Marie Curie, Boîte courrier 187, 75252 Paris Cedex 05, France
e-mail: despres@ann.jussieu.fr

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_52,

1 Finite Volume Lagrangian Scheme in Compressible Hydrodynamic

In compressible hydrodynamic, the conservation laws of momentum and energy

$$\rho D_t \begin{pmatrix} \mathbf{v} \\ e \end{pmatrix} = div \begin{pmatrix} -p\mathbf{I} \\ -p\mathbf{v} \end{pmatrix} \tag{1}$$

are discretized with cell-centered Lagrangian Finite Volume methods by

$$\frac{M_j}{\Delta t} \begin{pmatrix} \mathbf{v}_j^{n+1} - \mathbf{v}_j \\ e_j^{n+1} - e_j \end{pmatrix} = \sum_r \begin{pmatrix} -p_{jr}\mathbf{C}_{jr} \\ -p_{jr}\mathbf{C}_{jr}.\mathbf{v}_r \end{pmatrix}. \tag{2}$$

The derivative in time D_t is the material derivative (along the flow). $\mathbf{v}$ is the velocity, $e = \epsilon + \frac{\mathbf{v}^2}{2}$ the total energy per mass, ρ is the mass density. The hydrodynamic equation of state $\epsilon = \phi_1(\tau, S)$ gives the internal energy as a function of specific volume $\tau = 1/\rho$ and entropy S. The pressure is $p = -\partial\phi_1/\partial\tau$.

The mesh is made of moving cells (index j) of constant mass M_j. Nodes (index r) are moved during the timestep $\Delta t = t^{n+1} - t^n$ with

$$\mathbf{x}_r^{n+1} = \mathbf{x}_r + \Delta t \mathbf{v}_r \tag{3}$$

and finally

$$\rho_j^{n+1} = \frac{M_j}{V_j^{n+1}} \tag{4}$$

V_j^{n+1} being the volume of the moved cell j. The vector $\mathbf{C}_{jr}$ is defined by

$$\mathbf{C}_{jr} = \nabla_{\mathbf{x}_r} V_j \tag{5}$$

so we have

$$V_j'(t) = \sum_r \mathbf{C}_{jr}.\mathbf{v}_r, \tag{6}$$

or with $V_j = M_j \tau_j$

$$M_j \tau_j'(t) = \sum_r \mathbf{C}_{jr}.\mathbf{v}_r. \tag{7}$$

The nodal solver gives p_{jr} and $\mathbf{v}_r$. It is chosen to satisfy 2 properties: the production of entropy and the conservativity. This is a major point for stability of the scheme and selection of entropic weak solutions.

2 Finite Volume Lagrangian Scheme in Hyperelasticity

To extend the preceding schemes to hyperelasticity, we have to discretize the jacobian matrix (or deformation gradient). Then we can generalize the hydrodynamic discretization to material strength.

2.1 Discretization of the Jacobian Matrix

2.1.1 The Jacobian Matrix

The Jacobian matrix discretization is necessary in the following situations: for any scheme in solid mechanics, based on deformations; for changes in coordinates (for example to pass from Eulerian to Lagrangian coordinates); for the characterization of the geometry of a deformed mesh in the context of Lagrangian schemes.

The motion brings a point in initial coordinates $\mathbf{X}$ to its actual coordinates $\mathbf{x} = \boldsymbol{\psi}(t, \mathbf{X})$. The Jacobian matrix $\mathbf{F} = \nabla_{\mathbf{X}} \boldsymbol{\psi}$ and the velocity $\mathbf{v} = D_t \boldsymbol{\psi}$ are linked by the following compatibility relations

$$D_t \mathbf{F} = \nabla_{\mathbf{X}} \mathbf{v}. \tag{8}$$

The mass conservation gives

$$J = \det(\mathbf{F}) = \rho_0 \tau \tag{9}$$

ρ_0 being the initial mass density.

2.1.2 Discretization

Two issues arise concerning the discretization of the Jacobian matrix. First, we want a discretization which uses the mesh at the beginning of the timestep, and not at the initial one : that's why we want an Eulerian equivalent for (8). Second, with (7) and (9) we have two definitions of τ which have to be compatible.

By chain rule, we have (see Fig. 1)

$$\mathbf{F}_j = \mathbf{G}_j \, \mathbf{F}_j^n. \tag{10}$$

Thus, we suggest to discretize at time $t > t^n$

$$\begin{cases} D_t \, \mathbf{G} = \nabla_{\mathbf{x}_n} \, \mathbf{v}, \\ \mathbf{G}(t^n) = \mathbf{I}, \\ \mathbf{F} = \mathbf{G} \, \mathbf{F}. \end{cases} \quad \text{in spite of} \quad \begin{cases} D_t \mathbf{F} = \nabla_{\mathbf{X}} \, \mathbf{v}, \\ \mathbf{F}(t^0) = \mathbf{I}. \end{cases} \tag{11}$$

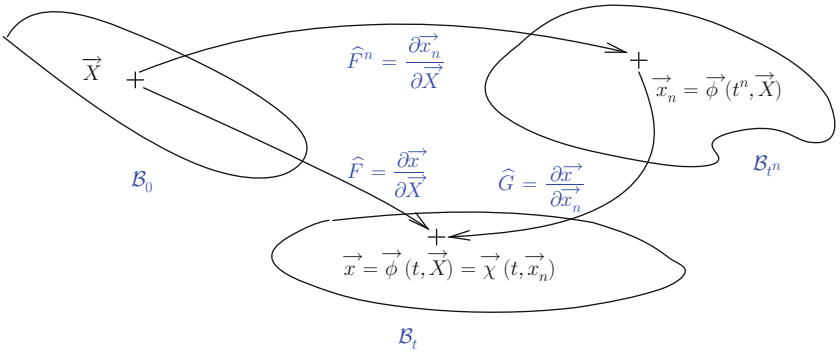


Fig. 1 Definition of the Jacobian matrix $\mathbf{G}$, from the configuration $\mathscr{B}_{t^n}$ at time t^n with coordinates $\mathbf{x_n}$, to the configuration $\mathscr{B}_t$ at time t with coordinates $\mathbf{x}$. The Jacobian matrix $\mathbf{F}$ characterizes the motion from configuration $\mathscr{B}_0$ to $\mathscr{B}_t$, $\mathbf{F^n}$ from $\mathscr{B}_0$ to $\mathscr{B}_{t^n}$

This gives the discretization

$$\mathbf{G}_j^{n+1} = \mathbf{I} + \frac{\Delta t}{V_j} \sum_r \mathbf{v}_r \otimes \mathbf{C}_{jr}, \tag{12}$$

and with (10)

$$\mathbf{F}_j^{n+1} = \mathbf{F}_j + \frac{\Delta t}{V_j} \sum_r \left(\mathbf{v}_r \otimes \mathbf{C}_{jr}\right) \mathbf{F}_j, \tag{13}$$

Proposition 1. *The discretization of the Jacobian matrix (13) with* $\tau_j = \frac{det(\mathbf{F})}{\rho_{0j}}$ *gives the discretization (7) at first order in time*

Proof. For simplicity, we do the proof in 2D. In 3D, the result remains true. Using (13) we have (we omitt indexes j and n)

$$\begin{aligned}(F_{11}F_{22}-F_{21}F_{12})^{n+1} = {} & F_{11}F_{22} - F_{21}F_{12} + F_{11}\left(\frac{\Delta t}{V_j}\sum_r v_1(C_1F_{11} + C_2F_{21})\right) \\ & + F_{22}\left(\frac{\Delta t}{V_j}\sum_r v_2(C_1F_{12} + C_2F_{22})\right) - F_{12}\left(\frac{\Delta t}{V_j}\sum_r v_1(C_1F_{12} + C_2F_{22})\right) \\ & - F_{21}\left(\frac{\Delta t}{V_j}\sum_r v_2(C_1F_{11} + C_2F_{21})\right) + O((\Delta t)^2).\end{aligned}$$

which gives

$$\tau_j^{n+1} = \tau_j + \frac{\Delta t}{\rho_{0j} V_j} \sum_r det(\mathbf{F_j})\mathbf{v}_r.\mathbf{C}_{jr} + O((\Delta t)^2) \tag{14}$$

and thus the result.

2.2 The Hyperelastic Scheme

The hyperelastic system is given by [2]

$$\rho D_t \begin{pmatrix} \mathbf{v} \\ e \end{pmatrix} = div \begin{pmatrix} \boldsymbol{\sigma} \\ \mathbf{v}^t \boldsymbol{\sigma} \end{pmatrix} \tag{15}$$

with the equation of state $\epsilon = \phi_2(\mathbf{F}, S)$ and the Cauchy stress tensor $\boldsymbol{\sigma} = \rho \nabla_{\mathbf{F}}(\phi_2)\mathbf{F}^t$. It is discretized by

$$\frac{M_j}{\Delta t} \begin{pmatrix} \mathbf{v}_j^{n+1} - \mathbf{v}_j \\ e_j^{n+1} - e_j \end{pmatrix} = \sum_r \begin{pmatrix} \boldsymbol{\sigma} C_{jr} \\ \boldsymbol{\sigma} C_{jr}.\mathbf{v}_r \end{pmatrix}. \tag{16}$$

We note $\boldsymbol{\sigma} C_{jr}$, and not $\boldsymbol{\sigma}_{jr} \mathbf{C}_{jr}$ because we keep the possibility to have a flux which is not aligned with $\mathbf{C}_{jr}$. We need now to determine the nodal stress $\boldsymbol{\sigma} C_{jr}$ and the nodal velocity $\mathbf{v_r}$.

Proposition 2. *By using the following nodal solver (which gives nodal quantities $\boldsymbol{\sigma} C_{jr}$ and $\mathbf{v_r}$)*

$$\boldsymbol{\sigma} C_{jr} = \boldsymbol{\sigma}_j \mathbf{C}_{jr} + \left|\mathbf{C}_{jr}\right| \mathbf{Q}_{jr}(\mathbf{v_r} - \mathbf{v}_j)$$

$$\sum_j \boldsymbol{\sigma} C_{jr} = \mathbf{0}$$

$\mathbf{Q}_{jr}$ *being symmetric and non negative, the scheme is conservative and entropic.*

The conservativity is immediately given by the second part of the solver. The proof of the entropic criterion is given in [6]. More precisely, it is shown that for the semi-discrete scheme (continous in time) we have

$$(M_j \partial_S \phi_2) S_j'(t) = \sum_r \left(\boldsymbol{\sigma} C_{jr} - \boldsymbol{\sigma}_j \mathbf{C}_{jr}\right) . \left(\mathbf{v_r} - \mathbf{v}_j\right). \tag{17}$$

In the hydrodynamic case $\epsilon = \phi_1(\tau, S)$, we have $\boldsymbol{\sigma} = -p\mathbf{I}$ and if we take

- $\mathbf{Q}_{jr} = \frac{(\rho c)_j}{\|\mathbf{C}_{jr}\|^2} \mathbf{C}_{jr} \otimes \mathbf{C}_{jr}$, we obtain the hydrocode [1] or [3],
- $\mathbf{Q}_{jr} = \frac{(\rho c)_j}{2\|\mathbf{C}_{jr}\|^2} \left(\frac{l_{r-1,r}}{|\mathbf{C}_{jr}|} \mathbf{n}_{r-1,r} \otimes \mathbf{n}_{r-1,r} + \frac{l_{r,r+1}}{|\mathbf{C}_{jr}|} \mathbf{n}_{r,r+1} \otimes \mathbf{n}_{r,r+1} \right)$, we obtain the hydrocode [9] or [8].

Above, $c_j > 0$ is the cell speed of sound, $l_{r-1,r}$ and $\mathbf{n}_{r-1,r}$ are respectively the length and the normal of the edge which bridges the node $r-1$ to its neighbor r in the cell j.

3 Extension to Finite Plasticity

The hyperelastic scheme, using the appropriate equation of state, allows to observe some plastic effects (those implied by constraining the stress), such as the split shock in flyer-plate experiment, the chronometry of an imploding plastic shell or the characteristic shape of a cylinder impacted on a wall. This equation of state is [7]

$$\epsilon = H(\tau, S) + \tau\psi(\mathbf{F}) \tag{18}$$

where H is an hydrodynamic equation of state, and ψ is

$$\psi = \begin{cases} \frac{\mu}{4} R & \text{if } R \leq R^*, \\ \frac{Y}{\sqrt{6}}(\sqrt{R} - \sqrt{R^*}) + \frac{\mu}{4} R^* & \text{if } R^* \leq R, \end{cases}$$

$$R = \iota_1 \iota_3^{-1/3} + \iota_2 \iota_3^{-2/3} - 6,$$

$$R^* = \frac{2}{3} \frac{Y^2}{\mu^2}$$

Y and μ being respectively the yield limit and the shear modulus. ι_1, ι_2 and ι_3 are the trace, the trace of the cofactor and the determinant, of the right Cauchy–Green tensor $\mathbf{B} = \mathbf{F}\mathbf{F}^t$.

4 Numerical Results

4.1 Hugoniot Experiments

These experiments are described in [4]. An analytic solution is given in [7] to which we compare our results. Two plates of steel are impacted symmetrically. We see Fig. 2 our numerical results with a 1D scheme. The 2D scheme converges to the same result.

4.2 Imploding Shell

This test case is taken from [5]. An elastic-plastic shell implodes, converting cinematic energy into internal one, until the shell stops at a given radius (we stop the simulation when 99.9% of the total initial energy is converted). See Fig. 3.

4.3 Taylor Test Case

The Taylor testcase is the impact of a cylinder on a rigid wall [10]. The final shape of the cylinder characterizes the solid material [12]. As our scheme is planar, results

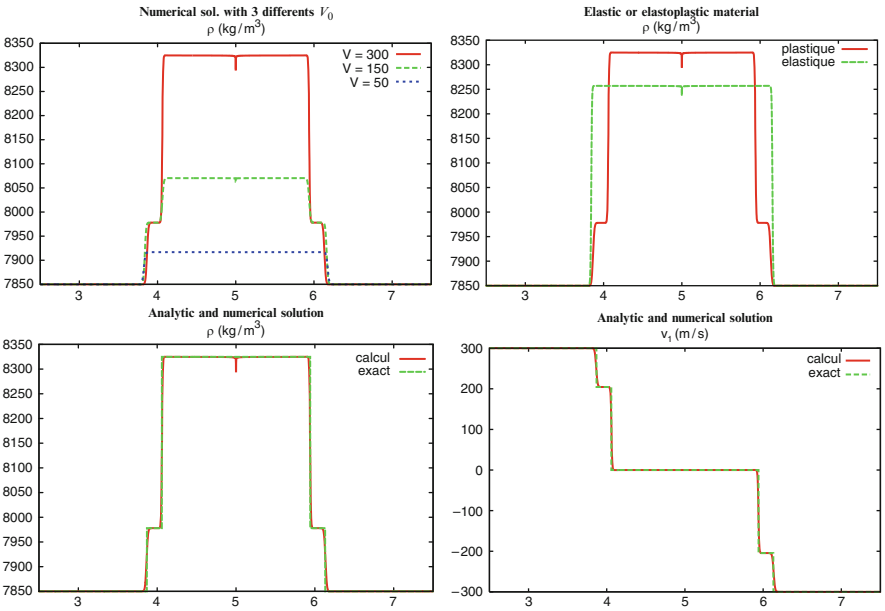
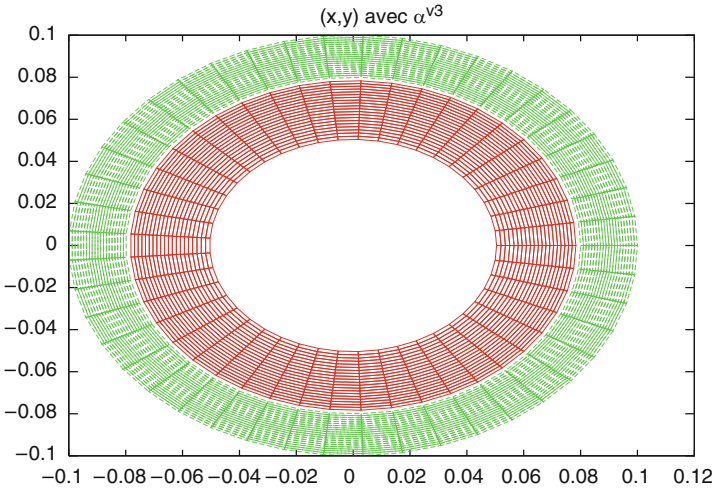


Fig. 2 On the *top*, 1D numerical results for three different impact velocities (on the *left*), for an elastic material by taking Y very high (on the *right*). On the *bottom*, the differences between analytic and numerical solutions



V_0	R_f^-	[5]	Diff. (%)	R_f^+	[5]	Diff. (%)	t_f
417.1	50.40	50	+0.008	78.33	78.1	+0.003	122.25
454.7	45.71	45	+0.015	75.38	75	+0.005	126.69
490.2	41.17	40	+0.029	72.14	75.12	−0.040	129.49

Fig. 3 The final mesh is inside the initial one. The table compares values from [5] (obtained by analytic solution of an incompressible model [11]) to our results, for three different initial radial velocities of norm V_0. R_f^- and R_f^+ are the final intern and extern radius

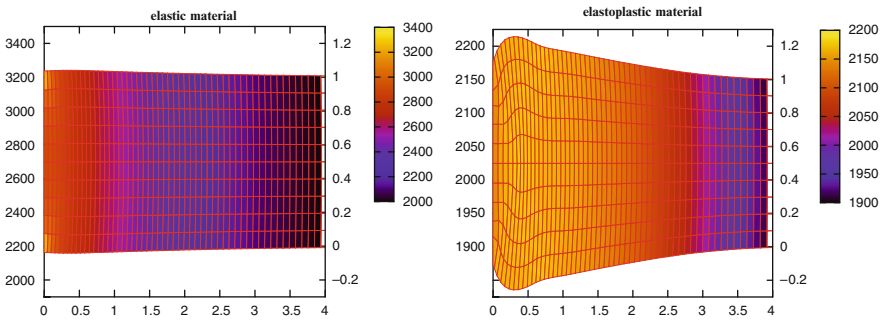


Fig. 4 An elastic material (with Y very high) on the *left*: the metal acts like a spring. A plastic material on the *right*

are difficult to interpret. Nevertheless, we observe qualitatively the consequences of plasticity. See Fig. 4.

References

1. Carre, G., Delpino, S., Després, B., Labourasse, E.: A cell-centered Lagrangian hydrodynamics scheme on general unstructured meshes in arbitrary dimension. J. Comput. Phys. **228**, 5160–5183 (2009)
2. Ciarlet, P.: Mathematical elasticity, volume I: three-dimensional elasticity. North-Holland, Amsterdam (1988)
3. Després, B., Mazeran, C.: Lagrangian gas dynamics in 2D and lagrangian systems. Comptes Rendus Académie des Sciences (Paris). **331**, 327–372 (2003)
4. Drumheller D.S.: Introduction to wave propagation in nonlinear fluids and solids. Cambridge University Press, Cambridge (1998)
5. Howell, B.P., Ball, G.J.: A free-lagrange augmented Godunov method for the simulation of elastic-plastic solids. J. Comput. Phys. **175**, 128–167 (2002)
6. Kluth, G.: Analyse mathématique et numérique de systèmes hyperélastiques, et introduction de la plasticité. PHD of Université Pierre et Marie Curie, Paris VI (2008)
7. Kluth, G., Després, B.: Perfect plasticity and hyperelastic models for isotropic materials. Cont. Mech. Thermo. **20**, 173–192 (2008)
8. Maire, P.H.: A high-order cell centered lagrangian scheme for two-dimensional compressible fluid flows on unstructured meshes. J. Comput. Phys. **228**(7), 2391–2425 (2009)
9. Maire, P.H., Abgrall, R., Breil, J., Ovadia, J.: A cell-centered lagrangian scheme for 2D compressible flow problems. Siam J. Sci. Comp. **29** (2007)
10. Taylor, G.I.: The use of flat-ended projectiles for determining dynamic yield stress I. Theoretical considerations. Proc. Roy. Soc. Lond. A **194**, 289 (1948)
11. Verney, D.: Evaluation de la limite élastique du cuivre et de l'uranium par des expériences d'implosion 'lente'. Behavior of Dense Media Under High Dynamic Pressures, Symposium H.D.P., Paris (1968)
12. Wilkins M.L., Guinan M.W.: Impact of cylinders on a rigid boundary. J. Appl. Phys. **44**, 1200 (1973)

Local Projection Method for Convection-Diffusion-Reaction Problems with Projection Spaces Defined on Overlapping Sets

Petr Knobloch

Abstract We extend the local projection finite element method for steady scalar convection-diffusion-reaction equations to local projection spaces defined on overlapping sets. This enables to define the local projection method without the need of a mesh refinement or an enrichment of the finite element space. For the streamline derivative based stabilization, we introduce a modification that leads to an optimal estimate of the consistency error even if the stabilization parameters scale correctly with respect to convection, diffusion and mesh width. The main result of the paper is an optimal error estimate with respect to the standard local projection norm.

1 Introduction

In this paper we deal with the application of the local projection finite element method to the numerical solution of the convection-diffusion-reaction problem

$$-\varepsilon\,\Delta u+\mathbf{b}\cdot\nabla u+c\,u=f\quad\text{in }\Omega,\qquad u=u_b\quad\text{on }\Gamma^D,\qquad \varepsilon\,\frac{\partial u}{\partial\mathbf{n}}=g\quad\text{on }\Gamma^N, \tag{1}$$

where $\Omega\subset\mathbb{R}^d$, $d=2,3$, is a bounded domain with a polyhedral Lipschitz-continuous boundary $\partial\Omega$ and $\Gamma^D,\Gamma^N\subset\partial\Omega$ are two relatively open disjoint sets satisfying $\overline{\Gamma^D\cup\Gamma^N}=\partial\Omega$ and $\mathrm{meas}_{d-1}(\Gamma^D)>0$. We denote by $\mathbf{n}$ the outer unit normal vector to $\partial\Omega$. We assume that ε is a positive constant and $\mathbf{b}\in W^{1,\infty}(\Omega)^d$, $c\in L^\infty(\Omega)$, $f\in L^2(\Omega)$, $u_b\in H^{1/2}(\Gamma^D)$ and $g\in H^{-1/2}(\Gamma^N)$ are given functions satisfying

$$\sigma:=c-\tfrac{1}{2}\,\mathrm{div}\,\mathbf{b}\ge\sigma_0>0\,,$$

P. Knobloch
Faculty of Mathematics and Physics, Department of Numerical Mathematics, Charles University, Sokolovská 83, 186 75 Praha 8, Czech Republic
e-mail: knobloch@karlin.mff.cuni.cz

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_53, © Springer-Verlag Berlin Heidelberg 2010

where σ_0 is a constant. Moreover, we assume that the inflow boundary is a part of the Dirichlet boundary, i.e.,

$$\{\mathbf{x} \in \partial\Omega;\ (\mathbf{b} \cdot \mathbf{n})(\mathbf{x}) < 0\} \subset \Gamma^D.$$

The standard weak formulation of (1) reads: Find $u \in H^1(\Omega)$ such that $u = u_b$ on Γ^D and

$$a(u, v) = (f, v) + \langle g, v\rangle_{\Gamma^N} \qquad \forall\ v \in V := \{v \in H^1(\Omega);\ v = 0 \ \text{ on } \ \Gamma^D\}, \tag{2}$$

where

$$a(u, v) := \varepsilon\,(\nabla u, \nabla v) + (\mathbf{b} \cdot \nabla u, v) + (c\,u, v),$$

$(\cdot, \cdot)$ denotes the inner product in $L^2(\Omega)$ or $L^2(\Omega)^d$ and $\langle \cdot, \cdot \rangle_{\Gamma^N}$ is the duality pairing between $H^{-1/2}(\Gamma^N)$ and $H^{1/2}(\Gamma^N)$. Since $a(v, v) \geq \varepsilon |v|^2_{1,\Omega}$ for any $v \in V$, the weak formulation has a unique solution.

The local projection finite element method was introduced by Becker and Braack [1] for the Stokes problem and later it was applied to many other problems including transport problems, convection-diffusion-reaction equations, Oseen equations and Navier–Stokes equations, see, e.g., [2, 7–9]. A drawback of the local projection method is that it requires (significantly) more degrees of freedom than, e.g., residual-based methods [10]. Indeed, for a residual-based stabilization, we can simply use a finite element space consisting of piecewise polynomials of some degree l on a given triangulation whereas, for the classical local projection approaches, the finite element space is either defined on a refined triangulation or enriched by additional bubble functions. In this paper we shall show that the increase of the number of degrees of freedom in not necessary if we allow local projection spaces defined on overlapping subsets of Ω.

The plan of the paper is as follows. In Sect. 2, we introduce the finite element spaces needed for defining the discrete problem and formulate all assumptions used in this paper. Then, in Sect. 3, we define the local projection discretization of (1). For the streamline derivative based stabilization, we introduce a simple modification that leads to an optimal estimate of the consistency error, not available before. Finally, in Sect. 4, we carry out our error analysis leading to an optimal error estimate. Throughout the paper we use standard notation which can be found, e.g., in [3]. Moreover, we use the notation $a \lesssim b$ if $a \leq Cb$ with $C > 0$ independent of the mesh parameter h and the data of (1). We write $a \sim b$ if $a \lesssim b$ and $b \lesssim a$.

2 Assumptions on Finite Element Spaces

Given $h > 0$, let $W_h \subset H^1(\Omega)$ be a finite element space approximating the space $H^1(\Omega)$ and set $V_h = W_h \cap V$. Furthermore, let $\mathscr{M}_h$ be a set consisting of a finite number of open subsets M of Ω such that $\overline{\Omega} = \cup_{M \in \mathscr{M}_h} \overline{M}$. We assume that, for

any $M \in \mathscr{M}_h$,

$$\operatorname{card}\{M' \in \mathscr{M}_h\,;\ M \cap M' \neq \emptyset\} \lesssim 1\,, \tag{3}$$

$$h_M := \operatorname{diam}(M) \lesssim h\,, \tag{4}$$

$$h_M \lesssim h_{M'} \qquad \forall\ M' \in \mathscr{M}_h,\ M \cap M' \neq \emptyset\,. \tag{5}$$

Moreover, we assume that, for any $M \in \mathscr{M}_h$, there is a nontrivial space $B_M \subset (W_h|_M) \cap H_0^1(M)$ such that $B_M \subset W_h$ if the functions from B_M are extended by zero outside M. For any $M \in \mathscr{M}_h$, we introduce a finite-dimensional space $D_M \subset L^2(M)$ and we assume that

$$\sup_{v \in B_M} \frac{(v,q)_M}{\|v\|_{0,M}} \gtrsim \|q\|_{0,M} \qquad \forall\ q \in D_M,\ M \in \mathscr{M}_h\,, \tag{6}$$

where $(\cdot,\cdot)_M$ is the inner product in $L^2(M)$. We shall also need the inverse inequality

$$|v_h|_{1,M} \lesssim h_M^{-1}\, \|v_h\|_{0,M} \qquad \forall\ v_h \in W_h,\ M \in \mathscr{M}_h\,. \tag{7}$$

In addition, we assume that there exist interpolation operators $i_h \in \mathscr{L}(H^2(\Omega), W_h) \cap \mathscr{L}(H^2(\Omega) \cap V, V_h)$ and $j_M \in \mathscr{L}(H^1(M), D_M)$, $M \in \mathscr{M}_h$, such that, for some constant $l \in \mathbb{N}$ and any $k \in \{1, \ldots, l\}$, we have

$$\|v - i_h v\|_{1,h} + h^{-1/2}\, \|v - i_h v\|_{0,\Gamma^N} \lesssim h^k\, |v|_{k+1,\Omega} \qquad \forall\ v \in H^{k+1}(\Omega)\,, \tag{8}$$

$$\|q - j_M q\|_{0,M} \lesssim h_M^k\, |q|_{k,M} \qquad \forall\ q \in H^k(M),\ M \in \mathscr{M}_h\,. \tag{9}$$

In (8), we use the mesh dependent norm

$$\|v\|_{1,h} = \left(\sum_{M \in \mathscr{M}_h} \{|v|_{1,M}^2 + h_M^{-2}\, \|v\|_{0,M}^2\} \right)^{1/2}.$$

Remark 1. Let $\mathscr{T}_h$ be a triangulation of Ω consisting of shape-regular cells (simplices, quadrilaterals or hexahedra) possessing the usual compatibility properties. Then the above assumptions can be satisfied if the sets M are defined as unions of cells of $\mathscr{T}_h$ sharing a common interior vertex, the space W_h consists of (mapped) piecewise polynomial functions of degree l on $\mathscr{T}_h$ and $D_M = P_{l-1}(M)$. Let us emphasize that for classical local projection methods, which do not allow overlapping sets M, a standard finite element space on $\mathscr{T}_h$ generally cannot be used as W_h (cf., e.g., [7]). In these approaches either $\mathscr{T}_h$ has to be refined or the finite element space has to be enriched by additional bubble functions.

3 Discrete Problem

Consider any $M \in \mathscr{M}_h$. We denote by π_M a continuous linear projection operator which maps the space $L^2(M)$ onto the space D_M such that $\|\pi_M\|_{\mathscr{L}(L^2(M),L^2(M))} \lesssim 1$. Using π_M, we introduce the so-called fluctuation operator $\kappa_M = id - \pi_M$, where id is the identity operator on $L^2(M)$. Then

$$\|\kappa_M\|_{\mathscr{L}(L^2(M),L^2(M))} \lesssim 1 . \tag{10}$$

An application of κ_M to a vector valued function means that κ_M is applied componentwise. In addition, we choose a constant $\mathbf{b}_M \in \mathbb{R}^d$ such that

$$|\mathbf{b}_M| \le \|\mathbf{b}\|_{0,\infty,M} , \qquad \|\mathbf{b} - \mathbf{b}_M\|_{0,\infty,M} \lesssim h_M \, |\mathbf{b}|_{1,\infty,M} . \tag{11}$$

Finally, we define the local projection stabilization term

$$s_M(u,v) = (\kappa_M(\mathbf{b}_M \cdot \nabla u), \kappa_M(\mathbf{b}_M \cdot \nabla v))_M \tag{12}$$

or

$$s_M(u,v) = (\kappa_M \nabla u, \kappa_M \nabla v)_M . \tag{13}$$

Now we set

$$s_h(u,v) = \sum_{M \in \mathscr{M}_h} \tau_M \, s_M(u,v) ,$$

where τ_M is a nonnegative stabilization parameter. It was shown in [5] that τ_M should satisfy

$$\tau_M \sim \min \left\{ \frac{h_M}{\|\mathbf{b}\|_{0,\infty,M}}, \frac{h_M^2}{\varepsilon} \right\} \frac{\|\mathbf{b}\|_{0,\infty,M}^2}{\gamma_M(\mathbf{b})} , \tag{14}$$

where

$$\begin{aligned} &\gamma_M(\mathbf{b}) = \|\mathbf{b}\|_{0,\infty,M}^2 && \text{if } s_M \text{ are given by (12),} \\ &\gamma_M(\mathbf{b}) = 1 && \text{if } s_M \text{ are given by (13).} \end{aligned}$$

To formulate a discrete Dirichlet boundary condition, we introduce a function $\tilde{u}_{bh} \in W_h$ such that its trace on Γ^D approximates the boundary condition u_b.

The local projection discretization of (1) is based on the weak formulation (2) and reads: Find $u_h \in W_h$ such that $u_h - \tilde{u}_{bh} \in V_h$ and

$$a_h(u_h, v_h) = (f, v_h) + \langle g, v_h \rangle_{\Gamma^N} \qquad \forall \; v_h \in V_h , \tag{15}$$

where $a_h(u,v) = a(u,v) + s_h(u,v)$.

It is natural to investigate the error of u_h with respect to the local projection norm

$$|||v|||_{LP} = \left(\varepsilon \, |v|_{1,\Omega}^2 + \|\sigma^{1/2} \, v\|_{0,\Omega}^2 + \frac{1}{2} \, \|(\mathbf{b} \cdot \mathbf{n})^{1/2} \, v\|_{0,\Gamma^N}^2 + s_h(v,v) \right)^{1/2}$$

since $a_h(v,v) = |||v|||_{LP}^2$ for any $v \in V$. This property shows that the discrete problem is uniquely solvable. Note also that, for any $v \in H^1(\Omega)$,

$$|||v|||_{LP} \lesssim \|(\mathbf{b}\cdot\mathbf{n})^{1/2} \, v\|_{0,\Gamma^N} + \left(\varepsilon + h \, \|\mathbf{b}\|_{0,\infty,\Omega} + h^2 \, \|\sigma\|_{0,\infty,\Omega}\right)^{1/2} \|v\|_{1,h} \,. \quad (16)$$

Remark 2. A standard choice is to use $\mathbf{b}$ instead of $\mathbf{b}_M$ in (12). However, it was demonstrated in [5] that then it is generally not possible to obtain optimal convergence results if (14) holds. We shall see in the next section that the use of $\mathbf{b}_M$ leads to an optimal error estimate.

4 Error Analysis

Let $u \in H^1(\Omega)$ be the weak solution of (1). The local projection discretization is not consistent and we have $a_h(u - u_h, v_h) = s_h(u, v_h)$ for any $v_h \in V_h$. Denoting

$$W_h^b = \{w_h \in W_h \,;\; w_h - \tilde{u}_{bh} \in V_h\} \,,$$

we obtain similarly as in the proof of the first Strang lemma (see, e.g., [3]) that

$$|||u - u_h|||_{LP} \le \inf_{w_h \in W_h^b} \left\{ |||u - w_h|||_{LP} + \sup_{v_h \in V_h} \frac{a_h(u - w_h, v_h)}{|||v_h|||_{LP}} \right\} + \sup_{v_h \in V_h} \frac{s_h(u, v_h)}{|||v_h|||_{LP}} \,. \quad (17)$$

In the classical local projection method, where the sets $M \in \mathscr{M}_h$ are non-overlapping, the infimum on the right-hand side of (17) is estimated by choosing w_h as a special interpolate of u for which the interpolation error is L^2-orthogonal to the spaces D_M. If the sets M overlap, such an interpolate cannot be constructed. Instead we shall employ the following lemma.

Lemma 1. *There is an operator $\varrho_h : L^2(\Omega) \to W_h \cap H_0^1(\Omega)$ such that, for any $u, v \in L^2(\Omega)$, we have*

$$|(u, v - \varrho_h v)| \lesssim \sum_{M \in \mathscr{M}_h} \|\kappa_M \, u\|_{0,M} \, \|v\|_{0,M} \,, \quad (18)$$

$$\|\varrho_h v\|_{1,h} \lesssim \left(\sum_{M \in \mathscr{M}_h} h_M^{-2} \, \|v\|_{0,M}^2 \right)^{1/2} . \tag{19}$$

Proof. First, let us write $\mathscr{M}_h$ in the form $\mathscr{M}_h = \{M_i\}_{i=1}^N$ and set

$$\widetilde{M}_1 = M_1 , \qquad \widetilde{M}_i = M_i \setminus \bigcup_{k=1}^{i-1} \widetilde{M}_k , \quad i = 2, \ldots, N .$$

Then

$$\widetilde{M}_i \cap \widetilde{M}_j = \emptyset \quad \forall \; i \neq j , \qquad \bigcup_{i=1}^N \widetilde{M}_i = \bigcup_{i=1}^N M_i .$$

To simplify the notation, we now again drop the indices and, for any $M \in \mathscr{M}_h$, we denote by $\widetilde{M}$ the subset of M constructed in the above way.

Consider any $v \in L^2(\Omega)$. The inf–sup conditions (6) imply that, for any $M \in \mathscr{M}_h$, there exists $z_M \in B_M$ such that (cf., e.g., [4])

$$(q, z_M)_M = (q, v)_{\widetilde{M}} \qquad \forall \; q \in D_M , \tag{20}$$

$$\|z_M\|_{0,M} \lesssim \|v\|_{0,\widetilde{M}} . \tag{21}$$

We set $\varrho_h v = \sum_{M \in \mathscr{M}_h} z_M$ (with $z_M = 0$ in $\Omega \setminus M$). Then $\varrho_h v \in W_h \cap H_0^1(\Omega)$ and, for any $u \in L^2(\Omega)$, we obtain

$$(u, v - \varrho_h v) = \sum_{M \in \mathscr{M}_h} \left[(u, v)_{\widetilde{M}} - (u, z_M)_M \right] = \sum_{M \in \mathscr{M}_h} \left[(\kappa_M u, v)_{\widetilde{M}} - (\kappa_M u, z_M)_M \right] ,$$

which gives (18) by applying the Cauchy–Schwarz inequality and (21). Using (3) and (5), we derive that

$$\sum_{M \in \mathscr{M}_h} h_M^{-2} \, \|\varrho_h v\|_{0,M}^2 \lesssim \sum_{\substack{M, M' \in \mathscr{M}_h, \\ M \cap M' \neq \emptyset}} h_M^{-2} \, \|z_{M'}\|_{0,M}^2 \lesssim \sum_{M' \in \mathscr{M}_h} h_{M'}^{-2} \, \|z_{M'}\|_{0,M'}^2 .$$

Analogously, using (3) and (7), we get

$$\sum_{M \in \mathscr{M}_h} |\varrho_h v|_{1,M}^2 \lesssim \sum_{\substack{M, M' \in \mathscr{M}_h, \\ M \cap M' \neq \emptyset}} |z_{M'}|_{1,M}^2 \lesssim \sum_{M' \in \mathscr{M}_h} h_{M'}^{-2} \, \|z_{M'}\|_{0,M'}^2 ,$$

which implies (19) in view of (21). □

The following lemma enables us to obtain an optimal estimate of the infimum on the right-hand side of (17).

Lemma 2. *For any $w \in H^1(\Omega)$ and any $v_h \in V_h \setminus \{0\}$, we have*

$$|||w - \varrho_h w|||_{LP} + \frac{a_h(w - \varrho_h w, v_h)}{|||v_h|||_{LP}} \lesssim \|(\mathbf{b}\cdot\mathbf{n})^{1/2} w\|_{0,\Gamma^N} + \left(\varepsilon + h\,\|\mathbf{b}\|_{0,\infty,\Omega} + h^2\,\|\sigma\|_{0,\infty,\Omega} + h^2\,|\mathbf{b}|^2_{1,\infty,\Omega}\,\sigma_0^{-1}\right)^{1/2} \|w\|_{1,h}\,.$$

Proof. Consider any $w \in H^1(\Omega)$ and $v_h \in V_h$ and set $z = w - \varrho_h w$. Then

$$(\mathbf{b}\cdot\nabla z, v_h) + (c\,z, v_h) = -(z, \mathbf{b}\cdot\nabla v_h) + (\sigma\,z, v_h) - \tfrac{1}{2}((\operatorname{div}\mathbf{b})\,z, v_h) + \langle(\mathbf{b}\cdot\mathbf{n})\,z, v_h\rangle_{\Gamma^N}\,.$$

Therefore, applying the Cauchy–Schwarz inequality, we obtain

$$a_h(z, v_h) \lesssim |||z|||_{LP}\,|||v_h|||_{LP} + \sigma_0^{-1/2}\,\|(\operatorname{div}\mathbf{b})\,z\|_{0,\Omega}\,|||v_h|||_{LP} - (z, \mathbf{b}\cdot\nabla v_h)\,.$$

According to (18), we have

$$|(z, \mathbf{b}\cdot\nabla v_h)| \lesssim \sum_{M\in\mathscr{M}_h} \|\kappa_M(\mathbf{b}\cdot\nabla v_h)\|_{0,M}\,\|w\|_{0,M}\,.$$

Applying (7), (10) and (11), we derive

$$\begin{aligned}\|\kappa_M(\mathbf{b}\cdot\nabla v_h)\|_{0,M} &\le \|\kappa_M((\mathbf{b}-\mathbf{b}_M)\cdot\nabla v_h)\|_{0,M} + \|\kappa_M(\mathbf{b}_M\cdot\nabla v_h)\|_{0,M}\\ &\lesssim |\mathbf{b}|_{1,\infty,M}\,\|v_h\|_{0,M} + \left(\|\mathbf{b}\|^2_{0,\infty,M}\,\gamma_M(\mathbf{b})^{-1}\,s_M(v_h, v_h)\right)^{1/2}.\end{aligned}$$

If $\varepsilon < \|\mathbf{b}\|_{0,\infty,M}\,h_M$, we have

$$\|\kappa_M(\mathbf{b}\cdot\nabla v_h)\|_{0,M} \lesssim |\mathbf{b}|_{1,\infty,M}\|v_h\|_{0,M} + (\|\mathbf{b}\|_{0,\infty,M} h_M\,\tau_M\,s_M(v_h, v_h))^{1/2}\,h_M^{-1}.$$

If $\varepsilon \ge \|\mathbf{b}\|_{0,\infty,M}\,h_M$, we obtain $\|\kappa_M(\mathbf{b}\cdot\nabla v_h)\|_{0,M} \le \varepsilon\,|v_h|_{1,M}\,h_M^{-1}$. Thus, in both cases, we arrive at the estimate

$$|(z, \mathbf{b}\cdot\nabla v_h)| \lesssim \left(\varepsilon + h\,\|\mathbf{b}\|_{0,\infty,\Omega} + h^2\,|\mathbf{b}|^2_{1,\infty,\Omega}\,\sigma_0^{-1}\right)^{1/2}\|w\|_{1,h}\,|||v_h|||_{LP}\,.$$

Now, to complete the proof, it suffices to apply (16) and (19). □

It remains to estimate the consistency error.

Lemma 3. *Let $u \in H^{k+1}(\Omega)$ for some $k \in \{1, \dots, l\}$. Then*

$$\sup_{v\in H^1(\Omega)} \frac{s_h(u, v)}{|||v|||_{LP}} \lesssim \|\mathbf{b}\|^{1/2}_{0,\infty,\Omega}\,h^{k+1/2}\,|u|_{k+1,\Omega}\,.$$

Proof. For any $u \in H^{k+1}(\Omega)$ and $v \in H^1(\Omega)$, we have

$$s_h(u,v) \le \sqrt{s_h(u,u)}\,\sqrt{s_h(v,v)} \le \sqrt{s_h(u,u)}\,|||v|||_{LP}$$

and hence it suffices to estimate $\tau_M\, s_M(u,u)$ with an arbitrary $M \in \mathscr{M}_h$. In view of (11), we obtain $\tau_M\, s_M(u,u) \lesssim h_M\, \|\mathbf{b}\|_{0,\infty,M}\, \|\kappa_M \nabla u\|_{0,M}^2$ for s_M defined by both (12) and (13). Since κ_M vanishes on D_M, the assumption (10) and the approximation property (9) imply that $\|\kappa_M \nabla u\|_{0,M} \lesssim \|\nabla u - j_M \nabla u\|_{0,M} \lesssim h_M^k\, |u|_{k+1,M}$ and the lemma follows using (3). □

Now we are in a position to prove the main result of this paper.

Theorem 1. *Let the weak solution of (1) satisfy $u \in H^{k+1}(\Omega)$ for some $k \in \{1,\dots,l\}$, let $\tilde{u}_b \in H^2(\Omega)$ be an extension of u_b and let $\tilde{u}_{bh} = i_h \tilde{u}_b$. Then the solution u_h of the local projection discretization (15) satisfies the error estimate*

$$|||u - u_h|||_{LP} \lesssim (\varepsilon + h\|\mathbf{b}\|_{0,\infty,\Omega} + h^2\|\sigma\|_{0,\infty,\Omega} + h^2|\mathbf{b}|_{1,\infty,\Omega}^2\sigma_0^{-1})^{1/2} h^k |u|_{k+1,\Omega}. \tag{22}$$

Proof. The theorem follows by setting $w_h = i_h u + \varrho_h(u - i_h u)$ in (17), employing Lemmas 2 and 3 and applying (8). □

Remark 3. Estimates of the type (22) can be proved for various stabilized finite element methods applied to the problem (1) and are known to be optimal, see, e.g., [10]. If we define the stabilization term $s_M(u,v)$ in (12) using $\mathbf{b}$ instead of $\mathbf{b}_M$, then Lemma 2 still holds but the consistency error cannot be estimated as in Lemma 3. Assuming $\mathbf{b} \cdot \nabla u \in H^k(\Omega)$ with $k \in \{1,\dots,l\}$, we obtain

$$\sup_{v_h \in V_h} \frac{s_h(u,v_h)}{|||v_h|||_{LP}} \lesssim h^k \left(\sum_{M \in \mathscr{M}_h} \min \left\{ \frac{|\mathbf{b} \cdot \nabla u|_{k,M}^2}{\sigma_0}, \frac{h_M\, |\mathbf{b} \cdot \nabla u|_{k,M}^2}{\|\mathbf{b}\|_{0,\infty,M}} \right\} \right)^{1/2},$$

see [5,6]. Thus, if $\mathbf{b} \neq \mathbf{0}$ in $\overline{\Omega}$, the optimal convergence order can be still proved but generally we only have the suboptimal convergence order k. Moreover, for small σ_0, the accuracy of the discrete solution may be significantly worse than for s_h defined using $\mathbf{b}_M$.

Acknowledgements This work 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. Becker, R., Braack, M.: A finite element pressure gradient stabilization for the Stokes equations based on local projections. Calcolo **38**, 173–199 (2001)

2. Braack, M., Burman, E.: Local projection stabilization for the Oseen problem and its interpretation as a variational multiscale method. SIAM J. Numer. Anal. **43**, 2544–2566 (2006)
3. Ciarlet, P.G.: Basic error estimates for elliptic problems. In: Ciarlet, P.G., Lions, J.L. (eds.) Handbook of Numerical Analysis, vol. 2 – Finite Element Methods (pt. 1), pp. 17–351. North-Holland, Amsterdam (1991)
4. Ern, A., Guermond, J.-L.: Theory and Practice of Finite Elements. Springer, New York (2004)
5. Knobloch, P.: On the application of local projection methods to convection-diffusion-reaction problems. In: Hegarty, A.F., Kopteva, N., O'Riordan, E., Stynes, M. (eds.) BAIL 2008 – Boundary and Interior Layers, Lect. Notes Comput. Sci. Eng., vol. 69, pp. 183–194. Springer, Berlin (2009)
6. Knobloch, P., Lube, G.: Local projection stabilization for advection–diffusion-reaction problems: One-level vs. two-level approach. Appl. Numer. Math. **59**, 2891–2907 (2009)
7. Matthies, G., Skrzypacz, P., Tobiska, L.: A unified convergence analysis for local projection stabilisations applied to the Oseen problem. M2AN **41**, 713–742 (2007)
8. Matthies, G., Skrzypacz, P., Tobiska, L.: Stabilization of local projection type applied to convection-diffusion problems with mixed boundary conditions. Electron. Trans. Numer. Anal. **32**, 90–105 (2008)
9. Rapin, G., Lube, G., Löwe, J.: Applying local projection stabilization to inf-sup stable elements. In: Kunisch, K., Of, G., Steinbach, O. (eds.) Numerical Mathematics and Advanced Applications, pp. 521–528. Springer, Berlin (2008)
10. Roos, H.-G., Stynes, M., Tobiska, L.: Robust Numerical Methods for Singularly Perturbed Differential Equations. Convection-Diffusion-Reaction and Flow Problems, 2nd edn. Springer, Berlin (2008)

Numerical Solution of Volterra Integral Equations with Weak Singularities

M. Kolk and A. Pedas

Abstract We propose a piecewise polynomial collocation method for solving linear Volterra integral equations of the second kind with kernels which, in addition to a weak 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 $C^k(\Omega)$ be the set of all k times continuously differentiable functions on Ω, $C^0(\Omega) = C(\Omega)$. By $C[a,b]$ we denote the Banach space of continuous functions f on $[a,b]$ with the usual norm $\|f\| = \max\{|f(x)| : x \in [a,b]\}$. Let

$$D_b = \{(x,y) : \ 0 < y < x \le b\}, \quad \overline{D}_b = \{(x,y) : 0 \le y \le x \le b\}.$$

In many practical applications there arise integral equations of the form

$$u(x) = \int_0^x K(x,y)u(y)dy + f(x), \quad 0 \le x \le b, \tag{1}$$

with $f \in C^m[0,b]$, $K(x,y) = g(x,y)(x-y)^{-\nu}$, $0 < \nu < 1$, $g \in C^m(\overline{D}_b)$, $m \in \mathbf{N} = \{1,2,\ldots\}$. The solution $u(x)$ to (1) is typically non-smooth at $x = 0$ where its derivatives become unbounded (see for example, [3–5]). In collocation methods the singular behaviour of the solution $u(x)$ can be taken into account by using polynomial splines on special graded grids $\Delta_N^r = \{x_0, \ldots, x_N : 0 = x_0 < \ldots < x_N = b\}$ with the nodes

$$x_i = b(i/N)^r, \quad i = 0, \ldots, N, \quad N \in \mathbf{N}, \quad r \in \mathbf{R}, \quad r \ge 1. \tag{2}$$

M. Kolk (✉) and A. Pedas
Institute of Mathematics, University of Tartu, Estonia
e-mail: marek.kolk@ut.ee, arvet.pedas@ut.ee

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_54, © Springer-Verlag Berlin Heidelberg 2010

The parameter r characterizes the degree of non-uniformity of the grid Δ_N^r: if $r > 1$, then the nodes $x_0, \ldots, x_N$ of the grid Δ_N^r are more densely clustered near the left endpoint of the interval $[0, b]$ where $u(x)$ may be singular. High order methods use large values of r, see [3–5]. However, in practice, the use of strongly graded grids Δ_N^r 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 behavior 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 [10] for details (see also [2, 7]). Note that in [9, 12] similar ideas for solving Fredholm integral equations have been used (see also [6, 14]).

In the present paper we extend the domain of applicability of this approach. To this aim, we examine a more complicated situation for (1) where the kernel $K(x, y)$, in addition to a diagonal singularity (a singularity as $y \to x$), may have a boundary singularity (a singularity as $y \to 0$). Actually we assume that the kernel $K(x, y)$ has the form

$$K(x, y) = g(x, y)(x - y)^{-\nu} y^{-\lambda}, \quad (x, y) \in D_b, \quad 0 < \nu < 1, \quad 0 \le \lambda < 1, \tag{3}$$

where $g \in C^m(\overline{D}_b)$, $m \in \{0\} \cup \mathbf{N}$. The set of kernels satisfying (3) will be denoted by $W^{m,\nu,\lambda}(D_b)$.

Throughout the paper c denotes a positive constant which may have different values by different occurrences.

2 Regularity of the Solution

For given $m \in \mathbf{N}$ and $0 < \theta < 1$ let $C^{m,\theta}(0, b]$ be the set of functions $u \in C[0, b] \cap C^m(0, b]$ such that

$$|u^{(j)}(x)| \le c x^{1-\theta-j}, \quad 0 < x \le b, \quad j = 1, \ldots, m. \tag{4}$$

It follows from [11] that the regularity of the solution to (1) can be characterized by the following result.

Lemma 1. *Assume that $K \in W^{m,\nu,\lambda}(D_b)$, $f \in C^{m,\nu+\lambda}(0, b]$ where $m \in \mathbf{N}$, $0 < \nu < 1$, $0 \le \lambda < 1$, $\nu + \lambda < 1$. Then (1) has a unique solution $u \in C^{m,\nu+\lambda}(0, b]$.*

3 Smoothing Transformation

For given $\varrho \in [1, \infty)$ denote

$$\varphi(s) = b^{1-\varrho} s^{\varrho}, \quad 0 \le s \le b. \tag{5}$$

Clearly, $\varphi \in C[0,b]$, $\varphi(0) = 0$, $\varphi(b) = b$ 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 $\varphi(s) \equiv s$ for $\varrho = 1$. We are interested in a transformation (5) with $\varrho > 1$ since it possesses a smoothing property for $u(\varphi(s))$ with respect to the singularities of $u'(x), \ldots, u^{(m)}(x)$ at $x = 0$ (see Lemma 2).

Lemma 2. *[10]. Let $u \in C^{m,\theta}(0,b]$, $m \in \mathbf{N}$, $0 < \theta < 1$, and let φ be defined by formula (5). Furthermore, let $u_\varphi(s) = u(\varphi(s)), 0 \le s \le b$. Then $u_\varphi \in C[0,b] \cap C^m(0,b]$ and*

$$|u_\varphi^{(j)}(s)| \le c s^{\varrho(1-\theta)-j}, \quad 0 < s \le b, \quad j = 1, \ldots, m. \tag{6}$$

4 Numerical Method

Using (5) 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.

For given integers $m, N \in \mathbf{N}$ let

$$S_{m-1}^{(-1)}(\Delta_N^r) = \left\{ v_N : v_N\big|_{[x_{j-1}, x_j]} \in \pi_{m-1},\ j = 1, \ldots, N \right\}$$

be the underlying spline spaces of piecewise polynomial functions on the grid Δ_N^r with the nodes (2). Here $v_N\big|_{[x_{j-1},x_j]}$ $(j = 1, \ldots, N)$ is the restriction of $v_N(t)$, $t \in [0,b]$, to the subinterval $[x_{j-1}, x_j] \subset [0,b]$ 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 $x_1, \ldots, x_{N-1}$ of the grid Δ_N^r. In every subinterval $[x_{j-1}, x_j]$ $(j = 1, \ldots, N)$ we introduce $m \in \mathbf{N}$ interpolation (collocation) points

$$x_{jl} = x_{j-1} + \eta_l (x_j - x_{j-1}), \quad l = 1, \ldots, m;\ j = 1, \ldots, N, \tag{8}$$

where $\eta_1, \ldots, \eta_m$ are some fixed (collocation) parameters such that

$$0 \le \eta_1 < \ldots < \eta_m \le 1. \tag{9}$$

We find an approximation $v_N = v_{N,m,r,\varphi}$ to u_φ, the solution of (7) (under the conditions of Theorems 1 and 2 below (1) and (7) are uniquely solvable), by collocation method from the following conditions:

$$v_N \in S_{m-1}^{(-1)}(\Delta_N^r), \quad N, m \in \mathbf{N},\ r \ge 1, \tag{10}$$

$$v_N(x_{jl}) = \int_0^{x_{jl}} K_\varphi(x_{jl}, s) v_N(s) ds + f_\varphi(x_{jl}), \quad l = 1, \ldots, m;\ j = 1, \ldots, N, \tag{11}$$

with x_{jl}, $l = 1, \ldots, m$; $j = 1, \ldots, N$, given by formula (8).

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 (1), setting

$$u_N(x) = v_N(\varphi^{-1}(x)), \quad 0 \le x \le b. \tag{12}$$

The settings (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 [10] for a convenient choice of it.

Theorem 1. *[8]. Let the following conditions be fulfilled:*
1. $K \in W^{m,\nu,\lambda}(D_b)$, $f \in C^{m,\nu+\lambda}(0, b]$, $m \in \mathbf{N}$, $0 < \nu < 1$, $0 \le \lambda < 1 - \nu$;
2. φ is defined by the formula (5);
3. The interpolation nodes (8) with grid points (2) and parameters (9) are used.

Then the settings (10)–(12) determine for $N \ge N_0$ a unique approximation u_N to u, the solution to (1), and

$$\|u_N - u\|_\infty \le c \begin{cases} N^{-r\varrho(1-\nu-\lambda)} & \text{for } 1 \le r < \frac{m}{\varrho(1-\nu-\lambda)}, \\ N^{-m} & \text{for } r \ge \frac{m}{\varrho(1-\nu-\lambda)},\ r \ge 1, \end{cases} \tag{13}$$

where $\|u_N - u\|_\infty = \sup_{0 \le x \le b} |u_N(x) - u(x)|$ and c is a positive constant not depending on N.

Theorem 1 proposes, in particular, how r and ρ should be chosen to achieve the highest convergence order $\|u_N - u\|_\infty = \|v_N - v\|_\infty \le cN^{-m}$ by splines of degree $m-1$. Especially, it follows from Theorem 1 that the accuracy $\|u_N - u\|_\infty \le cN^{-m}$ can be achieved on a mildly graded or uniform grid. As an example, if we assume that $\nu = 1/4$, $\lambda = 1/2$, $m = 3$ (the case of piecewise quadratic polynomials), $\varrho \ge 12$, the maximal convergence order $\|u_N - u\|_\infty \le cN^{-3}$ is available for $r \ge 1$. In particular, the uniform grid with nodes (2), $r = 1$, may be used.

5 Superconvergence Results

In this section we see that, in addition to Theorem 1, assuming some additional smoothness of f and g (see (3)) and choosing more carefully the collocation parameters (9), the superconvergence of v_N at the collocation points (8) can be established, cf. [1, 3–5, 10, 13].

Theorem 2. *Assume that $K \in W^{m+1,\nu,\lambda}(D_b)$, $f \in C^{m+1,\nu+\lambda}(0,b]$, $m \in \mathbf{N}$, $0 < \nu < 1$, $0 \le \lambda < 1-\nu$, and let the interpolation nodes (8) be generated by the grid points (2) and by the node points $\eta_1, \ldots, \eta_m$ of a quadrature approximation*

$$\int_0^1 z(s)ds \approx \sum_{l=1}^{m} w_l z(\eta_l), \quad 0 \le \eta_1 < \ldots < \eta_m \le 1, \tag{14}$$

which, with appropriate weights $\{w_l\}$, is exact for all polynomials of degree m. Then for sufficiently large N, say $N \ge N_0$, method (10)–(12) determines a unique approximation u_N to u, the solution of (1), and the following error estimate holds:

$$\max_{\substack{l=1,\ldots,m;\\ j=1,\ldots,N}} |u_N(\varphi(x_{jl})) - u(\varphi(x_{jl}))| = \max_{\substack{l=1,\ldots,m;\\ j=1,\ldots,N}} |v_N(x_{jl}) - u_\varphi(x_{jl})| \le c E_N^{(m,\nu,\lambda,\rho,r)}. \tag{15}$$

Here c is a positive constant not depending on N and

$$E_N^{(m,\nu,\lambda,\rho,r)} = \begin{cases} N^{-2\varrho r(1-\nu-\lambda)} & \text{for} \quad 1 \le r < \frac{m+1-\nu}{2\varrho(1-\nu-\lambda)}, \\ N^{-m-(1-\nu)} & \text{for} \quad r \ge \frac{m+1-\nu}{2\varrho(1-\nu-\lambda)}, \quad r \ge 1. \end{cases} \tag{16}$$

Proof. We outline the basic ideas on which the proof is based. Let T_{K_φ} be defined by the formula $(T_{K_\varphi} u_\varphi)(t) = \int_0^t K_\varphi(t,s) u_\varphi(s)\, ds$, $0 \le t \le b$. Using T_{K_φ}, we write (7) in the form $u_\varphi = T_{K_\varphi} u_\varphi + f_\varphi$, where $f_\varphi \in C[0,b]$ and T_{K_φ} is compact as an operator from $L^\infty(0,b)$ into $C[0,b]$. Since the corresponding homogenous equation $u_\varphi = T_{K_\varphi} u_\varphi$ has only the trivial solution $u_\varphi = 0$, equation $u_\varphi = T_{K_\varphi} u_\varphi + f_\varphi$ has a unique solution $u_\varphi \in C[0,b]$, $u_\varphi(t) = u(\varphi(t))$, $0 \le t \le b$, with u, the solution to (1). On the basis of Lemmas 1 and 2 we find that u_φ belongs to $C^m(0,b]$ and satisfies (6). Further, conditions (10) and (11) have the operator equation representation $v_N = P_N T_{K_\varphi} v_N + P_N f_\varphi$, where P_N is an operator which assigns to any function $z \in C[0,b]$ its piecewise interpolation function $P_N z \in S_{m-1}^{(-1)}(\Delta_N^r)$ interpolating z at the points (8). Using a standard arguing (cf. [10, 12, 13]), we obtain that equation $v_N = P_N T_{K_\varphi} v_N + P_N f_\varphi$ has a unique solution $v_N \in S_{m-1}^{(-1)}(\Delta_N^r)$ for $N \ge N_0$ and $\|v_N - P_N u_\varphi\|_\infty \le c \|T_{K_\varphi}(P_N u_\varphi - u_\varphi)\|_\infty$. We have

$$\begin{aligned} |u_N(\varphi(x_{jl})) - u(\varphi(x_{jl}))| &= |v_N(x_{jl}) - u_\varphi(x_{jl})| = |v_N(x_{jl}) - (P_N u_\varphi)(x_{jl})| \\ &\le \|v_N - P_N u_\varphi\|_\infty, \quad l = 1, \ldots, m;\ j = 1, \ldots, N. \end{aligned}$$

Therefore, by $N \geq N_0$,

$$\max_{l=1,\ldots,m;\, j=1,\ldots,N} |u_N(\varphi(x_{jl})) - u(\varphi(x_{jl}))| \leq c \|T_{K_\varphi}(P_N u_\varphi - u_\varphi)\|_\infty, \tag{17}$$

where

$$\begin{aligned} \|T_{K_\varphi}(P_N u_\varphi - u_\varphi)\|_\infty &= \sup_{0\leq t\leq b} \left| \int_0^t K_\varphi(t,s) \left[(P_N u_\varphi)(s) - u_\varphi(s) \right] ds \right| \\ &\leq \sup_{0\leq t\leq x_1} |S(t)| + \sup_{x_1\leq t\leq b} |S(t)|, \end{aligned} \tag{18}$$

with

$$S(t) = \int_0^t K_\varphi(t,s) \left[(P_N u_\varphi)(s) - u_\varphi(s) \right] ds, \quad 0 \leq t \leq b. \tag{19}$$

Due to the hypothesis of theorem we can prove that $\sup_{0\leq t\leq x_1} |S(t)| \leq c E_N^{(m,\nu,\lambda,\rho,r)}$, $\sup_{x_1\leq t\leq b} |S(t)| \leq c E_N^{(m,\nu,\lambda,\rho,r)}$, with a positive constant c which is independent of N. This together with (17)–(19) yields (15). □

6 Numerical Example

Let us consider the following equation:

$$u(x) = \int_0^x (x-y)^{-\nu} y^{-\lambda} u(y)\, dy + f(x), \quad 0 \leq x \leq 1, \tag{20}$$

where $0 < \nu < 1$, $0 \leq \lambda < 1$, $\nu + \lambda < 1$. The forcing function f is selected so that $u(x) = x^{1-\nu-\lambda}$ is the exact solution to (20). Actually, this is a problem of the form (1) and (3) where $b = 1$, $g(x,y) \equiv 1$, $K(x,y) = (x-y)^{-\nu} y^{-\lambda}$,

$$f(x) = x^{1-\nu-\lambda} - x^{2(1-\nu-\lambda)} \frac{\Gamma(1-\nu)\,\Gamma(2(1-\lambda)-\nu)}{\Gamma(3-2(\nu+\lambda))}, \quad 0 \leq x \leq 1,$$

with $\Gamma(t) = \int_0^\infty e^{-s} s^{t-1}\, ds$, $t > 0$. It is easy to check that in this case $K \in W^{m,\nu,\lambda}(D_1)$ and $f \in C^{m,\nu+\lambda}(0,1]$ for arbitrary $m \in \mathbf{N}$.

Equation (20) was solved numerically by method (10)–(12) for $\nu = 1/4$, $\lambda = 1/2$, $m = 3$, $\eta_1 = (5-\sqrt{15})/10$, $\eta_2 = 1/2$, $\eta_3 = (5+\sqrt{15})/10$. Here η_1, η_2, η_3 are the node points of the Gauss–Legendre quadrature rule (14) by $m = 3$. This formula 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. The quantities $\varepsilon_N^{(\varrho,r)}$ in Table 1 are approximate values of the norm

Table 1 $\left(m = 3, \nu = \frac{1}{4}, \lambda = \frac{1}{2}, \eta_1 = \frac{5-\sqrt{15}}{10}, \eta_2 = \frac{1}{2}, \eta_3 = \frac{5+\sqrt{15}}{10}\right)$

N	$\delta_N^{(1,1)}$	$\delta_N^{(3,2)}$	$\delta_N^{(7,1)}$	$\delta_N^{(12,1)}$	$\widetilde{\delta}_N^{(1,1)}$	$\widetilde{\delta}_N^{(3,2)}$	$\widetilde{\delta}_N^{(7,1)}$	$\widetilde{\delta}_N^{(12,1)}$
	$\epsilon_N^{(1,1)}$	$\epsilon_N^{(3,2)}$	$\epsilon_N^{(7,1)}$	$\epsilon_N^{(12,1)}$	$\gamma_N^{(1,1)}$	$\gamma_N^{(3,2)}$	$\gamma_N^{(7,1)}$	$\gamma_N^{(12,1)}$
	1.71	2.83	3.37	8.62	1.71	8.05	11.35	12.44
32	8.4 E – 1	3.5 E – 4	3.1 E – 5	1.7 E – 6	8.4 E – 1	1.6 E – 7	1.4 E – 8	1.1 E – 7
	1.68	2.83	3.36	8.40	1.67	8.02	11.32	12.93
64	5.0 E – 1	1.2 E – 4	9.3 E – 6	2.1 E – 7	5.0 E – 1	2.0 E – 8	1.2 E – 9	8.2 E – 9
	1.66	2.83	3.36	8.25	1.66	8.01	11.32	13.13
128	3.0 E – 1	4.3 E – 5	2.8 E – 6	2.5 E – 8	3.0 E – 1	2.5 E – 9	1.1 E – 10	6.3 E – 10
	1.65	2.83	3.36	8.16	1.65	8.00	11.32	13.45
256	1.8 E – 1	1.5 E – 5	8.2 E – 7	3.1 E – 9	1.8 E – 1	3.1 E – 10	9.5 E – 12	4.7 E – 11
	1.65	2.83	3.36	8.10	1.65	8.00	11.31	14.78
512	1.1 E – 1	5.4 E – 6	2.4 E – 7	3.8 E – 10	1.1 E – 1	3.9 E – 11	8.4 E – 13	3.2 E – 12
	1.19	2.83	3.36	8.00	1.41	8.00	11.31	13.45

$\|u_N - u\|_\infty$, calculated as follows:

$$\varepsilon_N^{(\varrho,r)} = \max_{\substack{l=0,\dots,10\\ j=1,\dots,N}} |u_N((\tau_{jl}^{(r)})^\varrho) - u((\tau_{jl}^{(r)})^\varrho)|\,,$$

where $\tau_{jl}^{(r)} = x_{j-1} + l(x_j - x_{j-1})/10, \quad l = 0,\dots,10; \quad j = 1,\dots,N$, with the grid points x_j, defined by formula (2) for $b = 1$. The last four columns of Table 1 show the dependance of

$$\gamma_N^{(\varrho,r)} = \max_{\substack{l=1,\dots,m;\\ j=1,\dots,N}} |u_N(\varphi(x_{jl})) - u(\varphi(x_{jl}))| = \max_{\substack{l=1,\dots,m;\\ j=1,\dots,N}} |v_N(x_{jl}) - u_\varphi(x_{jl})|$$

on the parameters N, ϱ and r (see (15)). The ratios

$$\delta_N^{(\varrho,r)} = \varepsilon_{N/2}^{(\varrho,r)} \Big/ \varepsilon_N^{(\varrho,r)}, \qquad \widetilde{\delta}_N^{(\varrho,r)} = \gamma_{N/2}^{(\varrho,r)} \Big/ \gamma_N^{(\varrho,r)},$$

characterizing the observed convergence rate, are also presented.

From Theorem 1 it follows that for sufficiently large N,

$$\varepsilon_N^{(\varrho,r)} \approx \|u_N - u\|_\infty \le c \begin{cases} N^{-\varrho r/4} & \text{if} \quad 1 \le \varrho r < 12, \\ N^{-3} & \text{if} \quad \varrho r \ge 12\,. \end{cases} \tag{21}$$

Due to (21), the ratio $\delta_N^{(\varrho,r)}$ ought to be approximately $(N/2)^{-\varrho r/4}/N^{-\varrho r/4} = 2^{\varrho r/4}$ for $1 \le \varrho r < 12$ and 8 for $\varrho r \ge 12$. In particular, $\delta_N^{(1,1)}$, $\delta_N^{(3,2)}$, $\delta_N^{(7,1)}$ and $\delta_N^{(12,1)}$ ought to be approximately 1.19, 2.83, 3.36, and 8.00, respectively.

In a similar way we obtain from Theorem 2 (see (15) and (16)) that $\widetilde{\delta}_N^{(1,1)}$, $\widetilde{\delta}_N^{(3,2)}$, $\widetilde{\delta}_N^{(7,1)}$, and $\widetilde{\delta}_N^{(12,1)}$ ought to be approximately 1.41, 8.00, 11.31, and 13.45, respectively. These values of $\delta_N^{(\varrho,r)}$ and $\widetilde{\delta}_N^{(\varrho,r)}$ are given in the last row of Table 1.

As we can see from Table 1, the numerical results are in good agreement with the theoretical estimates.

Acknowledgements This work has been supported by Estonian Science Foundation, grant No. 7353.

References

1. K.E. Atkinson. The Numerical Solution of Integral Equations of the Second Kind. Cambridge University Press, Cambridge, 1997
2. 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
3. H. Brunner. Collocation Methods for Volterra Integral and Related Functional Equations, Cambridge Monographs on Applied and Computational Mathematics, 15. Cambridge University Press, Cambridge, 2004
4. H. Brunner, P.J. van der Houwen. The Numerical Solution of Volterra Equations, CWI Monographs, 3. North Holland, Amsterdam, 1986
5. H. Brunner, A. Pedas, G. Vainikko. The piecewise polynomial collocation method for nonlinear weakly singular Volterra eqnations. Math. Comput., 68:1079–1095, 1999
6. Y. Cao, M. Huang, L. Liu, Y. Xu. Hybrid collocation methods for Fredholm integral equations with weakly singular kernels. Appl. Numer. Math, 57:549–567, 2007
7. 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
8. 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
9. G. Monegato, L. Scuderi. High order methods for weakly singular integral equations with nonsmooth input functions. Math. Comput., 67:1493–1515, 1998
10. A. Pedas, G. Vainikko. Smoothing transformation and piecewise polynomial collocation for weakly singular Volterra integral equations. Comput., 73:271–293, 2004
11. A. Pedas, G. Vainikko. Integral equations with diagonal and boundary singularities of the kernel. ZAA, 25(4):487–516, 2006
12. A. Pedas, G. Vainikko. Smoothing transformation and piecewise polynomial projection methods for weakly singular Fredholm integral equations. Commun. Pure Appl. Math., 5:395–413, 2006
13. G. Vainikko. Multidimensional Weakly Singular Integral Equations. Springer, Berlin, 1993
14. E. Vainikko, G. Vainikko. A spline product quasi-interpolation method for weakly singular Fredholm integral equations. SIAM J. Numer. Anal., 46:1799–1820, 2008

Non-Conforming Finite Element Method for the Brinkman Problem

Juho Könnö and Rolf Stenberg

Abstract The Brinkman equations describe the flow of a viscous fluid in a porous matrix. Mathematically the Brinkman model is a parameter-dependent combination of the Darcy and Stokes models. A dual mixed framework is introduced for the problem, and $H(\mathrm{div})$-conforming finite elements are used with Nitsche's method to obtain a stable formulation. We show the formulation to be stable in a mesh-dependent norm for all values of the parameter and introduce a postprocessing scheme for the pressure, which gives optimal convergence for the pressure.

1 Introduction

We study the application of $H(\mathrm{div})$-conforming finite elements designed for the Darcy problem to the more complicated Brinkman problem. This constitutes a non-conforming approximation of the Brinkman problem. For an analysis of a conforming method see e.g., [4] and the references therein. To obtain a stable method, the so-called Nitsche's method first introduced in [8] is used. This in turn requires the use of a mesh-dependent bilinear form. The motivation behind using this non-conforming approximation is the fact that $H(\mathrm{div})$-conforming elements are widely used in industry for solving the Darcy equation, and we want to easily show a way of incorporating viscosity to the existing implementations.

2 The Brinkman Model

The Brinkman model describes the flow of a fluid in a porous medium [3]. For a derivation of the Brinkman equations, see e.g., [1, 6]. The main difference to the simpler Darcy problem is the introduction of viscosity to the equations. Let $\boldsymbol{u}$ be the

J. Könnö (✉) and R. Stenberg
Helsinki University of Technology, Department of Mathematics and Systems Analysis,
P.O. Box 1100, 02015 TKK, Finland
e-mail: juho.konno@tkk.fi, rolf.stenberg@tkk.fi

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_55,

velocity field of the fluid, p the pore pressure, and $\Omega \in \mathbb{R}^n$, with $n = 2, 3$. Denoting by the parameter t the effective viscosity of the fluid, the Brinkman equations are

$$-t^2 \Delta \boldsymbol{u} + \boldsymbol{u} - \nabla p = \boldsymbol{f}, \quad \text{in } \Omega \tag{1}$$

$$\operatorname{div} \boldsymbol{u} = g, \quad \text{in } \Omega \tag{2}$$

For $t > 0$, the equations have formally the same structure as the Stokes problem. The solution $(\boldsymbol{u}, p)$ is sought in $\boldsymbol{V} \times Q = [H^1(\Omega)]^n \times L_0^2(\Omega)$. For the case $t = 0$ we get the Darcy problem, and accordingly the solution is sought in $\boldsymbol{V} \times Q = H(\operatorname{div}, \Omega) \times L_0^2(\Omega)$.

We define the following bilinear forms

$$a(\boldsymbol{u}, \boldsymbol{v}) = t^2(\nabla \boldsymbol{u}, \nabla \boldsymbol{v}) + (\boldsymbol{u}, \boldsymbol{v}), \tag{3}$$

$$b(\boldsymbol{v}, p) = (\operatorname{div} \boldsymbol{v}, p), \tag{4}$$

and

$$\mathscr{B}(\boldsymbol{u}, p; \boldsymbol{v}, q) = a(\boldsymbol{u}, \boldsymbol{v}) + b(\boldsymbol{v}, p) + b(\boldsymbol{u}, q). \tag{5}$$

The Brinkman problem in the weak formulation then reads: Find $(\boldsymbol{u}, p) \in \boldsymbol{V} \times Q$ such that

$$\mathscr{B}(\boldsymbol{u}, p; \boldsymbol{v}, q) = (\boldsymbol{f}, \boldsymbol{v}) + (g, q), \quad \forall (\boldsymbol{v}, q) \in \boldsymbol{V} \times Q. \tag{6}$$

3 Solution by Mixed Finite Elements

3.1 Mesh Dependent Norms

We introduce the following mesh-dependent norms for the problem. Note, that both of the norms are also parameter-dependent. We denote the jump in the value of a generic function f on the edge E by $[\![f]\!] = f|_{K_1} - f|_{K_2}$, where $E = \partial K_1 \bigcap \partial K_2$. Similarly, the average on the edge is denoted $\{f\} = \frac{1}{2}(f|_{K_1} + f|_{K_2})$. For the velocity we use the norm

$$\|\boldsymbol{u}\|_{t,h}^2 = \|\boldsymbol{u}\|^2 + t^2 \sum_{K \in \mathscr{K}_h} \|\nabla \boldsymbol{u}\|_{0,K}^2 + t^2 \sum_{E \in \mathscr{E}_h} \frac{1}{h_E} \|[\![\boldsymbol{u} \cdot \boldsymbol{\tau}]\!]\|_{0,E}^2, \tag{7}$$

and for the pressure

$$|\!|\!| p |\!|\!|_{t,h}^2 = \sum_{K \in \mathscr{K}_h} \frac{h_K^2}{h_K^2 + t^2} \|\nabla p\|_{0,K}^2 + \sum_{E \in \mathscr{E}_h} \frac{h_E}{h_E^2 + t^2} \|[\![p]\!]\|_{0,E}^2. \tag{8}$$

3.2 Mixed Method

For simplicity, we only prove the stability results for the Raviart–Thomas spaces. All presented results also hold for the Brezzi–Douglas–Marini family of elements, since $\boldsymbol{V}_h^{RT} \subset \boldsymbol{V}_h^{BDM}$ and $Q_h^{BDM} = Q_h^{RT}$. The spaces of order k are [2]

$$\boldsymbol{V}_h^{RT} = \{\boldsymbol{v} \in H(\mathrm{div}, \Omega) \mid \boldsymbol{v}|_K \in [P_{k-1}(K)]^n \oplus \boldsymbol{x}\widetilde{P}_{k-1}(K)\ \forall K \in \mathscr{K}_h\}, \quad (9)$$

$$\boldsymbol{V}_h^{BDM} = \{\boldsymbol{v} \in H(\mathrm{div}, \Omega) \mid \boldsymbol{v}|_K \in [P_k(K)]^n\ \forall K \in \mathscr{K}_h\}, \quad (10)$$

$$Q_h = \{q \in L^2(\Omega) \mid q|_K \in P_{k-1}(K)\ \forall K \in \mathscr{K}_h\}, \quad (11)$$

in which $\widetilde{P}_{k-1}(K)$ denotes the homogeneous polynomials of degree $k-1$. The spaces are chosen such that the following equilibrium property holds:

$$\mathrm{div}\, \boldsymbol{V}_h \subset Q_h. \quad (12)$$

To obtain a stable non-conforming method, Nitsche's method with a suitably chosen stabilization parameter α is used. We define the following mesh-dependent bilinear form

$$\mathscr{B}_h(\boldsymbol{u}, p; \boldsymbol{v}, q) = a_h(\boldsymbol{u}, \boldsymbol{v}) + b(\boldsymbol{v}, p) + b(\boldsymbol{u}, q), \quad (13)$$

in which

$$a_h(\boldsymbol{u}, \boldsymbol{v}) = (\boldsymbol{u}, \boldsymbol{v}) + t^2 \sum_{K \in \mathscr{K}_h} (\nabla \boldsymbol{u}, \nabla \boldsymbol{v})_K + t^2 \sum_{E \in \mathscr{E}_h} \{\frac{\alpha}{h_E}\langle [\![\boldsymbol{u}]\!], [\![\boldsymbol{v}]\!]\rangle_E - \langle \{\frac{\partial \boldsymbol{u}}{\partial n}\}, [\![\boldsymbol{v}]\!]\rangle_E - \langle \{\frac{\partial \boldsymbol{v}}{\partial n}\}, [\![\boldsymbol{u}]\!]\rangle_E\}. \quad (14)$$

Then the discrete problem is to find $\boldsymbol{u}_h \in \boldsymbol{V}_h$ and $p_h \in Q_h$ such that

$$\mathscr{B}_h(\boldsymbol{u}_h, p_h; \boldsymbol{v}, q) = (\boldsymbol{f}, \boldsymbol{v}) + (g, q), \quad \forall (\boldsymbol{v}, q) \in \boldsymbol{V}_h \times Q_h. \quad (15)$$

The modified method is consistent by the following theorem.

Theorem 1. *For the exact solution* $(\boldsymbol{u}, p) \in \boldsymbol{V} \times Q$ *it holds*

$$\mathscr{B}_h(\boldsymbol{u}, p; \boldsymbol{v}, q) = (\boldsymbol{f}, \boldsymbol{v}) + (g, q), \quad \forall (\boldsymbol{v}, q) \in \boldsymbol{V}_h \times Q_h. \quad (16)$$

Next we prove the stability of $a_h(\cdot, \cdot)$ in the mesh-dependent norm (7). The stability only holds in the discrete space $\boldsymbol{V}_h$, since we need to use the inverse inequality. Estimating the negative term using Young's inequality in the following expression from below gives

$$a_h(\boldsymbol{v},\boldsymbol{v}) = \|\boldsymbol{v}\|_0^2 + t^2 \sum_{K\in\mathscr{K}_h} \|\nabla \boldsymbol{v}\|_{0,E}^2 + t^2 \times \sum_{E\in\mathscr{E}_h} \left(\frac{\alpha}{h_E}\|[\![\boldsymbol{v}]\!]\|_{0,E}^2 - 2\langle \frac{\partial \boldsymbol{v}}{\partial n}, [\![\boldsymbol{v}]\!]\rangle_E\right)$$
$$\geq \min\{1 - \frac{C_I}{2\epsilon}, \alpha - \frac{\epsilon}{2}\}\|\boldsymbol{v}\|_{t,h}^2. \tag{17}$$

Here C_I is the constant from the discrete trace inequality. Since ϵ and α are free parameters, choosing $\epsilon > C_I/2$ and $\alpha > \epsilon/2$ gives

$$a_h(\boldsymbol{v},\boldsymbol{v}) \geq C\|\boldsymbol{v}\|_{t,h}^2, \quad \forall \boldsymbol{v} \in \boldsymbol{V}_h, \tag{18}$$

with a constant $C > 0$. The method also satisfies the discrete Brezzi–Babuska stability condition [2] in the mesh-dependent norms (7) and (8). One only has to prove the condition in the Raviart–Thomas case since $\boldsymbol{V}_h^{RT} \subset \boldsymbol{V}_h^{BDM}$. The proof follows the lines of [7].

Lemma 1. *There exists a positive constant C such that*

$$\sup_{\boldsymbol{v}\in\boldsymbol{V}_h} \frac{b(\boldsymbol{v},q)}{\|\boldsymbol{v}\|_{t,h}} \geq C|\!|\!|q|\!|\!|_{t,h}, \quad \forall q \in Q_h. \tag{19}$$

Combining the above stability results for $a_h(\cdot,\cdot)$ and $b(\cdot,\cdot)$ yields the following full stability result.

Lemma 2. *There exists a positive constant C such that*

$$\sup_{(\boldsymbol{v},q)\in\boldsymbol{V}_h\times Q_h} \frac{\mathscr{B}_h(\boldsymbol{r},s;\boldsymbol{v},q)}{\|\boldsymbol{v}\|_{t,h} + |\!|\!|q|\!|\!|_{t,h}} \geq C(\|\boldsymbol{r}\|_{t,h} + |\!|\!|s|\!|\!|_{t,h}), \quad \forall (\boldsymbol{r},s) \in \boldsymbol{V}_h \times Q_h. \tag{20}$$

Next we recall the interpolation operator $\boldsymbol{R}_h : H(\mathrm{div},\Omega) \to \boldsymbol{V}_h$ [9] satisfying

$$(\mathrm{div}\,(\boldsymbol{v} - \boldsymbol{R}_h\boldsymbol{v}), q) = 0, \quad \forall q \in Q_h. \tag{21}$$

We denote by $P_h : L^2(\Omega) \to V_h$ the L^2-projection. The equilibrium property (12) implies

$$(\mathrm{div}\,\boldsymbol{v}, q - P_h q) = 0, \quad \forall \boldsymbol{v} \in \boldsymbol{V}_h. \tag{22}$$

Furthermore, we have the commuting diagram property:

$$\mathrm{div}\,\boldsymbol{R}_h = P_h \mathrm{div}. \tag{23}$$

We then have the following quasioptimal a priori result. The result is quasioptimal in the sense that the error of the finite element solution is limited by, but not necessarily equal to, the interpolation error.

Theorem 2. *There is a positive constant C such that*

$$\|\boldsymbol{u} - \boldsymbol{u}_h\|_{t,h} + |\!|\!|P_h p - p_h|\!|\!|_{t,h} \leq C\|\boldsymbol{u} - \boldsymbol{R}_h\boldsymbol{u}\|_{t,h}. \tag{24}$$

Proof. By Lemma 2 there exists functions $(\boldsymbol{v}, q) \in \boldsymbol{V}_h \times Q_h$ *with* $\|\boldsymbol{v}\|_{t,h} + |||q|||_{t,h} \leq C$*, such that*

$$\begin{aligned} \|\boldsymbol{u}_h - \boldsymbol{R}_h\boldsymbol{u}\|_{t,h} + |||p_h - P_h p|||_{t,h} &\leq \mathscr{B}_h(\boldsymbol{u}_h - \boldsymbol{R}_h\boldsymbol{u}, p_h - P_h p; \boldsymbol{v}, q) \\ &= a_h(\boldsymbol{u}_h - \boldsymbol{R}_h\boldsymbol{u}, \boldsymbol{v}) + (\operatorname{div} \boldsymbol{v}, p_h - P_h p) + (\operatorname{div} (\boldsymbol{u}_h - \boldsymbol{R}_h\boldsymbol{u}), q) \end{aligned}$$

By using the interpolation properties (21) *and* (22) *along with the consistency property of Theorem 1, we arrive at*

$$\|\boldsymbol{u}_h - \boldsymbol{R}_h\boldsymbol{u}\|_{t,h} + |||p_h - P_h p|||_{t,h} \leq a_h(\boldsymbol{u} - \boldsymbol{R}_h\boldsymbol{u}, \boldsymbol{v}) \leq C\,\|\boldsymbol{u} - \boldsymbol{R}_h\boldsymbol{u}\|_{t,h} \tag{25}$$

Using the triangle inequality yields the result of the theorem. □

Equivalently to the dual mixed Poisson problem, we have a superconvergence result for $|||p_h - P_h p|||_{t,h}$. This implies that the pressure solution can be improved by local postprocessing. Assuming full regularity, one has

$$\|\boldsymbol{u} - \boldsymbol{u}_h\|_{t,h} + |||P_h p - p_h|||_{t,h} \leq \begin{cases} C h^k(\|\boldsymbol{u}\|_k + t\|\boldsymbol{u}\|_{k+1}), & \text{for RT,} \\ C h^{k+1}(\|\boldsymbol{u}\|_{k+1} + t\|\boldsymbol{u}\|_{k+2}), & \text{for BDM.} \end{cases} \tag{26}$$

4 Postprocessing Method

In this section we present a postprocessing method for the pressure in the spirit of [7]. We seek the postprocessed pressure in an augmented space $Q_h^* \supset Q_h$, defined as

$$Q_h^* = \begin{cases} \{q \in L^2(\Omega) \mid q|_K \in P_k(K)\ \forall K \in \mathscr{K}_h\}, & \text{for RT,} \\ \{q \in L^2(\Omega) \mid q|_K \in P_{k+1}(K)\ \forall K \in \mathscr{K}_h\}, & \text{for BDM.} \end{cases} \tag{27}$$

The postprosessing method is: Find $p_h^* \in Q_h^*$ such that

$$P_h p_h^* = p_h, \tag{28}$$

$$(\nabla p_h^*, \nabla q)_K = (-t^2 \Delta \boldsymbol{u}_h + \boldsymbol{u}_h - \boldsymbol{f}, \nabla q)_K, \quad \forall q \in (I - P_h) Q_h^*|_K. \tag{29}$$

The method can be compactly treated as an integral part of the problem by embedding it into the bilinear form. Thus we introduce the modified bilinear form

$$\begin{aligned} \widetilde{\mathscr{B}}_h(\boldsymbol{u}, p^*; \boldsymbol{v}, q^*) &= \mathscr{B}_h(\boldsymbol{u}, p^*; \boldsymbol{v}, q^*) \\ &+ \sum_{K \in \mathscr{K}_h} \frac{h_K^2}{h_K^2 + t^2} (-\nabla p^* + \boldsymbol{u} - t^2 \Delta \boldsymbol{u}, \nabla (I - P_h) q^*)_K. \end{aligned} \tag{30}$$

The postprocessed problem is then: Find $(\boldsymbol{u}_h, p_h^*) \in \boldsymbol{V}_h \times Q_h^*$ such that for every pair $(\boldsymbol{v}, q^*) \in \boldsymbol{V}_h \times Q_h^*$ it holds

$$\widetilde{\mathscr{B}}_h(\boldsymbol{u}_h, p_h^*; \boldsymbol{v}, q^*) = L^2(\Omega)_h(\boldsymbol{f}, P_h g; \boldsymbol{v}, q^*), \tag{31}$$

in which

$$L^2(\Omega)_h(\boldsymbol{f}, g; \boldsymbol{v}, q^*) = (\boldsymbol{f}, \boldsymbol{v}) + (g, q^*) + \sum_{K \in \mathscr{K}_h} \frac{h_K^2}{h_K^2 + t^2} (\boldsymbol{f}, \nabla(I - P_h) q^*)_K. \tag{32}$$

We have the following theorem relating the solution of the postprocessed problem to the original problem.

Theorem 3. *Let $(\boldsymbol{u}_h, p_h^*) \in \boldsymbol{V}_h \times Q_h^*$ be the solution of the problem* (31) *and set $p_h = P_h p_h^*$. Then $(\boldsymbol{u}_h, p_h) \in \boldsymbol{V}_h \times Q_h$ is the solution of the original problem* (15)*. Conversely, if $(\boldsymbol{u}_h, p_h) \in \boldsymbol{V}_h \times Q_h$ is the solution of the original problem* (15) *and p_h^* is defined as above, then $(\boldsymbol{u}_h, p_h^*) \in \boldsymbol{V}_h \times Q_h^*$ is the solution to* (31)*.*

The postprocessed method also has full stability in the mesh-dependent norms.

Theorem 4. *There exists a constant $C > 0$ such that for every $(\boldsymbol{u}, p^*) \in \boldsymbol{V}_h \times Q_h^*$ it holds*

$$\sup_{(\boldsymbol{v}, q^*) \in \boldsymbol{V}_h \times Q_h^*} \frac{\widetilde{\mathscr{B}}_h(\boldsymbol{u}, p^*; \boldsymbol{v}, q^*)}{\|\boldsymbol{v}\|_{t,h} + |\!|\!| q^* |\!|\!|_{t,h}} \geq C(\|\boldsymbol{u}\|_{t,h} + |\!|\!| p^* |\!|\!|_{t,h}). \tag{33}$$

We have the following quasioptimal a priori result.

Theorem 5. *For the postprocessed solution it holds*

$$\begin{aligned} \|\boldsymbol{u} - \boldsymbol{u}_h\|_{t,h} + |\!|\!| p - p_h |\!|\!|_{t,h} \leq C \inf_{q^* \in Q_h^*} \{ & \|\boldsymbol{u} - R_h \boldsymbol{u}\|_{t,h} + |\!|\!| p - q^* |\!|\!|_{t,h} \\ & + \Big(\sum_{K \in \mathscr{K}_h} \frac{h_K^2}{h_K^2 + t^2} \| - \nabla q^* + R_h \boldsymbol{u} - t^2 \Delta R_h \boldsymbol{u} - \boldsymbol{f} \|_{0,K}^2 \Big)^{1/2} \}. \end{aligned} \tag{34}$$

Proof. *Let $q^* \in Q_h^*$. From Theorem 4 it follows that we have a pair $(\boldsymbol{v}, r^*) \in \boldsymbol{V}_h \times Q_h^*$ such that $\|\boldsymbol{v}\|_{t,h} + |\!|\!| r^* |\!|\!|_{t,h} \leq C$ and*

$$\|\boldsymbol{u}_h - R_h \boldsymbol{u}\|_{t,h} + |\!|\!| p_h^* - q^* |\!|\!|_{t,h} \leq C \widetilde{\mathscr{B}}_h(\boldsymbol{u}_h - R_h \boldsymbol{u}, p_h^* - q^*; \boldsymbol{v}, r^*).$$

Combining the definition of the postprocessed problem and the consistency result of Theorem 1 gives

$$\|\boldsymbol{u}_h - R_h\boldsymbol{u}\|_{t,h} + |\!|\!| p_h^* - q^* |\!|\!|_{t,h} \le C\,\widetilde{\mathscr{B}}_h(\boldsymbol{u} - R_h\boldsymbol{u},\, p - q^*; \boldsymbol{v}, r^*) - (g - P_h g, r^*)$$
$$= a_h(\boldsymbol{u} - R_h\boldsymbol{u}, \boldsymbol{v}) + (\operatorname{div} \boldsymbol{v},\, p - q^*) + (\operatorname{div}\,(\boldsymbol{u} - R_h\boldsymbol{u}), r^*) - (g - P_h g, r^*)$$
$$\sum_{K\in\mathscr{K}_h} \frac{h_K^2}{h_K^2 + t^2}(-\nabla(p - q^*) + (\boldsymbol{u} - R_h\boldsymbol{u}) + -t^2\Delta(\boldsymbol{u} - R_h\boldsymbol{u}), \nabla(I - P_h)r^*)_K.$$

The last two terms on the second line cancel by the commuting diagram property (23)*. Inserting* $\boldsymbol{f}$ *into the last line of the previous equation we have*

$$\|\boldsymbol{u}_h - R_h\boldsymbol{u}\|_{t,h} + |\!|\!| p_h^* - q^* |\!|\!|_{t,h} \le C\{\|\boldsymbol{u} - R_h\boldsymbol{u}\|_{t,h}\|\boldsymbol{v}\|_{t,h} + |\!|\!| p - q^* |\!|\!|_{t,h}\|\boldsymbol{v}\|_{t,h}$$
$$+ \Big(\sum_{K\in\mathscr{K}_h} \frac{h_K^2}{h_K^2 + t^2}\|\nabla q^* - R_h\boldsymbol{u} + t^2\Delta R_h\boldsymbol{u} + \boldsymbol{f}\|_{0,K}^2\Big)^{1/2} |\!|\!| r^* |\!|\!|_{t,h},$$

thus the assertion is proved. □

Theorem 5 shows that the postprocessed pressure converges to the exact solution in the mesh-dependent norm at an optimal convergence rate. This gives the proposed method a good balance between accuracy and keeping the number of degrees of freedom for the pressure relatively low in the original system. This is a particularly important property, since the pressure space is discontinuous.

5 Conclusions

It was shown that Nitsche's method can be successfully applied to using $H(div)$-conforming elements as a non-conforming approximation for the Brinkman problem. The method is stable for all values of the viscosity parameter t. We were able to extend the postprocessing scheme introduced for the Darcy case to the Brinkman problem, thus recovering optimal convergence rates for both the variables. The postprocessing procedure is also essential in deriving a reliable and sharp residual-based a posteriori estimator for the problem. This issue will be addressed in an upcoming article by the authors [5].

Furthermore, since the postprocessing is performed elementwise, the procedure is numerically lightweight adding very little computational cost to the original problem. This makes the non-conforming approximation a viable alternative to the Stokes-based approach, even though it adds complexity to the implementation.

Both the a priori and a posteriori performance of the method will be investigated numerically in an upcoming paper by the authors.

Acknowledgements This work has been supported by the KYT 2010 Finnish Research Programme on Nuclear Waste Management. The first author has been supported by a grant from the Finnish Cultural Foundation.

References

1. T. Arbogast and H. L. Lehr. Homogenization of a Darcy-Stokes system modeling vuggy porous media. *Comput. Geosci.*, 10(3):291–302, 2006
2. F. Brezzi and M. Fortin. *Mixed and Hybrid Finite Element Methods*. Springer, Berlin, 1991
3. P. Hansbo and M Juntunen. Weakly imposed Dirichlet boundary conditions for the Brinkman model of porous media flow. *Appl. Numer. Math.*, 59(6):1274–1289, 2009
4. M. Juntunen and R. Stenberg. Analysis of finite element methods for the Brinkman problem. *Calcolo*, 2009. doi:10.1007/s10092-009-0017-6
5. J. Könnö and R. Stenberg. Analysis of H(div)-conforming finite elements for the Brinkman problem. *Helsinki University of Technology Institute of Mathematics Research Report A 582 (2010)*
6. T. Lévy. Loi de Darcy ou loi de Brinkman? *C. R. Acad. Sci. Paris Sér. II Méc. Phys. Chim. Sci. Univers Sci. Terre*, 292(12):871–874, Erratum (17):1239, 1981
7. C. Lovadina and R. Stenberg. Energy norm a posteriori error estimates for mixed finite element methods. *Math. Comp.*, 75(256):1659–1674 (electronic), 2006
8. J. Nitsche. Über ein Variationsprinzip zur Lösung von Dirichlet-Problemen bei Verwendung von Teilräumen, die keinen Randbedingungen unterworfen sind. *Abh. Math. Sem. Univ. Hamburg*, 36:9–15, 1971. Collection of articles dedicated to Lothar Collatz on his sixtieth birthday
9. J. Schöberl. Commuting quasi-interpolation operators for mixed finite elements. Preprint ISC-01-10-MATH, Institute for Scientific Computing, Texas AM University, 2001

Error Control for Simulations of a Dissociative Quantum System

Katharina Kormann and Anna Nissen

Abstract We present a framework for solving the Schrödinger equation modeling the interaction of a dissociative quantum system with a laser field. A perfectly matched layer (PML) is used to handle non-reflecting boundaries and the Schrödinger equation is discretized with high-order finite differences in space and an h, p-adaptive Magnus–Arnoldi propagator in time. We use a posteriori error estimation theory to control the global error of the numerical discretization. The parameters of the PML are chosen to meet the same error tolerance. We apply our framework to the IBr molecule, for which numerical experiments show that the total error can be controlled efficiently. Moreover, we provide numerical evidence that the Magnus–Arnoldi solver outperforms the implicit Crank–Nicolson scheme by far.

1 Introduction

The quantum system of a molecule can be described by the time-dependent Schrödinger equation (TDSE),

$$\mathrm{i}\hbar\frac{\partial\psi}{\partial t} = H_0\psi,$$

where $H_0 = -\frac{\hbar^2}{2m}\Delta + V(x)$ is the Hamiltonian, $\hbar$ Planck's reduced constant, m the mass of the system, and $V(x)$ is a space-dependent potential. To describe the interaction of the molecule with a time-dependent field, one adds a time-dependent coupling term to the Hamiltonian. We consider the nuclear motion with various (fixed) electronic states. These states can be coupled statically through crossings in

K. Kormann (✉) and A. Nissen
Division of Scientific Computing, Department of Information Technology, Uppsala University, Sweden
e-mail: katharina.kormann@it.uu.se,anna.nissen@it.uu.se

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_56,

the potential energy, or dynamically by a laser field. We refer to [12] for a more detailed description of the chemical model.

In this article, we are especially interested in the case where unstable electronic states are included, so that the wave function is not concentrated to a bounded domain. For the numerical simulations we need to truncate the domain and, in order to avoid reflections from the numerical boundary, artificial damping needs to be imposed close to the boundary. In [11] approximate error formulas were derived for a damping based on a perfectly matched layer (PML). Here, we combine this boundary model with efficient Magnus–Arnoldi time-propagation, see [8]. We propose a procedure of balancing errors that arise from the boundary modeling as well as the spatial and temporal discretization.

As a sample system, we consider a three-state IBr system with a Hamiltonian of the form

$$H = \begin{pmatrix} -\frac{\hbar^2}{2m}\Delta + V_1(x) & \mu(x)\varepsilon(t) & 0 \\ \mu(x)\varepsilon(t) & -\frac{\hbar^2}{2m}\Delta + V_2(x) & V_c(x) \\ 0 & V_c(x) & -\frac{\hbar^2}{2m}\Delta + V_3(x) \end{pmatrix}, \tag{1}$$

where μ is the dipole moment, ε is the time-dependent field, and V_c is the statical coupling term. In our example, the second and the third states are dissociative.

2 Domain Truncation and Discretization

Firstly, we truncate the computational domain to $\Omega = [\underline{x}, \overline{x}] \subset R^+$ which should contain the domain of interest for the purpose of the respective computation. The left boundary is chosen such that the wave function is (almost) zero for smaller values of x. The right boundary is chosen so that the space-dependent potentials are slowly varying at the boundary and that Ω includes the range of the involved bounded states. By assuming that the potentials are constant in space, we derive a PML using a modal ansatz as an extension to the right boundary (cf. [5]). The modified TDSE with boundary treatment is rewritten as a complex symmetric expression (see [7]), where the second derivatives for the dissociative second and third state in (1) are replaced by

$$-\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2} \to -\frac{\hbar^2}{2m}\left\{\frac{1}{f(x)}\frac{\partial^2}{\partial x^2}\frac{1}{f(x)} + F(x)\right\},$$

and solved for the new wave function $\sqrt{f(x)}\psi(x,t)$, where $F(x) = \frac{3f'(x)^2 - 2f''(x)f(x)}{4f(x)^4}$ and $f(x) = 1 + e^{i\pi/4}\sigma(x)$. $\sigma(x)$ is a polynomial damping profile. Note that such a modeling truncates the self-adjointness of the Hamiltonian.

The TDSE formulated on a bounded domain is now discretized based on the method of lines. A spatial grid with step size Δx is introduced and the second

derivatives are computed with a standard eighth order central stencil. Dirichlet conditions are posed at the outer boundaries and central finite difference schemes of order 2, 4, and 6 are used for the points closest to the boundaries.

After discretization in space, the TDSE becomes a system of ordinary differential equations. Common methods to propagate the semi-discretized TDSE are partitioned Runge–Kutta methods and exponential integrators [10, Chap. III]. We use a combination of the Magnus expansion [1] and the Arnoldi algorithm [6]. For this propagator, an h, p-adaptive implementation with global error control was devised in [9].

3 Error Control

In this section, we discuss how to estimate the errors that arise from the boundary model and the discretization in both space and time.

3.1 PML Errors

The boundary model gives rise to a modeling error due to the finite width of the PML as well as numerical reflections due to the discretization of the modified TDSE. Approximate error formulas for a polynomial absorption function of order r, $\sigma(x) = \sigma_{max}\left(\frac{x-\overline{x}}{d}\right)^r$, were derived in [11] for the modeling error and the numerical reflections, respectively, see (3) and (2). Here, $d = \Delta x \cdot n$ is the width of the layer, σ_{max} the maximal value of $\sigma(x)$, and $\overline{x}$ the PML interface.

The error due to numerical reflections, $\varepsilon_{PML,2}$, is approximately

$$\varepsilon_{PML,2} \approx C\frac{\sigma_{max}}{n^r}. \tag{2}$$

We use $r = 2q$, where $2q$ is the order of the finite difference scheme. The constant C can be determined numerically by using a sufficiently wide PML with strong damping and fine spatial and temporal discretization, so that other error sources than $\varepsilon_{PML,2}$ are negligible, see [11].

The modeling error of the PML, $\varepsilon_{PML,1}$, for the polynomial profile is approximately

$$\varepsilon_{PML,1} \approx \mathrm{e}^{\frac{-\sqrt{2}k\sigma_{max}d}{r+1}}. \tag{3}$$

Here, k is the dominating frequency of the system, which we need to determine in order to estimate $\varepsilon_{PML,1}$.

These error estimates can be used to determine the number of points in the PML, n, and the strength, σ_{max}, in a way that the errors are within a prescribed tolerance and of equal size. We only need to solve the problem once in order to determine C in

(2) and we have to identify a suitable value for k based on the particular properties of the system.

For the three state IBr system, the dissociation process takes place in the second and third states and we are thus interested in the energies of the parts of the initial wave packet that have been excited to a higher state by the laser pulse. For simplicity, we calculate k from the second state, since this is where the main part dissociates.

The dominating frequency, k, is determined from the time-independent part of the Hamiltonian for the second state, $H_2 = -\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2} + V_2(x)$. Eigenvectors, ϕ_2, and eigenvalues, E_2, of H_2 are given by $H_2\phi_2 = E_2\phi_2$.

Let ψ_2 be the wave function in the second state. The energy distribution is determined in terms of the Fourier coefficients for the second state, $a_2 =< \phi_2|\psi_2 >$.

By assuming that the potential is constant near the PML interface, we can consider the free particle case, where k is given by

$$k = \sqrt{2m(E_2 - V_2)}. \tag{4}$$

3.2 Discretization Errors

Our aim is to derive an a posteriori error estimate [2] which tells us how to compute the error at final time from the residual at each point in space and time. For this purpose, we view the numerical solution in terms of its continuous interpolant, $\widetilde{\psi}$, which solves a perturbed version of the Schrödinger equation,

$$\mathrm{i}\hbar\frac{\partial}{\partial t}\widetilde{\psi} = H\widetilde{\psi} + R(x,t), \qquad \widetilde{\psi}(x,0) = \psi(x) + R(x,0),$$

where $R(x,t)$ is the perturbation due to numerical approximation. Let us look at the error in some functional defined as $\varepsilon_{\mathrm{dis}} = \int_\Omega \varphi(x)^* \left(\psi(x,t_f) - \widetilde{\psi}(x,t_f)\right)\mathrm{d}x$. As we focus on the L_2 norm of the error, we choose $\varphi = \frac{\varepsilon_{\mathrm{dis}}}{\|\varepsilon_{\mathrm{dis}}\|}$. Defining the adjoint problem,

$$\mathrm{i}\hbar\frac{\partial}{\partial t}\chi = H^*\chi, \qquad \chi(x,t_f) = \varphi(x),$$

we can find the following expression for the error in terms of the residual

$$\begin{aligned}\varepsilon_{\mathrm{dis}} &= \int_0^{t_f}\int_\Omega \chi^*(x,t)R(x,t)\mathrm{d}x\mathrm{d}t + \Delta x\int_\Omega \chi^*(x,0)R(x,0)\mathrm{d}x \\ &\approx \int_0^{t_f} \Delta x\sum_j \chi^*(x_j,t)R(x_j,t)\mathrm{d}t + \Delta x\sum_j \chi^*(x_j,0)R(x_j,0).\end{aligned}$$

In the second step, we have approximated the integral over Ω on the same mesh as the TDSE. In order to be able to compute the error, we have to evaluate the residual. To make this procedure easier, we examine one time step at a time. Then, we can consider the error from previous steps as a perturbation of the initial value. In this way, we only have to estimate the local perturbations (cf. [9]).

Firstly, we look at the residual due to spatial approximation. Instead of computing the second derivative at point x_j exactly, we approximate it by finite differences, i.e., the spatial perturbation at grid point x_j reads $R_s(x_j,t) = -\frac{\hbar^2}{2m}\big((D_{2q}v)_j - \Delta\widetilde{\psi}(x_j,t)\big)$, where D_{2q} denotes a $(2q)$th order accurate finite difference operator and v the vector of the numerical solution at the grid points. Since we do not know the value of $\Delta\widetilde{\psi}(x_j,t)$ and are looking for an easy-to-compute estimate, we use the $(2q+2)$th order finite difference approximation as a reference, that is,

$$R_s(x_j,t) = -\frac{\hbar^2}{2m}\left((D_{2q}v)_j - (D_{2q+2}v)_j\right).$$

Since we suppose that R_s is small compared to $H\widetilde{\psi}$, we assume that we can get a good estimate of the temporal residual R_t when computing it based on the discretized Hamiltonian and neglecting mixed spatial and temporal terms in the residual. This splitting gives the following error estimate

$$\varepsilon_{\text{dis}} \approx \Delta t\,\Delta x \sum_j \sum_l \chi^*(x_j,t_l)R_s(x_j,t_l) + \Delta x \sum_j \sum_l \int_{t_{l-1}}^{t_l} \chi^*(x_j,t)R_t(x_j,t)\mathrm{d}t, \tag{5}$$

where temporal and spatial influences are split.

We consider $\varphi = \frac{\varepsilon_{\text{dis}}}{\|\varepsilon_{\text{dis}}\|}$ which is unknown. Therefore, we cannot actually solve the dual problem. Instead, we split the inner products in (5) using the Cauchy–Schwarz inequality. We also used that $\|\chi\| \approx 1$ for all $t \in [0,t_f]$ since we expect the dual solution to be concentrated on the computational domain where H is self-adjoint (see the discussion in [9]). This yields

$$|\varepsilon_{\text{dis}}| \le \underbrace{\Delta t \sum_l \sqrt{\Delta x \sum_j |R_s(x_j,t_l)|^2}}_{:=\,\varepsilon_{\text{dis,s}}} + \underbrace{\sum_l \int_{t_{l-1}}^{t_l} \sqrt{\Delta x \sum_j |R_t(x_j,t)]^2}\mathrm{d}t}_{:=\,\varepsilon_{\text{dis,t}}}.$$

In order to identify the spatial error, we have to solve the TDSE once and evaluate R_s in each time step. Since we are using a $(2q)$th order stencil, we expect the error to be of the order of the $(2q)$th power of the mesh size. By extrapolation, we can thus compute a guess for the mesh size that should be small enough to meet the given tolerance (tol),

$$\Delta x_{\text{new}} = \sqrt[2q]{\frac{\text{tol}}{\varepsilon_{\text{dis,s}}}}\Delta x.$$

We now turn to the temporal error. We integrate over each time interval to get a sum of local errors instead of local residuals. If we now distribute the error equally over the time intervals, that is, if we make sure that the local error in each interval is less than the tolerance weighted by $\frac{\Delta t}{t_f}$, we can bound the global error in the propagation. As proposed in [9], we control the error in the Magnus expansion by adjusting the step size Δt and the error in the Arnoldi method by the size of the underlying Krylov space (cf. [6]).

3.3 Error Balancing

We conclude this section by formulating a procedure that includes the control of each of the three error sources with the aim of computing the solution of the TDSE to a given accuracy. We suppose that the errors are accumulated and therefore take one third of the total tolerance for each error source. Computing the error estimate for the spatial discretization error requires to solve the problem on a coarse initial mesh. On this mesh one has to identify the parameter C in (2) which facilitates us to find suitable PML parameters. Then the equation can be solved with the adaptive Magnus–Arnoldi method, and an estimate of the spatial error can be computed. Since we expect this first computation to be of low accuracy, it is reasonable to choose a loose tolerance also for boundary modeling and temporal error.

As the next step, we estimate the mesh size Δx suitable for the given tolerance. We then update the PML parameters accordingly and solve the problem again with the adaptive Magnus–Arnoldi solver. If the error of the first run is much larger than the tolerance, one should compute the error estimate again and check whether the extrapolated value for Δx was indeed small enough.

This procedure can be summarized as follows:

1. Set a tolerance and choose an initial spatial grid size Δx.
2. Compute the value of the parameter C in (2) and identify the optimal PML parameters.
3. Solve the TDSE with adaptive Magnus–Arnoldi time stepping and estimate the inner discretization error.
4. Adjust Δx and update the PML parameters correspondingly.
5. Solve the TDSE with adaptive Magnus–Arnoldi on the new grid.

4 Numerical Experiments

We have tested our algorithm for a three-state IBr system. Starting with the lowest eigenfunction of the ground state, we excite the molecule to the $B^3\Pi_0^+$ state with a laser with a wavelength of 500 nm, a width of 50 fs, and strength 219 cm^{-1}. This state is statically coupled to the Y_0^+ dissociative state. For the parameters of the potential energy curves, we refer to [3]. In order to check our results, we compute a reference solution on a spatial mesh with $\Delta x = 5.5 \cdot 10^{-3}$ a.u. ($2.9 \cdot 10^{-3}$ Å) over a large spatial interval where domain truncation is not necessary. All experiments are done with Matlab.

In order to demonstrate the potential of the Magnus–Arnoldi method, we compare the performance of the 2nd and 4th order Magnus–Arnoldi solver with the implicit Crank–Nicolson scheme for simulations on a grid with $\Delta x = 5.5 \cdot 10^{-3}$ a.u. Figure 1 shows that the Magnus–Arnoldi schemes outperform Crank–Nicolson and that the Magnus–Arnoldi propagator is the more efficient the smaller the time step.

We now use the procedure described in Sect. 3.3 for error control. For the preliminary computations, we choose the step size $\Delta x = 0.022$ a.u. (0.012 Å) which corresponds to $N = 125$ inner discretization points. Then, we compute the PML error constant C in (2) to be $3.79 \cdot 10^5$. We choose 0.44 a.u. (12.0 eV) as a threshold value for E_2 in (4), since a majority of the Fourier coefficients has energies that are larger and the performance of the PML increases with increasing wave number. V_2 in (4) is set to the value of the second potential energy curve at the PML interface, $x = \overline{x}$, to $V_2(\overline{x}) = 0.08$ a.u. (2.2 eV). Using that the mass m for the IBr system is $m = 89379$ a.u. ($1.48 \cdot 10^{-22}$ kg) and the corresponding length scale $\tilde{x} = 2.75$ a.u. (1.46 Å) gives us the scaled dominating frequency in atomic units, $k = 92.2$, from (4). Next, we identify the inner discretization error on this mesh to be 0.012.

With these preliminary computations, we have all the parameters at hand to adjust Δx and the PML-parameters for computations with arbitrary accuracy. In Table 1,

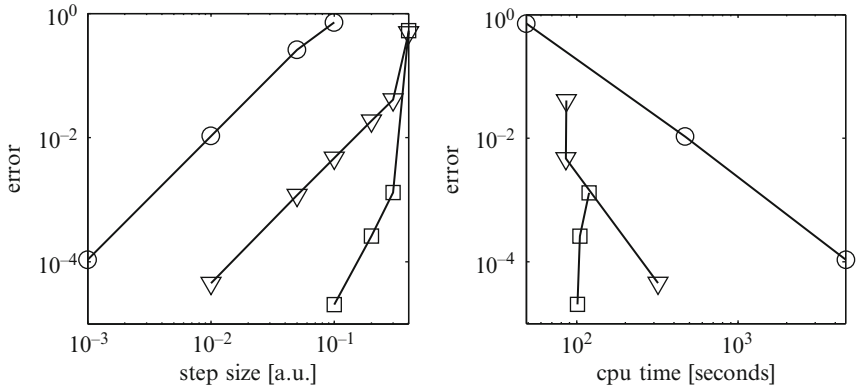


Fig. 1 Performance of Magnus–Arnoldi 4 (□), Magnus–Arnoldi 2 (▽), and Crank–Nicolson (○). The times were taken on an AMD Opteron 2216 (2.4 GHz)

Table 1 Comparison of effort and actual value of the ℓ_2 error for various tolerances

Tolerance	ℓ_2 error	N	n	M	Total no. MVP
10^{-2}	$3.9 \cdot 10^{-3}$	148	9	2,843	15,698
10^{-3}	$4.5 \cdot 10^{-4}$	196	13	6,977	29,328
10^{-4}	$7.9 \cdot 10^{-5}$	261	17	19,069	74,926

we report the results for three different tolerances. The effort for the computations depends on the number of spatial grid points, $N + n$ (inner + PML points), and the number of time steps, M, in combination with the size of the Krylov space in each step (we report the total number of matrix-vector products (MVP)). Table 1 compares the amount of work needed and shows that the total error estimation is both effective and efficient. Considering each error source separately shows that the spatial error dominantes and the other two are of similar order (overestimation by a factor 20 at most).

5 Summary and Outlook

We have suggested a procedure on how to balance the errors that arise due to boundary treatment as well as spatial and temporal discretization. We have observed that the explicit Magnus–Arnoldi method outperforms the implicit Crank–Nicolson scheme already in one dimension.

In this article, we have only considered the L_2 error and adaptivity in time, but spatial adaptivity should also be included in future work. To achieve even better performance, the error control has to be included in a parallel implementation (cf. [4]). In this way, high dimensional problems are supposed to be tackled.

References

1. Blanes,S., Casas, F., Ros, J.: Improved high order integrators based on the Magnus expansion, BIT Numer. Math. **40**, 434–450 (2000)
2. Cao,Y., Petzold, L.: A posteriori error estimate and global error control for ordinary differential equations by the adjoint method. SIAM J. Sci. Comput. **26**, 359–374 (2004)
3. Guo, H.: The effect of nonadiabatic coupling in the predissociation dynamics of IBr. J. Chem. Phys. **99**, 1685–1692 (1993)
4. Gustafsson, M.: A PDE solver framework optimized for clusters of multicore processors. Master's thesis, UPTEC Report F09 004, Uppsala University (2009)
5. Hagstrom, T.: New results on absorbing layers and radiation boundary conditions. In: M. Ainsworth, P. Davies, D. Duncan, P. Martin, B. Rynne (Eds.), Topics in Computational Wave Propagation, vol. 31 of Lecture Notes in Computational Science and Engineering, pp. 1–42. Springer, New York (2003)
6. Hochbruck, M., Lubich, C., Selhofer, H.: Exponential integrators for large systems of differential equations. SIAM J. Sci. Comput. **19**, 1552–1574 (1998)

7. Karlsson, H.O.: Accurate resonances and effective absorption of flux using smooth exterior scaling. J. Chem. Phys. **109**, 9366–9371 (1998)
8. Kormann, K., Holmgren, S., Karlsson, H.O.: Accurate time propagation for the Schrödinger equation with an explicitly time-dependent Hamiltonian. J. Chem. Phys. **128**, 184101 (2008)
9. Kormann, K., Holmgren, S., Karlsson, H.O.: Global error control of the time-propagation for the Schrödinger equation with a time-dependent Hamiltonian. Technical Report 2009-021, Uppsala University (2009)
10. Lubich, C.: From quantum to classical molecular dynamics: Reduced models and numerical analysis. Eur. Math. Soc., Zürich **9**, 147–179 (2011)
11. Nissen, A., Kreiss, G.: An optimized perfectly matched layer for the Schrödinger equation. Commun. Comput. Phys.
12. Tannor, D.J.: Introduction to Quantum Mechanics: A Time-Dependent Perspective. University Science Books, Sausalito (2007)

A Comparison of Simplicial and Block Finite Elements

Sergey Korotov and Tomáš Vejchodský

Abstract In this note we discuss and compare the performance of the finite element method (FEM) on two popular types of meshes – simplicial and block ones. A special emphasis is put on the validity of discrete maximum principles and on associated (geometric) mesh generation/refinement issues in higher dimensions. As a result, we would recommend to carefully reconsider the common belief that the simplicial finite elements are very convenient to describe complicated geometries (which appear in real-life problems), and also that the block finite elements, due to their simplicity, should be used if the geometry of the solution domain allows that.

1 Introduction

Geometrically, there are two types of finite elements (FEs) which can be naturally generalized to any dimension – simplices and blocks, where by blocks we mean Cartesian products of intervals. In what follows, we shall only consider the lowest-order finite elements, i.e., linear functions on simplices and multilinear functions on blocks. In 1D, the only reasonable element is an interval which can be understood both as a simplex and a block. Therefore, we shall make comparison for the case of two and more dimensions. Namely, we concentrate on validity of discrete maximum principles and on associated geometrical issues for mesh generation and adaptivity.

S. Korotov (✉)
Institute of Mathematics, Tampere University of Technology, P.O. Box 553,
FI–33101 Tampere, Finland
e-mail: sergey.korotov@tut.fi

T. Vejchodský
Institute of Mathematics, Academy of Sciences, Žitná 25, CZ–115 67 Prague 1, Czech Republic
e-mail: vejchod@math.cas.cz

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_57,

2 Model Problem at Its Finite Element Discretization

We consider the following test problem: Find a function u such that

$$-\Delta u + cu = f \quad \text{in } \Omega, \qquad u = g \quad \text{on } \partial\Omega, \tag{1}$$

where $\Omega \subset \mathbb{R}^d$ is a bounded polytopic domain with Lipschitz boundary $\partial\Omega$ and $c \geq 0$. The classical solution $u \in C^2(\Omega) \cap C(\overline{\Omega})$ of (1) satisfies the maximum principle:

$$f \leq 0 \qquad \Longrightarrow \qquad \max_{x\in\overline{\Omega}} u(x) \leq \max\{0\,,\, \max_{s\in\partial\Omega} g(s)\}. \tag{2}$$

Most of FE schemes are based on the weak formulation: Find $u \in H^1(\Omega)$ such that the boundary condition $u = g$ is satisfied in the sense of traces on $\partial\Omega$ and

$$a(u,v) = \mathscr{F}(v) \quad \forall v \in H^1_0(\Omega),$$

where $a(u,v) = \int_\Omega (\nabla u \cdot \nabla v + cuv)\,\mathrm{d}x$, $\mathscr{F}(v) = \int_\Omega fv\,\mathrm{d}x$, $c \in L^\infty(\Omega)$, and $f \in L^2(\Omega)$.

Let $\mathscr{T}_h$ be a conforming (block or simplicial) FE mesh on $\overline{\Omega}$ with interior nodes $B_1, \ldots, B_N$ lying in Ω and boundary nodes $B_{N+1}, \ldots, B_{N+N^\partial}$ lying on $\partial\Omega$. Further, let V_h be a finite-dimensional subspace of $H^1(\Omega)$, associated with $\mathscr{T}_h$ and its nodes, being spanned by the basis functions $\phi_1, \phi_2, \ldots, \phi_{N+N^\partial}$ with the following properties: $\phi_i \geq 0$ in $\overline{\Omega}$ (nonnegativity), $\phi_i(B_j) = \delta_{ij}$ (delta property), $i, j = 1, \ldots, N + N^\partial$, and $\sum_{i=1}^{N+N^\partial} \phi_i \equiv 1$ in $\overline{\Omega}$ (partition of unity). Notice that the lowest-order finite elements on simplices and on blocks meet these requirements. We also assume that the basis functions $\phi_1, \phi_2, \ldots, \phi_N$ vanish on the boundary $\partial\Omega$. Thus, they span a finite-dimensional subspace V_h^0 of $H^1_0(\Omega)$. Let, in addition, $g_h = \sum_{i=1}^{N^\partial} g_{N+i}\phi_{N+i} \in V_h$ be a suitable approximation of the function g, for example its nodal interpolant.

The FE approximation is a function $u_h = u_h^0 + g_h$ such that $u_h^0 \in V_h^0$ and

$$a(u_h, v_h) = \mathscr{F}(v_h) \quad \forall v_h \in V_h^0, \tag{3}$$

whose existence and uniqueness is also provided by the Lax–Milgram lemma.

Algorithmically, $u_h = \sum_{i=1}^{N+N^\partial} y_i\phi_i$, where y_i are the entries of the solution $\bar{\mathbf{y}} = [y_1, \ldots, y_{N+N^\partial}]^\top$ of the square system of $N + N^\partial$ linear algebraic equations

$$\bar{\mathbf{A}}\bar{\mathbf{y}} = \bar{\mathbf{F}}, \quad \text{where} \quad \bar{\mathbf{A}} = \begin{bmatrix} \mathbf{A} & \mathbf{A}^\partial \\ \mathbf{0} & \mathbf{I} \end{bmatrix}, \quad \text{and} \quad \bar{\mathbf{F}} = \begin{bmatrix} \mathbf{F} \\ \mathbf{F}^\partial \end{bmatrix}. \tag{4}$$

In the above, $\mathbf{A} \in \mathbb{R}^{N\times N}$, $\mathbf{A}^\partial \in \mathbb{R}^{N\times N^\partial}$, $\mathbf{0}$ and $\mathbf{I}$ stand for the zero and unit matrices of appropriate sizes. The nontrivial entries of $\bar{\mathbf{A}}$ are $a_{ij} = a(\phi_j, \phi_i)$,

$i = 1, \ldots, N,\ j = 1, \ldots, N + N^{\partial}$. The block $\mathbf{F}$ consists of entries $f_i = \mathscr{F}(\phi_i)$, $i = 1, \ldots, N$, and the block-vector $\mathbf{F}^{\partial}$ has entries $f_i^{\partial} = f_{N+i} = g_{N+i}$, $i = 1, \ldots, N^{\partial}$, given by the boundary data.

3 Discrete Maximum Principles for FEM

In this section we compare simplicial and block finite elements with respect to the so-called discrete maximum principle (DMP). For a fixed mesh $\mathscr{T}_h$, we say that the discretization (3) satisfies the DMP if

$$f \leq 0 \qquad \Longrightarrow \qquad \max_{x \in \overline{\Omega}} u_h(x) \leq \max\{0\,,\ \max_{s \in \partial\Omega} g_h(s)\}. \tag{5}$$

In the case of the lowest-order finite elements, it is well known [4] that the DMP is satisfied if (i) the stiffness matrix $\bar{\mathbf{A}}$ is monotone and if (ii) the row sums of $\bar{\mathbf{A}}$ are nonnegative. Condition (ii) is satisfied, because the basis functions form the partition of unity and the coefficient c is nonnegative. Sufficient conditions for (i) can be obtained from the theory of M-matrices [7]. This, in particular, requires the nonpositivity of the off-diagonal entries in the FE matrix $\bar{\mathbf{A}}$. Matrix $\bar{\mathbf{A}}$ is assembled from the local (element) FE matrices, $\bar{\mathbf{A}} = \sum_{K \in \mathscr{T}_h} \bar{\mathbf{A}}^K$, and hence it suffices to guarantee the nonpositivity of the off-diagonal entries of each $\bar{\mathbf{A}}^K$. This observation yields various geometric limitations for the finite elements which we discuss in what follows.

3.1 On Entries of FE Matrices for Simplices

For simplicity, let us consider the Laplace operator only, i.e., $c \equiv 0$. In this case the off-diagonal entries a_{ij}^K $(i \neq j)$ of the local stiffness matrices $\bar{\mathbf{A}}^K$ for simplicial elements can be expressed in any dimension by the following formula [1]

$$a_{ij}^K = \int_K \nabla\phi_j \cdot \nabla\phi_i \,\mathrm{d}x = -\frac{\mathrm{meas}_{d-1}(F_i)\,\mathrm{meas}_{d-1}(F_j)}{d^2\,\mathrm{meas}_d(K)} \cos\alpha_{ij}\,,$$

where $\alpha_{ij} \in (0, \pi)$ stands for the dihedral angle between the facets F_i and F_j of the simplex $K \in \mathscr{T}_h$, see Fig. 1 (left).

Clearly, $a_{ij}^K \leq 0$ if and only if $\alpha_{ij} \leq \pi/2$. This nonobtuseness condition is well known for triangles and for tetrahedra, and it is crucial for the validity of DMPs [2]. For the case of general coefficients the conditions on meshes for DMP are stricter. Thus, if e.g., $c > 0$ then all dihedral angles in meshes have to be acute and, in addition, the meshes themselves have to be sufficiently fine due to the positive terms

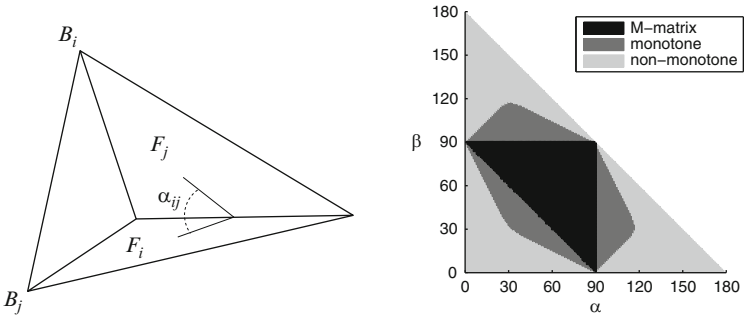


Fig. 1 The dihedral angle α_{ij} between faces F_i and F_j of a tetrahedron K (*left*). Results of the experiment for triangles (*right*)

$$\int_K \phi_j \phi_i \, \mathrm{d}x = \frac{d!}{(d+2)!} \operatorname{meas}_d(K), \quad i \neq j,$$

additionally appearing in computations, see e.g., [2, 5] for details.

Further, generalization can be obtained by requiring the stiffness matrix not to be M-matrix but to be monotone only. Theoretical handling of monotone matrices is difficult, but it can be checked numerically. Figure 1 (right) shows results of an experiment, where we consider the Poisson problem with homogeneous Dirichlet boundary conditions. Hence, the block $\mathbf{A}$ of $\bar{\mathbf{A}}$ only is relevant. The domain Ω is a triangle. The axis in Fig. 1 (right) correspond to two angles of Ω. For each pair of angles α and β, we construct a triangulation by three steps of uniform red refinement of Ω. Then we assemble the stiffness matrix $\mathbf{A}$, and color the corresponding point according to its properties. If $\mathbf{A}$ is M-matrix (has off-diagonal entries nonpositive) then the point is black. If $\mathbf{A}$ is monotone and not M-matrix then the point is dark gray. If $\mathbf{A}$ is not monotone then the point is light gray. We clearly see that in this case the stiffness matrix is M-matrix if and only if all angles are nonobtuse (black area). Further we observe that the DMP is satisfied under favorable circumstances even for angles up to 117° (dark gray area), see also [12] for a similar 3D test.

3.2 On Entries of FE Matrices for Blocks

The analysis of the DMP for block FE partitions can be done in the same fashion as for the simplices. The results, however, strongly depend on the dimension. For simplicity we again consider the Laplacian with homogeneous Dirichlet boundary condition. Let K be an element of a d-dimensional block mesh with edges of lengths $b_1, b_2, \ldots, b_d$. If B_i and B_j are its two vertices connected by the edge of length b_1 then the corresponding entry of the local stiffness matrix $\bar{\mathbf{A}}^K$ is

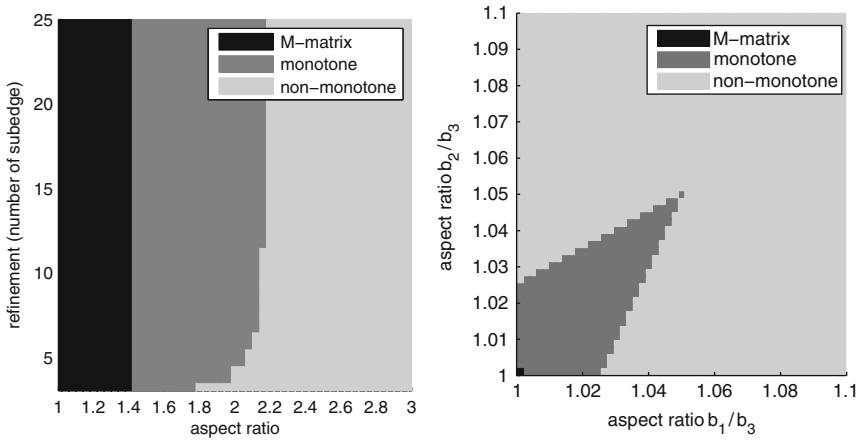


Fig. 2 The influence of the aspect ratio to the properties of the stiffness matrix $\mathbf{A}$. *Left*: Ω is a rectangle $(0, b_1) \times (0, b_2)$. *Right*: Ω is a rectangular cuboid $(0, b_1) \times (0, b_2) \times (0, b_3)$

$$a_{ij}^K = \frac{b_1 b_2 \dots b_d}{3^{d-1}} \left(\sum_{k=2}^{d} \frac{1}{2b_k^2} - \frac{1}{b_1^2} \right), \quad i \neq j. \tag{6}$$

In 2D we immediately see that $a_{ij}^K \leq 0$ if and only if $b_1/b_2 \leq \sqrt{2}$. This yields the well-known nonnarrow condition for the DMP. A rectangle K is nonnarrow if $1/\sqrt{2} \leq b_1/b_2 \leq \sqrt{2}$, where b_1 and b_2 stand for the lengths of its sides. It can be shown [9] that the DMP is satisfied if all rectangles in the mesh $\mathcal{T}_h$ are nonnarrow.

The nonnarrow condition guarantees that the corresponding stiffness matrix is M-matrix. A similar experiment as before reveals that this condition can be weakened if the stiffness matrix is required to be monotone only. In this experiment, we again consider $c \equiv 0$ and $g = 0$. The domain is a rectangle $\Omega = (0, b_1) \times (0, b_2)$. The finite element mesh is obtained by the uniform refinement of Ω into N_{sub}^2 elements, where N_{sub} is the number of subedges induced on each edge of Ω. The axes in Fig. 2 (left) correspond to the aspect ratio b_1/b_2 of the rectangle Ω (and of all elements) and to the value N_{sub}. The results in Fig. 2 (left) indicate that the value $\sqrt{2}$ in the nonnarrow condition can be increased up to about 2.16 provided the mesh is sufficiently fine.

The 3D analysis of the trilinear elements on blocks based on (6) gives a bit pessimistic conclusion. The stiffness matrix is M-matrix (and the DMP is satisfied) if all the elements are cubes [9]. Similar experiment as before, see Fig. 2 (right), indicates that the cubes cannot be distorted much in order to retain the stiffness matrix monotone and to satisfy the DMP. The two possible aspect ratios we have in rectangular cuboids can be at most around 1.05.

In dimensions 4 and higher, certain contributions form the local stiffness matrices are always positive. Indeed, without loss of generality we may assume that $b_1 \geq b_2 \geq \dots \geq b_d$. If a_{ij}^K was nonpositive then (6) would yield

$$\frac{1}{b_1^2} \geq \sum_{k=2}^{d} \frac{1}{2b_k^2} \geq \frac{d-1}{2b_2^2} > \frac{1}{b_2^2},$$

where the last inequality holds true for $d \geq 4$. This inequality, however, contradicts the fact that $b_1 \geq b_2$. Furthermore, considering the longest edge in the mesh, we see that all the contributions from all the elements surrounding this edge are positive and, hence, the corresponding off-diagonal entry in the stiffness matrix $\mathbf{A}$ is positive. Consequently, $\mathbf{A}$ is not an M-matrix. Similar experiments as before reveal that the stiffness matrix is neither monotone even on hyper-cubes. Thus, from the point of the DMP, the block finite elements are less advantageous than the simplicial elements especially for 3D and higher dimensional problems.

4 On Mesh Generation and Adaptivity

Modern FE computations require treatment of issues like generation of a mesh with desired geometric properties and its global and local refinements preserving those properties. In the following two subsections we shall discuss these issues for both, simplices and blocks, with respect to geometric limitations imposed by the DMP.

4.1 Simplicial FE Meshes (Acuteness and Nonobtuseness)

The practical realization of angle conditions (nonobtuseness and acuteness) is not easy. Even in 2D, an initial generation of reasonable nonobtuse and acute triangulations, especially for complicated domains, is algorithmically a hard task, see e.g., [3] for examples and literature on the subject. In 3D it is becoming even more difficult. Some results on generation and proper refinements of nonobtuse tetrahedral meshes are reported e.g., in [11] (see also [3]). But the only known positive (and very recent results) on acute meshes are the acute face-to-face tetrahedralization of the whole 3D Euclidean space [17], an infinite slab [6], some types of tetrahedra and a regular octahedron [10], and a cube [10, 18]. It is worth to mention that the last two works (the only relevant for real-life computations which are mostly done in bounded domains) were published in summer of 2009 only! Moreover, the acute tetrahedralization of a cube require many tetrahedra. In addition, these tetrahedra are very densely placed in the interior of the cube which is not so good for real computations as meshes used in practice should be dense mainly in vertices and along edges. Concerning higher dimensions, the situation with acute simplices is getting even more pessimistic. For example, it was shown in [10, 13] that the space $\mathbb{R}^d$ ($d \geq 4$) cannot (surprisingly!) be filled face-to-face by acute simplices at all, which means that, in general, it is not possible to generate (reasonable fine) acute

simplicial meshes for most of domains in higher dimensions, even for such simple as hypercubes.

In order to get more accurate FE approximations one needs to make various global and local refinements of the meshes preserving the desired geometric properties. For example, a triangle can be naturally split into four similar triangles using midlines (2D red refinement) (and thus acuteness or nonobtuseness are preserved), but a tetrahedron cannot be, in general, partitioned face-to-face into several similar tetrahedrons by a similar technique. After cutting four vertices of the tetrahedron off (and thus producing four similar tetrahedra), an interior octahedron remains, which can be split into four tetrahedra in three different ways. And in most of cases the resulting tetrahedra are not similar to the original one, moreover, the acuteness property cannot be preserved in any case. In addition, all further refinements should be done with a special care in order to avoid producing degenerating subtetrahedra, see [20] for details. An alternative can be one of bisection algorithms, see e.g., [15] and references therein. However, bisection cannot obviously produce all acute angles. Concerning local refinements, the only results in dimension 3 and higher are known for nonobtuse simplicial partitions, see [1].

4.2 Block FE Meshes (Preserving the Aspect Ratio)

In the case of block elements global refinement is obvious. Further, one can perform local refinements with or without hanging nodes [16]. However, local refinements without hanging nodes require forced refinements far from the targeted area and, moreover, elements with high aspect ratios are actually forming. Hanging nodes are practically more demanding to use, but they overcome these difficulties. The advantage is that the resulting meshes are nested and that the aspect ratio of subelements remains unchanged. Let us remark that the sufficient geometric conditions for the DMP are the same for meshes both with and without hanging nodes.

5 Conclusions

In 2D both triangular and rectangular meshes seem to be comparable in the sense that generation and refinement of meshes yielding the DMP is well treatable in both cases. Anyway, the triangles provide more flexibility for complicated domains (e.g., for those having non-right corners). In higher dimension, block elements can be recommended if the geometry of the domain allows them and if the DMP is not an issue. In the opposite case, the simplices should be used, but then we face the above described problems with mesh generation and local refinements constrained by the dihedral angle conditions. These problems are sometimes treatable by path-simplicial meshes [1], which guarantee the DMP at least for the Poisson problems. In addition, the practical implementation of simplicial meshes is technically more

demanding than the implementation of the blocks. This fact must be weighted as well. Let us remark that it is geometrically advantageous to use simplices and blocks together in the hybrid meshes. However, from the point of the DMP the hybrid meshes inherit the discussed disadvantages of all used types of elements. Moreover, the practical implementation of hybrid meshes is technically very demanding. For example, a 3D hybrid mesh with tetrahedra and blocks requires also right triangular prisms and pyramids to join the elements face-to-face [19]. The DMP on prismatic meshes has been analyzed in [8]. However, up to the authors' knowledge the DMP for pyramidal elements (and therefore on hybrid 3D meshes) has not been analyzed yet.

Another type of comparison of the same finite elements (but in 3D only) is done in [14].

Finally, it is interesting to mention that angle and aspect ratio conditions similar to those we discussed above also appear in the analysis of the convergence of FE approximations [5].

Acknowledgements The first author has been supported by Project no. 124619 from the Academy of Finland. The second author has been supported by Grant no. IAA100760702 of the Grant Agency of the Czech Academy of Sciences, Grant no. 102/07/0496 of the Czech Science Foundation, and by Institutional Research Plan no. AV0Z10190503 of the Czech Academy of Sciences.

References

1. Brandts, J., Korotov, S., Křížek, M.: Dissection of the path-simplex in $\mathbf{R}^n$ into n path-subsimplices. Linear Algebra Appl. **421**, 382–393 (2007)
2. Brandts, J., Korotov, S., Křížek, M.: The discrete maximum principle for linear simplicial finite element approximations of a reaction-diffusion problem. Linear Algebra Appl. **429**, 2344–2357 (2008)
3. Brandts, J., Korotov, S., Křížek, M., Šolc, J.: On nonobtuse simplicial partitions. SIAM Rev. **51**, 317–335 (2009)
4. Ciarlet, P.G.: Discrete maximum principle for finite-difference operators. Aequationes Math. **4**, 338–352 (1970)
5. Ciarlet, P.G.: The Finite Element Method for Elliptic Problems. North-Holland, Amsterdam (1978)
6. Eppstein, D., Sullivan, J.M., Üngör, A.: Tiling space and slabs with acute tetrahedra. Comput. Geom. Theor. Appl. **27**, 237–255 (2004)
7. Fiedler, M.: Special Matrices and Their Applications in Numerical Mathematics. Martinus Nijhoff Publishers, Dordrecht (1986)
8. Hannukainen, A., Korotov, S., Vejchodský, T.: Discrete maximum principle for FE-solutions of the diffusion-reaction problem on prismatic meshes. J. Comput. Appl. Math. **226**, 275–287 (2009)
9. Karátson, J., Korotov, S., Křížek, M.: On discrete maximum principles for nonlinear elliptic problems. Math. Comput. Simulation **76**, 99–108 (2007)
10. Kopczyński, E., Pak, I., Przytycki, P.: Acute triangulations of polyhedra and $\mathbf{R}^n$. arXiv: 0909.3706 (2009)
11. Korotov, S., Křížek, M.: Acute type refinements of tetrahedral partitions of polyhedral domains. SIAM J. Numer. Anal. **39**, 724–733 (2001)
12. Korotov, S., Křížek, M., Neittaanmäki, P.: Weakened acute type condition for tetrahedral triangulations and the discrete maximum principle. Math. Comp. **70**, 107–119 (2001)

13. Křížek, M.: There is no face-to-face partition of R^5 into acute simplices. Discrete Comput. Geom. **36**, 381–390 (2006)
14. Lin Qun: Tetrahedral or cubic mesh? In: Proc. Finite element methods (Jyväskylä, 2000), 160–182, GAKUTO Internat. Ser. Math. Sci. Appl., 15, Gakkotosho, Tokyo, 2001
15. Rivara M.-C.: Lepp-bisection algorithms, applications and mathematical properties. Appl. Numer. Math. **59**, 2218–2235 (2009)
16. Šolín, P., Červený, J., Doležel, I.: Arbitrary-level hanging nodes and automatic adaptivity in the hp-FEM. Math. Comput. Simulation **77**, 117–132 (2008)
17. Üngör, A.: Tiling 3D Euclidean space with acute tetrahedra. In: Proc. Canadian Conf. Comput. Geom., Waterloo, 169–172 (2001)
18. VanderZee, E., Hirani, A.N., Zharnitsky, V., Guoy, D.: A dihedral acute triangulation of the cube. arXiv:0905.3715 (2009)
19. Wieners, C.: Conforming discretizations on tetrahedrons, pyramids, prisms and hexahedrons. Univ. Stuttgart, Bericht 97/15, 1–9 (1997)
20. Zhang, S.: Successive subdivisions of tetrahedra and multigrid methods on tetrahedral meshes. Houston J. Math. **21**, 541–556 (1995)

Five-Dimensional Euclidean Space Cannot be Conformly Partitioned into Acute Simplices

Michal Křížek

Abstract We prove that a point in the Euclidean space $\mathbb{R}^5$ cannot be surrounded by a finite number of acute simplices. This fact implies that there does not exist a face-to-face partition of $\mathbb{R}^5$ into acute simplices.

1 Introduction

Acute simplicial partitions (defined in Sect. 2 below) are very useful in numerical analysis, since they yield monotone and irreducibly diagonally dominant stiffness matrices (see [4, 5]), when solving the equation

$$-\Delta u + bu = f$$

by standard linear conforming finite elements in a bounded polytopic domain in $\mathbb{R}^d$ with some boundary conditions and $b \geq 0$ small enough. In 2001, Alper Üngör [8] proved that there exists a face-to-face partition of $\mathbb{R}^3$ into acute tetrahedra (for an acute tetrahedralization of the cube see [9]). However, in [6] we showed that Üngör's result cannot be generalized into $\mathbb{R}^d$ for $d \geq 5$. Our proof resembles Fermat's method of infinite descent. In this paper we give a simpler proof for $d = 5$ which does not use the Euler-Poincaré formula as in [6]. The case $d = 4$ has not been solved, yet. Heuristics given in [1, p. 323] indicate that a four-dimensional Euclidean space probably cannot be partitioned into acute simplices either.

2 Acute Partitions

The convex hull of $d+1$ points in $\mathbb{R}^d$ for $d \in \{1, 2, 3, \dots\}$, which are not contained in a hyperplane (of dimension $d-1$), is called a *simplex* or *d-simplex*. Its $(d-1)$-dimensional faces are called *facets*. For $d > 1$ the inner angle α_{ij} between two

M. Křížek
Institute of Mathematics, Academy of Sciences, Žitná 25, CZ–115 67 Prague 1, Czech Republic
e-mail: krizek@math.cas.cz

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_58,

facets F_i and F_j for $i \neq j$ is defined by means of the scalar product of their unit outward normals n_i and n_j,

$$\cos \alpha_{ij} = -n_i \cdot n_j, \tag{1}$$

and it is called a *dihedral angle*. There are $\binom{d+1}{2}$ such angles. A simplex is said to be *acute* if all its dihedral angles are less than $\pi/2$.

Definition 1. A set of simplices is said to be a *partition* of $\mathbb{R}^d$ into simplices, if

i. The union of all these simplices is $\mathbb{R}^d$,
ii. The interiors of these simplices are mutually disjoint,
iii. Any facet of any simplex in the partition is facet of another simplex in the partition,
iv. The set of vertices of all simplices from the partition has no accumulation point in $\mathbb{R}^d$.

The condition (iii) says that all partitions are conforming, i.e., face-to-face. A partition is said to be *acute* if all its simplices are acute. We say that simplices $S_1, \ldots, S_k$ from a partition of $\mathbb{R}^d$ *surround a point* A if A is a vertex of each S_i and A lies in the interior of $\bigcup_i S_i$.

3 Auxiliary Lemmas

First we recall an elementary result for a 4-simplex.

Lemma 1. *The sum of all dihedral angles in a 4-simplex is greater than* 4π.

The proof immediately follows from a more general result [3, p. 96] that states the optimal lower and upper bounds for an arbitrary d-simplex

$$\pi \lfloor d^2/4 \rfloor < \sum_{1 \le i < j \le d+1} \alpha_{ij} < \pi d(d-1)/2, \tag{2}$$

where $d \geq 3$ and $\lfloor r \rfloor$ stands for the integer part of a real number r.

Now let F_1, F_2, and F_3 be arbitrary facets of a d-simplex S with $d \geq 3$. Since F_1 is a $(d-1)$-simplex, its inner angle φ between its $(d-2)$-dimensional faces $F_1 \cap F_2$ and $F_1 \cap F_3$ is defined similarly to (1), but in the hyperplane containing F_1.

The intersection $I = F_1 \cap F_2 \cap F_3$ has dimension $d-3$. Let L be a three-dimensional space orthogonal to I (for $d = 3$, let $L = \mathbb{R}^3$). Then $S \cap L$ is a tetrahedron. Applying the Cosine theorem from spherical trigonometry to a sufficiently small sphere centred at one of tetrahedron vertices contained in I, we get (see [2, p. 465])

$$\cos \alpha = -\cos \beta \cos \gamma + \sin \beta \sin \gamma \cos \varphi,$$

where the dihedral angles $\alpha = \angle F_2F_3$, $\beta = \angle F_1F_3$, and $\gamma = \angle F_1F_2$ are defined by (1). In Lemma 2 below we prove that the angle φ of an acute simplex is always less than the associated dihedral angle α.

Lemma 2. *Let $d \geq 3$. If a simplex is acute, then under the above notation we have*

$$\varphi < \alpha.$$

Proof. Since all dihedral angles α, β, and γ are less than $\pi/2$, we find by the above Cosine theorem that

$$\cos\varphi = \frac{\cos\alpha + \cos\beta\cos\gamma}{\sin\beta\sin\gamma} > \cos\alpha + \cos\beta\cos\gamma > \cos\alpha. \qquad \square$$

4 The Proposed Proof Technique

The proof technique used for five-dimensional space will be first illustrated on two lower-dimensional examples.

Example 1. Let A be a vertex of an acute tetrahedral partition of $\mathbb{R}^3$ (cf. [8]). Set

$$P = \bigcup_{i=1}^{t} S_i,$$

where $S_1, \ldots, S_t$ are all tetrahedra containing A. We see that P is a convex polyhedron. Denote by v, e, and t the number of vertices, edges, and triangles on the boundary ∂P, respectively. Since each edge is shared by exactly two triangular faces, we get (cf. Table below)

$$2e = 3t. \tag{3}$$

So the number of triangles on ∂P, and thus also the number of all tetrahedra sharing the vertex A is always even.

Since each S_i is acute, at least five tetrahedra will share each inner edge. Hence, each vertex from ∂P has to be surrounded by at least five edges from ∂P, i.e.,

$$5v \leq 2e. \tag{4}$$

Denote by $\alpha_1^T, \alpha_2^T, \alpha_3^T$ angles of a given triangle $T \subset \partial P$ and by $\alpha_1^V, \ldots, \alpha_{n_V}^V$ angles about vertex V. Then

$$\pi t = \sum_T \sum_{i=1}^{3} \alpha_i^T = \sum_V \sum_{j=1}^{n_V} \alpha_j^V < 2\pi v, \tag{5}$$

where the last inequality follows from Lemma 2 and the sums $\sum_T$ and $\sum_V$ are taken over all triangles T and vertices V from ∂P, respectively. Consequently, from (4), (3), and (5) we find that the number of edges and triangular faces is quite limited by the number of vertices,

$$5v \le 2e = 3t < 6v.$$

Example 2. For the time being we do not know if there exists an acute simplicial partition of $\mathbb{R}^4$. Anyway, a given point A can be surrounded by acute simplices. They can be defined, e.g., as the convex hull of the centre of the regular 600-cell (see [7]) and 600 regular tetrahedra on its three-dimensional surface.

So let

$$P = \bigcup_{i=1}^{c} S_i,$$

where $S_1, \dots, S_c$ are all 4-simplices containing the given vertex A. We see that P is a convex polytope, since it can be represented as the intersection

$$P = \bigcap_{i=1}^{c} H_i,$$

where H_i are closed half-spaces such that $S_i \subset H_i$ and ∂H_i contains that facet of S_i, which is opposite to A. Denote by v, e, t, and c the number of vertices, edges, triangles, and tetrahedra on the boundary ∂P, respectively. Since each facet is a tetrahedron, it has four triangular faces, and since each triangular face belongs to exactly two adjacent tetrahedra, we get the equality (cf. Table below)

$$2t = 4c. \tag{6}$$

Each interior triangle has to be surrounded by at least five simplices, because each S_i is acute. Hence, each edge from ∂P has to be shared by at least five triangular faces from ∂P, i.e.,

$$5e \le 3t. \tag{7}$$

Denote by $\alpha_1^C, \dots, \alpha_6^C$ all dihedral angles of a given tetrahedron C. Using the lower bound from (2), we have

$$2\pi < \sum_{i=1}^{6} \alpha_i^C.$$

Moreover, by Lemma 2 the sum of all dihedral angles $\alpha_1^E, \dots, \alpha_{n_E}^E$ of tetrahedra around a given edge E from ∂P is less than 2π. Therefore,

$$2\pi c < \sum_C \sum_{i=1}^{6} \alpha_i^C = \sum_E \sum_{j=1}^{n_E} \alpha_j^E < 2\pi e, \tag{8}$$

where the sums $\sum_C$ and $\sum_E$ are taken over all tetrahedra C and edges E from ∂P, respectively. Consequently, by (7), (6), and (8), we get

$$5e \leq 3t = 6c < 6e$$

which represents a very sharp bound on the number of triangles and tetrahedra from ∂P. In the next section we show that similar bounds for $d = 5$ lead to a contradiction.

The following table shows the simplicial equalities and acuteness inequalities for every $d \in \{2, 3, 4, 5\}$ (cf. (2), (3), (6), and (7)). The case $d = 2$ is obvious.

d	Simplicial equality	Acuteness inequality
2	$2v = 2e$	$5 \leq v$
3	$2e = 3t$	$5v \leq 2e$
4	$2t = 4c$	$5e \leq 3t$
5	$2c = 5f$	$5t \leq 4c$

5 The Nonexistence of Acute Partitions in $\mathbb{R}^5$

Theorem 1. *There is no acute partition of $\mathbb{R}^5$ into simplices.*

Proof. Assume, to the contrary, that such an acute partition exists and choose an arbitrary vertex $A \in \mathbb{R}^5$ of simplices from this partition. Set

$$P = \bigcup_{i=1}^{f} S_i,$$

where $S_1, \ldots, S_f$ are all simplices containing the given vertex A. We see again that P is a convex polytope, since it can be represented as the intersection

$$P = \bigcap_{i=1}^{f} H_i,$$

where H_i are closed half-spaces such that $S_i \subset H_i$ and ∂H_i contains that facet of S_i, which is opposite to A. Denote by v, e, t, c, and f the number of vertices, edges, triangles, tetrahedra, and facets on the boundary ∂P, respectively. Since each facet is a 4-simplex, it has five tetrahedral faces, and since each tetrahedral face belongs to exactly two adjacent facets, we get the equality (cf. Table above)

$$2c = 5f. \tag{9}$$

Each interior tetrahedron has to be surrounded by at least five 4-simplices S_i, because each S_i is acute. Hence, each triangular face from ∂P has to be shared by at least five tetrahedra from ∂P (each having four triangular faces), i.e.,

$$5t \le 4c. \tag{10}$$

Denote by $\alpha_1^F, \dots, \alpha_{10}^F$ all dihedral angles of a given 4-simplex F. Then from Lemma 1 we have

$$4\pi < \sum_{i=1}^{10} \alpha_i^F .$$

Moreover, by Lemma 2 the sum of all dihedral angles $\alpha_1^T, \dots, \alpha_{n_T}^T$ of 4-simplices around a given triangle T from ∂P is less than 2π. Therefore,

$$4\pi f < \sum_F \sum_{i=1}^{10} \alpha_i^F = \sum_T \sum_{j=1}^{n_T} \alpha_j^T < 2\pi t,$$

where the sums $\sum_F$ and $\sum_T$ are taken over all 4-simplices F and triangles T from ∂P, respectively. Consequently,

$$2f < t. \tag{11}$$

From (10), (9), and (11) we get a contradiction

$$5t \le 4c = 10f < 5t. \qquad \square$$

The above Theorem can be extented to higher dimensions $d > 5$, see [6, p. 388].

Acknowledgements The author is indebted to Jan Brandts for fruitful discussions which made the proof of Theorem simpler. The research was supported by grant No. IAA 100190803 of the Grant Agency of the Academy of Sciences of the Czech Republic.

References

1. Brandts, J., Korotov, S., Křížek, M., Šolc, J.: On nonobtuse simplicial partitions. SIAM Rev. **51**, 317–335 (2009)
2. Fiedler, M.: Über qualitative Winkeleigenschaften der Simplexe. Czechoslovak Math. J. **7**, 463–478 (1957)
3. Gaddum, J. W.: Distance sums on a sphere and angle sums in a simplex. Am. Math. Mon. **63**, 91–96 (1956)
4. Karátson, J., Korotov, S.: Discrete maximum principles for finite element solutions of nonlinear elliptic problems with mixed boundary conditions. Numer. Math. **99**, 669–698 (2005)
5. Korotov, S., Křížek, M., Šolc, J.: On a discrete maximum principle for linear FE solutions of elliptic problems with a nondiagonal coefficient matrix. Proc. of the 7th Internat. Conf. on Numerical Analysis and Applications (NAA'08), Rousse, 2008, LNCS 5434, S. Margenov et al. (eds.), Springer, Berlin, 384–391 (2009)
6. Křížek, M.: There is no face-to-face partition of R^5 into acute simplices. Discrete Comput. Geom. **36**, 381–390 (2006), Erratum 40 (2010)

7. Schläfli, L.: Theorie der vielfachen Kontinuität. Aufträge der Denkschriften-Kommission der Schweizer naturforschender Gesellschaft, Zurcher & Furre (1901). In: Gesammelte mathematische Abhandlungen, Birkhäuser, Basel (1950)
8. Üngör, A.: Tiling 3D Euclidean space with acute tetrahedra. Proc. of the 13th Canadian Conf. on Comput. Geometry, Waterloo, 169–172 (2001)
9. VanderZee, E., Hirani, A. N., Zharnitsky, V., Guoy, D.: A dihedral acute triangulation of the cube. Comput. Geom. **43**, 445–452 (2010)

The Discontinuous Galerkin Method for Convection-Diffusion Problems in Time-Dependent Domains

Václav Kučera, Miloslav Feistauer, and Jaroslava Prokopová

Abstract This paper is concerned with the numerical treatment of convection-diffusion problems in time-dependent domains. A suitable formulation of the governing equations is derived using the Arbitrary Lagrangian–Eulerian (ALE) method. The equations are then discretized in space using the discontinuous Galerkin method. The resulting space-semidiscretization scheme is numerically tested on the compressible Navier–Stokes equations describing the flow of viscous gases. The particular form of these equations allows the use of a semi-implicit time discretization, which has already been extensively studied in the case of stationary computational domains.

1 Introduction and Problem Formulation

This work deals with the solution of viscous compressible flows in domains with moving boundaries. Such problems arise in various fields of research where the interaction of fluids and solids take place. Applications range from purely technical problems (vibrations of aircraft components due to interaction with air flow, see e.g., [9]) to medical problems (air flow through human vocal folds). As our physical model we take the compressible Navier–Stokes equations in ALE form and discretize in space using the discontinuous Galerkin finite element method (DGFEM), which uses spaces of piecewise polynomial, in general discontinuous functions. The DGFEM represents a natural connection between the finite volume and finite element methods, yielding a high order method with good stability properties. For an overview of DG techniques, see e.g., [1, 2, 7, 10].

In this paper we shall consider only two-dimensional problems. Let $\Omega_t \subset \mathbb{R}^2$ be a bounded domain depending on time $t \in [0, T]$. We assume that the boundary

V. Kučera (✉), M. Feistauer, and J. Prokopová
Charles University Prague, Faculty of Mathematics and Physics, Sokolovska 83, 186 75 Praha 8
e-mail: vaclav.kucera@email.cz

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_59, © Springer-Verlag Berlin Heidelberg 2010

of Ω_t consists of three disjoint parts $\Gamma_I, \Gamma_O, \Gamma_{W_t}$: $\partial\Omega_t = \Gamma_I \cup \Gamma_O \cup \Gamma_{W_t}$, where Γ_I and Γ_O represent the time-independent inlet and outlet, respectively, and Γ_{W_t} represents moving impermeable walls.

As the governing equations we take the viscous compressible Navier–Stokes equations written in the conservative form

$$\frac{\partial \mathbf{w}}{\partial t} + \sum_{s=1}^{2} \frac{\partial \mathbf{f}_s(\mathbf{w})}{\partial x_s} = \sum_{s=1}^{2} \frac{\partial \mathbf{R}_s(\mathbf{w}, \nabla \mathbf{w})}{\partial x_s} \quad \text{in } \Omega_t,\ t \in (0, T), \tag{1}$$

where

$$\begin{aligned}
\mathbf{w} &= (\rho, \rho v_1, \rho v_2, E)^{\mathrm{T}} \in \mathbb{R}^4,\\
\mathbf{w} &= \mathbf{w}(x, t),\ x \in \Omega_t,\ t \in (0, T),\\
\mathbf{f}_i(\mathbf{w}) &= (f_{i1}, \ldots, f_{i4})^{\mathrm{T}} = (\rho v_i, \rho v_1 v_i + \delta_{1i}\, p, \rho v_2 v_i + \delta_{2i}\, p, (E + p) v_i)^{\mathrm{T}},\\
\mathbf{R}_i(\mathbf{w}, \nabla \mathbf{w}) &= (R_{i1}, \ldots, R_{i4})^{\mathrm{T}} = (0, \tau_{i1}, \tau_{i2}, \tau_{i1}\, v_1 + \tau_{i2}\, v_2 \kappa \partial\theta/\partial x_i)^{\mathrm{T}},\\
\tau_{ij} &= \lambda\, div \mathbf{v} \delta_{ij} + 2\mu\, d_{ij}(\mathbf{v}),\ d_{ij}(\mathbf{v}) = \frac{1}{2}\Big(\frac{\partial v_i}{\partial x_j} + \frac{\partial v_j}{\partial x_i}\Big).
\end{aligned}$$

We use the following notation: ρ – density, p – pressure, E – total energy, $\mathbf{v} = (v_1, \ldots, v_N)$ – velocity, θ – absolute temperature, $\gamma > 1$ – Poisson adiabatic constant, $c_v > 0$ – specific heat at constant volume, $\mu > 0, \lambda = -2\mu/3$ – viscosity coefficients, κ – heat conduction.

The above system is completed by the thermodynamical relations

$$p = (\gamma - 1)(E - \rho|\mathbf{v}|^2/2), \quad \theta = \Big(\frac{E}{\rho} - \frac{1}{2}|\mathbf{v}|^2\Big)\Big/ c_v.$$

The complete system is equipped with the initial condition

$$\mathbf{w}(x, 0) = \mathbf{w}^0(x), \quad x \in \Omega_0,$$

and the following boundary conditions:

$$\begin{aligned}
&\text{a) } \rho|_{\Gamma_I} = \rho_D, \quad \text{b) } \mathbf{v}|_{\Gamma_I} = \mathbf{v}_D = (v_{D1}, v_{D2})^{\mathrm{T}}, \quad \text{c) } \theta|_{\Gamma_I} = \theta_D,\\
&\text{a) } \mathbf{v}|_{\Gamma_{W_t}} = \mathbf{z}_D = \text{ velocity of a moving wall}, \quad \text{b) } \frac{\partial \theta}{\partial n}\Big|_{\Gamma_{W_t}} = 0,\\
&\text{a) } \sum_{i=1}^{2} \tau_{ij} n_i = 0,\ j = 1, 2, \quad \text{b) } \frac{\partial \theta}{\partial n} = 0 \text{ on } \Gamma_O.
\end{aligned}$$

2 ALE Formulation

In order to treat the time dependance of the domain, we use the so-called *arbitrary Lagrangian–Eulerian* ALE technique ([8]). We define a reference domain Ω_0 and introduce a regular one-to-one ALE mapping of Ω_0 onto Ω_t (cf. [8–10])

$$\mathcal{A}_t : \overline{\Omega}_0 \longrightarrow \overline{\Omega}_t, \ i.e. \ \mathbf{X} \in \overline{\Omega}_0 \longmapsto \mathbf{x} = \mathbf{x}(\mathbf{X}, t) = \mathcal{A}_t(\mathbf{X}) \in \overline{\Omega}_t.$$

Here we use the notation $\mathbf{X}$ for points in $\overline{\Omega}_0$ and $\mathbf{x}$ for points in $\overline{\Omega}_t$.

Further, we define the domain velocity:

$$\tilde{\mathbf{z}}(\mathbf{X}, t) = \frac{\partial}{\partial t}\mathcal{A}_t(\mathbf{X}), \quad t \in [0, T], \ \mathbf{X} \in \Omega_0,$$
$$\mathbf{z}(\mathbf{x}, t) = \tilde{\mathbf{z}}(\mathcal{A}^{-1}(\mathbf{x}), t), \quad t \in [0, T], \ \mathbf{x} \in \Omega_t$$

and the ALE derivative of a function $f = f(\mathbf{x}, t)$ defined for $\mathbf{x} \in \Omega_t$ and $t \in [0, T]$:

$$\frac{D^A}{Dt} f(\mathbf{x}, t) = \frac{\partial \tilde{f}}{\partial t}(\mathbf{X}, t), \tag{2}$$

where

$$\tilde{f}(\mathbf{X}, t) = f(\mathcal{A}_t(\mathbf{X}), t), \ \mathbf{X} \in \Omega_0, \ \mathbf{x} = \mathcal{A}_t(\mathbf{X}).$$

The following relation is a direct consequence of the chain rule:

$$\frac{D^A f}{Dt} = \frac{\partial f}{\partial t} + \operatorname{div}(\mathbf{z} f) - f \operatorname{div} \mathbf{z}.$$

This leads to our formulation of the Navier–Stokes equations in ALE form

$$\frac{D^A \mathbf{w}}{Dt} + \sum_{s=1}^{2} \frac{\partial \mathbf{g}_s(\mathbf{w})}{\partial x_s} + \mathbf{w} \operatorname{div} \mathbf{z} = \sum_{s=1}^{2} \frac{\partial \mathbf{R}_s(\mathbf{w}, \nabla \mathbf{w})}{\partial x_s}, \tag{3}$$

where $\mathbf{g}_s, s = 1, 2$, are modified inviscid fluxes

$$\mathbf{g}_s(\mathbf{w}) := \mathbf{f}_s(\mathbf{w}) - z_s \mathbf{w}.$$

3 Space Semidiscretization

In what follows we shall assume that Ω_t is a polygonal domain for all t. Let $\mathcal{T}_{ht}$ be a partition of the closure $\overline{\Omega}_t$ into a finite number of closed triangles with mutually disjoint interiors. We shall call $\mathcal{T}_{ht}$ a triangulation of Ω_t. We do not require the standard conforming properties of $\mathcal{T}_{ht}$ used in the finite element method.

This means that we admit the so-called hanging nodes. We shall use the following notation. By ∂K we denote the boundary of an element $K \in \mathcal{T}_{ht}$ and set $h_K = diam(K)$, $h = max_{K\in\mathcal{T}_{ht}} h_K$. By ρ_K we denote the radius of the largest circle inscribed into K and by $|K|$ we denote the area of K.

Let $K, K' \in \mathcal{T}_{ht}$. We say that K and K' are *neighbours*, if the set $\partial K \cap \partial K'$ has positive length. We say that $\Gamma \subset K$ is a *face* of K, if it is a maximal connected open subset either of $\partial K \cap \partial K'$, where K' is a neighbour of K, or of $\partial K \cap \partial\Omega_t$. By $\mathcal{F}_{ht}$ we denote the system of all faces of all elements $K \in \mathcal{T}_{ht}$. Further, we define the set of all inner faces by

$$\mathcal{F}_{ht}^I = \{\Gamma \in \mathcal{F}_{ht};\ \Gamma \subset \Omega_t\}$$

and the set of all boundary faces by

$$\mathcal{F}_{ht}^B = \{\Gamma \in \mathcal{F}_{ht};\ \Gamma \subset \partial\Omega_t\}\,.$$

For each $\Gamma \in \mathcal{F}_{ht}$ we define a unit normal vector $\mathbf{n}_\Gamma$. We assume that for $\Gamma \in \mathcal{F}_{ht}^B$ the normal $\mathbf{n}_\Gamma$ has the same orientation as the outer normal to $\partial\Omega$, otherwise the orientation of $\mathbf{n}_\Gamma$ is arbitrary. Finally, by $d(\Gamma)$ we denote the length of $\Gamma \in \mathcal{F}_{ht}$.

For each face $\Gamma \in \mathcal{F}_{ht}^I$ there exist two neighbours $K_\Gamma^{(L)}$, $K_\Gamma^{(R)} \in \mathcal{T}_{ht}$ such that $\Gamma \subset K_\Gamma^{(L)} \cap K_\Gamma^{(R)}$. We use the convention that $\mathbf{n}_\Gamma$ is the outer normal to the element $K_\Gamma^{(L)}$ and the inner normal to the element $K_\Gamma^{(R)}$.. Let $p \geq 1$ be an integer. The approximate solution will be sought in the space of discontinuous piecewise polynomial functions

$$\mathbf{S}_{ht} = \{v;\ v|_K \in P^p(K), \forall K \in \mathcal{T}_{ht}\}^4,$$

where $P^p(K)$ denotes the space of all polynomials on K of degree $\leq p$. For $v \in \mathbf{S}_{ht}$ and $\Gamma \in \mathcal{F}_{ht}^I$ we introduce the following notation:

$$v|_\Gamma^{(L)} = \text{ the trace of } v|_{K_\Gamma^{(L)}} \text{ on } \Gamma, \qquad v|_\Gamma^{(R)} = \text{ the trace of } v|_{K_\Gamma^{(R)}} \text{ on } \Gamma,$$
$$\langle v\rangle_\Gamma = \tfrac{1}{2}\big(v|_\Gamma^{(L)} + v|_\Gamma^{(R)}\big), \qquad [v]_\Gamma = v|_\Gamma^{(L)} - v|_\Gamma^{(R)}.$$

If $[\cdot]_\Gamma$ and $\langle\cdot\rangle_\Gamma$ appear in an integral of the form $\int_\Gamma \ldots \mathrm{d}S$, we omit the subscript Γ and write simply $[\cdot]$ and $\langle\cdot\rangle$.

4 Derivation of the Discrete Problem

In order to derive the discrete problem, we assume that $\mathbf{w}$ is a sufficiently regular solution of system (3), multiply (3) by a test function $\boldsymbol{\varphi} \in \mathbf{S}_{ht}$, integrate over any element K apply Green's theorem and sum over all $K \in \mathcal{T}_{ht}$. In this way we get

the following identity:

$$\sum_{K\in\mathcal{T}_{ht}}\int_K \frac{D^A\mathbf{w}}{Dt}\cdot\boldsymbol{\varphi}_h\,dx + b_h(\mathbf{w},\boldsymbol{\varphi}_h) + a_h(\mathbf{w},\boldsymbol{\varphi}_h) + J_h(\mathbf{w},\boldsymbol{\varphi}_h) + d_h(\mathbf{w},\boldsymbol{\varphi}_h) = \ell_h(\mathbf{w},\boldsymbol{\varphi}_h).$$

Here

$$b_h(\mathbf{w},\boldsymbol{\varphi}_h) = -\sum_{K\in\mathcal{T}_{ht}}\int_K\sum_{s=1}^{2}\mathbf{g}_s(\mathbf{w})\cdot\frac{\partial\boldsymbol{\varphi}_h}{\partial x_s}\,dx \tag{4}$$

$$+\sum_{\Gamma\in\mathcal{F}_{ht}^I}\int_\Gamma \mathbf{H}_g(\mathbf{w}^{(L)},\mathbf{w}^{(R)},\mathbf{n}_\Gamma)\cdot[\boldsymbol{\varphi}_h]\,dS + \sum_{\Gamma\in\mathcal{F}_{ht}^B}\int_\Gamma \mathbf{H}_g(\mathbf{w}^{(L)},\mathbf{w}^{(R)},\mathbf{n}_\Gamma)\cdot\boldsymbol{\varphi}_h^{(L)}dS$$

is the convection form, where $\mathbf{H}_g$ is an appropriate numerical flux, cf. Sect. 5. The state $\mathbf{w}_\Gamma^{(R)}$, for $\Gamma\subset\partial\Omega_{th}$, is determined with the aid of the Dirichlet data and the solution of a linearized local initial-boundary value Riemann problem, cf. [7].

Further, we define the viscous form

$$a_h(\mathbf{w},\boldsymbol{\varphi}) = \sum_{K\in\mathcal{T}_{ht}}\int_K\sum_{s=1}^{2}\mathbf{R}_s(\mathbf{w},\nabla\mathbf{w})\cdot\frac{\partial\boldsymbol{\varphi}}{\partial x_s}\,dx \tag{5}$$

$$-\sum_{\Gamma\in\mathcal{F}_{ht}^I}\int_\Gamma\sum_{s=1}^{2}\langle\mathbf{R}_s(\mathbf{w},\nabla\mathbf{w})\rangle(\mathbf{n}_\Gamma)_s\cdot[\boldsymbol{\varphi}]\,dS - \sum_{\Gamma\in\mathcal{F}_{ht}^D}\int_\Gamma\sum_{s=1}^{2}\mathbf{R}_s(\mathbf{w},\nabla\mathbf{w})(\mathbf{n}_\Gamma)_s\cdot\boldsymbol{\varphi}\,dS,$$

(we use the incomplete discretization of viscous terms - the so-called IIPG version), the interior and boundary penalty terms and the right-hand side form, respectively,

$$J_h(\mathbf{w},\boldsymbol{\varphi}) = \sum_{\Gamma\in\mathcal{F}_{ht}^I}\int_\Gamma\sigma[\mathbf{w}]\cdot[\boldsymbol{\varphi}]\,dS + \sum_{\Gamma\in\mathcal{F}_{ht}^D}\int_\Gamma\sigma\mathbf{w}\cdot\boldsymbol{\varphi}\,dS, \tag{6}$$

$$\ell_h(\mathbf{w},\boldsymbol{\varphi}) = \sum_{\Gamma\in\mathcal{F}_{ht}^D}\int_\Gamma\sum_{s=1}^{2}\sigma\mathbf{w}_B\cdot\boldsymbol{\varphi}\,dS. \tag{7}$$

Here $\sigma|_\Gamma = C_W\mu/d(\Gamma)$ and $C_W>0$ is a sufficiently large constant. The boundary state $\mathbf{w}_B$ is defined on the basis of the Dirichlet boundary conditions and extrapolation. The source form reads

$$d_h(\mathbf{w},\boldsymbol{\varphi}) = \sum_{K\in\mathcal{T}_{ht}}\int_K(\mathbf{w}\cdot\boldsymbol{\varphi}_h)\,\mathrm{div}\mathbf{z}\,dx. \tag{8}$$

5 Time Discretization

Our goal is to develop a numerical scheme, which would be accurate and robust, with good stability properties. One possibility is to use a implicit discretization in time, but this would lead to the solution of large nonlinear systems of algebraic equations. Therefore, we proceed similarly as in [5] and use a partial linearization of the forms b_h and a_h. This approach requires the solution of only one large sparse linear system per time level.

Let us construct a partition $0 = t_0 < t_1 < t_2 \ldots$ of the time interval $[0, T]$ and define the time step $\tau_k = t_{k+1} - t_k$. We use the approximations $\mathbf{w}_h(t_n) \approx \mathbf{w}_h^n \in \mathbf{S}_{ht_n}$, $\mathbf{z}(t_n) \approx \mathbf{z}^n$ and introduce the function $\hat{\mathbf{w}}_h^k = \mathbf{w}_h^k \circ \mathscr{A}_{t_k} \circ \mathscr{A}_{t_{k+1}}^{-1}$, which is defined in the domain $\Omega_{t_{k+1}}$. The ALE derivative at time t_{k+1} can be approximated due to (2) by the finite difference

$$\frac{D^A \mathbf{w}_h}{Dt}(\mathbf{x}, t_{k+1}) \approx \frac{\tilde{\mathbf{w}}_h^{k+1}(\mathbf{X}) - \tilde{\mathbf{w}}_h^k(\mathbf{X})}{\tau_k} = \frac{\mathbf{w}_h^{k+1}(\mathbf{x}) - \hat{\mathbf{w}}_h^k(\mathbf{x})}{\tau_k}, \quad \mathbf{x} = \mathscr{A}_{t_{k+1}}(\mathbf{X}) \in \Omega_{t_{k+1}}.$$

The linearization of the first term of the form b_h is based on the relations

$$\mathbf{g}_s(\mathbf{w}_h^{k+1}) = (\mathbf{A}_s(\mathbf{w}_h^{k+1}) - z_s^{k+1}\mathbb{I})\mathbf{w}_h^{k+1} \approx (\mathbf{A}_s(\hat{\mathbf{w}}_h^k) - z_s^{k+1}\mathbb{I})\mathbf{w}_h^{k+1},$$

where $\mathbf{A}_s(\mathbf{w})$ is the Jacobi matrix of $\mathbf{f}_s(\mathbf{w})$, cf. [6]. The second term of b_h is linearized with the aid of the Vijayasundaram numerical flux (cf. [11]):

$$\mathbf{H}_g(\mathbf{w}_{h\Gamma}^{k+1(L)}, \mathbf{w}_{h\Gamma}^{k+1(R)}, \mathbf{n}_\Gamma) \approx \mathbb{P}_g^+(\langle \hat{\mathbf{w}}_h^k \rangle_\Gamma, \mathbf{n}_\Gamma)\mathbf{w}_{h\Gamma}^{k+1(L)} + \mathbb{P}_g^-(\langle \hat{\mathbf{w}}_h^k \rangle_\Gamma, \mathbf{n}_\Gamma)\mathbf{w}_{h\Gamma}^{k+1(R)},$$

where $\mathbb{P}_g^+$ and $\mathbb{P}_g^-$ are positive and negative parts of the matrix $\mathbb{P}_g(\mathbf{w}, \mathbf{n}_\Gamma) = \sum_{s=1}^2 (\mathbf{A}_s(\mathbf{w}) - z_s^{k+1}\mathbb{I})(\mathbf{n}_\Gamma)_s$ (see [6]). In this way we get the form

$$\begin{aligned} b_h(\hat{\mathbf{w}}_h^k, \mathbf{w}_h^{k+1}, \boldsymbol{\varphi}_h) = &- \sum_{K \in \mathscr{T}_{ht_{k+1}}} \int_K \sum_{s=1}^2 (\mathbf{A}_s(\hat{\mathbf{w}}^k(x)) - z_s^{k+1}(x))\mathbb{I})\mathbf{w}^{k+1}(x)) \cdot \frac{\partial \boldsymbol{\varphi}_h(x)}{\partial x_s} dx, \\ &+ \sum_{\Gamma \in \mathscr{F}_{ht_{k+1}}^I} \int_\Gamma \left(\mathbb{P}_g^+\left(\langle \hat{\mathbf{w}}_h^k \rangle, \mathbf{n}_\Gamma\right)\mathbf{w}_h^{k+1(L)} + \mathbb{P}_g^-\left(\langle \hat{\mathbf{w}}_h^k \rangle, \mathbf{n}_\Gamma\right)\mathbf{w}_h^{k+1(R)} \right) \cdot [\boldsymbol{\varphi}_h] dS \\ &+ \sum_{\Gamma \in \mathscr{F}_{ht_{k+1}}^B} \int_\Gamma \left(\mathbb{P}_g^+\left(\langle \hat{\mathbf{w}}_h^k \rangle, \mathbf{n}_\Gamma\right)\mathbf{w}_h^{k+1(L)} + \mathbb{P}_g^-\left(\langle \hat{\mathbf{w}}_h^k \rangle, \mathbf{n}_\Gamma\right)\hat{\mathbf{w}}_h^{k(R)} \right) \cdot \boldsymbol{\varphi}_h dS. \end{aligned}$$

The linearization of the form a_h is based on the fact that $\mathbf{R}_s(\mathbf{w}_h, \nabla \mathbf{w}_h)$ is linear in $\nabla \mathbf{w}$ and nonlinear in $\mathbf{w}$. Hence,

$$a_h(\mathbf{w}_h^{k+1}, \boldsymbol{\varphi}_h) \approx \hat{a}_h(\hat{\mathbf{w}}_h^k, \mathbf{w}_h^{k+1}, \boldsymbol{\varphi}_h) := \sum_{K \in \mathscr{T}_{ht_{k+1}}} \int_K \sum_{s=1}^{2} \mathbf{R}_s(\hat{\mathbf{w}}_h^k, \nabla \mathbf{w}_h^{k+1}) \cdot \frac{\partial \boldsymbol{\varphi}_h}{\partial x_s}\, dx$$

$$- \sum_{\Gamma \in \mathscr{F}^I_{ht_{k+1}}} \int_\Gamma \sum_{s=1}^{2} \left\langle \mathbf{R}_s(\hat{\mathbf{w}}_h^k, \nabla \mathbf{w}^{k+1}) \right\rangle (\mathbf{n}_\Gamma)_s \cdot [\boldsymbol{\varphi}_h]\, dS$$

$$- \sum_{\Gamma \in \mathscr{F}^D_{ht_{k+1}}} \int_\Gamma \sum_{s=1}^{2} \mathbf{R}_s(\hat{\mathbf{w}}_h^k, \nabla \mathbf{w}_h^{k+1})(\mathbf{n}_\Gamma)_s \cdot \boldsymbol{\varphi}_h\, dS.$$

As a result we get the following semi-implicit discrete formulation:

$$\left(\frac{\mathbf{w}_h^{k+1} - \hat{\mathbf{w}}_h^k}{\tau_k}, \boldsymbol{\varphi}_h\right) + b_h(\hat{\mathbf{w}}_h^k, \mathbf{w}_h^{k+1}, \boldsymbol{\varphi}_h) + a_h(\hat{\mathbf{w}}_h^k, \mathbf{w}_h^{k+1}, \boldsymbol{\varphi}_h) \tag{9}$$

$$+ J_h(\mathbf{w}_h^{k+1}, \boldsymbol{\varphi}_h) + d_h\left(\mathbf{w}_h^{k+1}, \boldsymbol{\varphi}_h\right) = \ell(\mathbf{w}_B^k, \boldsymbol{\varphi}) \quad \forall \boldsymbol{\varphi}_h \in \mathbf{S}_{ht_{k+1}},\ k = 0, 1, \ldots.$$

This relation represents a system of linear algebraic equations on each time level which is solved either iteratively using the block-Jacobi preconditioned GMRES or a direct method (e.g., the direct unsymmetric solver UMFPACK, cf. [3]).

Several issues must be addressed for an efficient implementation. In high-speed flow with shock waves and contact discontinuities the so-called Gibbs phenomenon can appear, manifested by spurious overshoots and undershoots at discontinuities in the numerical solution. It is avoided with the aid of a discontinuity indicator and the addition of local artificial viscosity into (9), cf. [5] and [7].

Finally, integrals over elements and edges, which appear in the discrete forms, must be evaluated using sufficiently accurate quadrature rules, for example Gaussian quadrature on triangles K and faces Γ.

6 Numerical Experiment

We consider compressible flow in a channel, whose shape is inspired by the shape of human vocal folds and supraglottal spaces. The lower and upper channel walls are changing their shape according to a given function of time and axial coordinate with a given frequency in order to mimic the movement of vocal folds during speech. This movement is interpolated to the rest of the domain resulting in the ALE mapping $\mathscr{A}_t$. Figure 1 shows streamlines at different time instants $t = 504, 531, 612, 666$ during the fourth period of the motion. In the solution we can observe the formation of vortices, which are then convected through the domain.

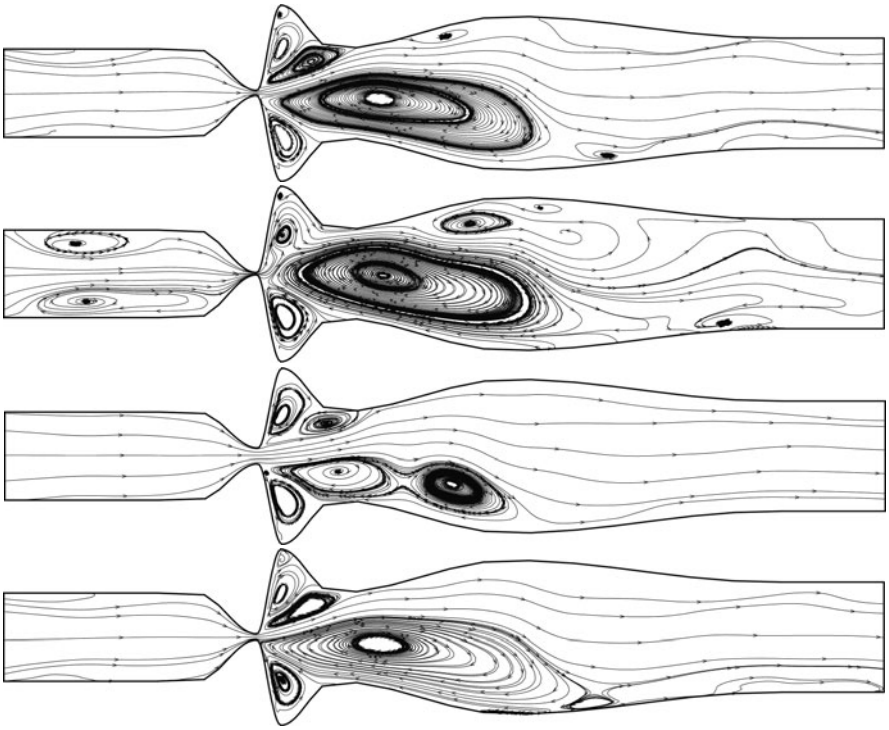


Fig. 1 Streamlines at time instants $t = 504, 531, 612, 666$

Acknowledgements This work is a part of the research project MSM 0021620839 financed by the Ministry of Education of the Czech Republic (MŠMT). The work of Václav Kučera was supported by the Nečas Center for Mathematical Modelling, project LC06052, financed by MŠMT.

References

1. Bassi, F., Rebay, S.: *High-order accurate discontinuous finite element solution of the 2D Euler equations*, J. Comput. Phys. 138 (1997) 251–285
2. Cockburn, B., Karniadakis, G. E., Shu, C.-W. (Eds.): *Discontinuous Galerkin methods*, Lecture Notes in Computational Science and Engineering 11, Springer, Berlin (2000)
3. Davis, T. A., Duff, I. S.:*A combined unifrontal/multifrontal method for unsymmetric sparse matrices*, ACM Transactions on Mathematical Software 25 (1999) 1–19
4. Dolejší, V., Feistauer, M., Schwab, C.: *On some aspects of the discontinuous Galerkin finite element method for conservation laws*, Math. Comput. Simul. 61 (2003) 333–346
5. Dolejší, V., Feistauer, M.: *A Semi-implicit discontinuous Galerkin finite element method for the numerical solution of inviscid compressible flow*, J. Comput. Phys. 198 (2004) 727–746
6. Feistauer, M., Felcman, J., Straškaraba, I.: *Mathematical and computational methods for compressible flow*, Clarendon Press, Oxford (2003)
7. Feistauer, M., Kučera, V.: *On a robust discontinuous Galerkin technique for the solution of compressible flow*, J. Comput. Phys. 224 (2007) 208–221

8. 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 (1992) 115–138
9. Sváček, P., Feistauer, M., Horáček, J.: *Numerical simulation of flow induced airfoil vibrations with large amplitudes*, J. Fluids Struct. 23 (2007) 391–411
10. van der Vegt, J. J. W., van der Ven, H.: *Space-time discontinuous Galerkin finite element method with dynamic grid motion for inviscid compressible flow*, J. Comput. Phys. 182 (2002) 546–585
11. Vijayasundaram, G.: *Transonic flow simulation using upstream centered scheme of Godunov type in finite elements*, J. Comput. Phys. 63 (1986) 416–433

A Spectral Time-Domain Method for Computational Electrodynamics

James V. Lambers

Abstract Block Krylov subspace spectral (KSS) methods have previously been applied to the variable-coefficient heat equation and wave equation, and have demonstrated high-order accuracy, as well as stability characteristic of implicit time-stepping schemes, even though KSS methods are explicit. KSS methods for scalar equations compute each Fourier coefficient of the solution using techniques developed by Gene Golub and Gérard Meurant for approximating elements of functions of matrices by Gaussian quadrature in the spectral, rather than physical, domain. We show how they can be generalized to non-self-adjoint systems of coupled equations, such as Maxwell's equations.

1 Introduction

We consider Maxwell's equation on the cube $[0, 2\pi]^3$, with periodic boundary conditions. Assuming nonconductive material with no losses, we have

$$\operatorname{div} \hat{\mathbf{E}} = 0, \quad \operatorname{div} \hat{\mathbf{H}} = 0, \tag{1}$$

$$\operatorname{curl} \hat{\mathbf{E}} = -\mu \frac{\partial \hat{\mathbf{H}}}{\partial t}, \quad \operatorname{curl} \hat{\mathbf{H}} = \varepsilon \frac{\partial \hat{\mathbf{E}}}{\partial t}, \tag{2}$$

where $\hat{\mathbf{E}}$, $\hat{\mathbf{H}}$ are the vectors of the electric and magnetic fields, and ε, μ are the electric permittivity and magnetic permeability, respectively.

By taking the curl of both sides of (2), we decouple the vector fields $\hat{\mathbf{E}}$ and $\hat{\mathbf{H}}$ and obtain the equations

J.V. Lambers
Department of Mathematics, University of Southern Mississippi, 118 College Dr #5045, Hattiesburg, MS 39406-0001
e-mail: James.Lambers@usm.edu

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_60,

$$\mu\varepsilon\frac{\partial^2\hat{\mathbf{E}}}{\partial t^2} = \Delta\hat{\mathbf{E}} + \mu^{-1}\text{curl}\,\hat{\mathbf{E}} \times \nabla\mu, \tag{3}$$

$$\mu\varepsilon\frac{\partial^2\hat{\mathbf{H}}}{\partial t^2} = \Delta\hat{\mathbf{H}} + \varepsilon^{-1}\text{curl}\,\hat{\mathbf{H}} \times \nabla\varepsilon. \tag{4}$$

In his 1966 paper [17], Yee proposed the original finite-difference time-domain method for solving the equations (1) and (2). This method uses a staggered grid to avoid solving simultaneous equations for $\hat{\mathbf{E}}$ and $\hat{\mathbf{H}}$, and also removes numerical dissipation. However, because it is an explicit finite-difference scheme, its time step is constrained by the CFL condition. In this paper, we introduce a new time-domain method for these equations.

In [10] a class of methods, called Krylov subspace spectral (KSS) methods, was introduced for the purpose of solving parabolic variable-coefficient PDE. These methods are based on techniques developed by Golub and Meurant in [5] for approximating elements of a function of a matrix by Gaussian quadrature in the *spectral* domain. In [8, 13], these methods were generalized to the second-order wave equation, for which these methods have exhibited even higher-order accuracy.

It has been shown in these references that KSS methods, by employing different approximations of the solution operator for each Fourier coefficient of the solution, achieve higher-order accuracy in time than other Krylov subspace methods (see, for example, [9]) for stiff systems of ODE, and, as shown in [13], they are also quite stable, considering that they are explicit methods. In [12,14], the accuracy and robustness of KSS methods were enhanced using block Gaussian quadrature.

Our goal is to extend the high-order accuracy achieved for the scalar wave equation to systems of coupled wave equations such as those described by Maxwell's equations. Section 2 reviews the main properties of KSS methods, including block KSS methods, as applied to the parabolic problems for which they were originally designed. Section 3 reviews their application to the wave equation, including previous convergence analysis. In Sect. 4, we discuss the modifications that must be made to block KSS methods in order to apply them to Maxwell's equations. Numerical results are presented in Sect. 5, and conclusions are stated in Sect. 6.

2 Krylov Subspace Spectral Methods

We first review KSS methods, which are easier to describe for parabolic problems. Let $S(t) = \exp[-Lt]$ represent the exact solution operator of the problem

$$u_t + Lu = 0, \quad t > 0, \tag{5}$$

with appropriate initial conditions and periodic boundary conditions. The operator L is a second-order, self-adjoint, positive definite differential operator.

Let $\langle\cdot,\cdot\rangle$ denote the standard inner product of functions defined on $[0, 2\pi]$. Krylov subspace spectral methods, introduced in [10], use Gaussian quadrature on the

spectral domain to compute the Fourier coefficients of the solution. These methods are time-stepping algorithms that compute the solution at time $t_1, t_2, \ldots$, where $t_n = n\Delta t$ for some choice of Δt. Given the computed solution $\tilde{u}(x, t_n)$ at time t_n, the solution at time t_{n+1} is computed by approximating the Fourier coefficients that would be obtained by applying the exact solution operator to $\tilde{u}(x, t_n)$,

$$\hat{u}(\omega, t_{n+1}) = \left\langle \frac{1}{\sqrt{2\pi}} e^{i\omega x}, S(\Delta t)\tilde{u}(x, t_n) \right\rangle. \tag{6}$$

In [5] Golub and Meurant describe a method for computing quantities of the form

$$\mathbf{u}^T f(A)\mathbf{v}, \tag{7}$$

where $\mathbf{u}$ and $\mathbf{v}$ are N-vectors, A is an $N \times N$ symmetric positive definite matrix, and f is a smooth function. Our goal is to apply this method with $A = L_N$ where L_N is a spectral discretization of L, $f(\lambda) = \exp(-\lambda t)$ for some t, and the vectors $\mathbf{u}$ and $\mathbf{v}$ are obtained from $\hat{\mathbf{e}}_\omega$ and $\mathbf{u}^n$, where $\hat{\mathbf{e}}_\omega$ is a discretization of $\frac{1}{\sqrt{2\pi}} e^{i\omega x}$ and $\mathbf{u}^n$ is the approximate solution at time t_n, evaluated on an N-point uniform grid.

The basic idea is as follows: since the matrix A is symmetric positive definite, it has real eigenvalues

$$b = \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N = a > 0, \tag{8}$$

and corresponding orthogonal eigenvectors $\mathbf{q}_j$, $j = 1, \ldots, N$. Therefore, the quantity (7) can be rewritten as

$$\mathbf{u}^T f(A)\mathbf{v} = \sum_{j=1}^{N} f(\lambda_j)\mathbf{u}^T \mathbf{q}_j \mathbf{q}_j^T \mathbf{v}. \tag{9}$$

which can also be viewed as a Riemann–Stieltjes integral

$$\mathbf{u}^T f(A)\mathbf{v} = I[f] = \int_a^b f(\lambda)\, d\alpha(\lambda). \tag{10}$$

As discussed in [5], the integral $I[f]$ can be approximated using Gaussian quadrature rules, which yields an approximation of the form

$$I[f] = \sum_{j=1}^{K} w_j f(\lambda_j) + R[f], \tag{11}$$

where the nodes λ_j, $j = 1, \ldots, K$, as well as the weights w_j, $j = 1, \ldots, K$, can be obtained using the symmetric Lanczos algorithm if $\mathbf{u} = \mathbf{v}$, and the unsymmetric Lanczos algorithm if $\mathbf{u} \neq \mathbf{v}$ (see [7]).

In the case $\mathbf{u} \neq \mathbf{v}$, there is a possibility that the weights may not be positive, which destabilizes the quadrature rule (see [1] for details). Instead, we consider

$$\begin{bmatrix}\mathbf{u} & \mathbf{v}\end{bmatrix}^T f(A) \begin{bmatrix}\mathbf{u} & \mathbf{v}\end{bmatrix}, \tag{12}$$

which results in the 2×2 matrix

$$\int_a^b f(\lambda)\, d\mu(\lambda) = \begin{bmatrix}\mathbf{u}^T f(A)\mathbf{u} & \mathbf{u}^T f(A)\mathbf{v} \\ \mathbf{v}^T f(A)\mathbf{u} & \mathbf{v}^T f(A)\mathbf{v}\end{bmatrix}, \tag{13}$$

where $\mu(\lambda)$ is a 2×2 matrix function of λ, each entry of which is a measure of the form $\alpha(\lambda)$ from (10).

In [5] Golub and Meurant showed how a block method can be used to generate quadrature formulas. We will describe this process here in more detail. The integral $\int_a^b f(\lambda)\, d\mu(\lambda)$ is now a 2×2 symmetric matrix and the most general K-node quadrature formula is of the form

$$\int_a^b f(\lambda)\, d\mu(\lambda) = \sum_{j=1}^{K} W_j f(T_j) W_j + error, \tag{14}$$

with T_j and W_j being symmetric 2×2 matrices. By diagonalizing each T_j, we obtain the simpler formula

$$\int_a^b f(\lambda)\, d\mu(\lambda) = \sum_{j=1}^{2K} f(\lambda_j)\mathbf{v}_j \mathbf{v}_j^T + error, \tag{15}$$

where, for each j, λ_j is a scalar and $\mathbf{v}_j$ is a 2-vector.

Each node λ_j is an eigenvalue of the matrix

$$\mathscr{T}_K = \begin{bmatrix} M_1 & B_1^T & & & \\ B_1 & M_2 & B_2^T & & \\ & \ddots & \ddots & \ddots & \\ & & B_{K-2} & M_{K-1} & B_{K-1}^T \\ & & & B_{K-1} & M_K \end{bmatrix}, \tag{16}$$

which is a block-triangular matrix of order $2K$. The vector $\mathbf{v}_j$ consists of the first two elements of the corresponding normalized eigenvector. To compute the matrices M_j and B_j, we use the block Lanczos algorithm, which was proposed by Golub and Underwood in [6].

We are now ready to describe block KSS methods. For each wave number $\omega = -N/2 + 1, \ldots, N/2$, we define

$$R_0(\omega) = \left[\hat{\mathbf{e}}_\omega \ \mathbf{u}^n \right]$$

and compute the QR factorization $R_0(\omega) = X_1(\omega) B_0(\omega)$. We then carry out block Lanczos iteration, applied to the discretized operator L_N, to obtain a block tridiagonal matrix $\mathscr{T}_K(\omega)$ of the form (16), where each entry is a function of ω. The recursion coefficents in $\mathscr{T}_K(\omega)$ can be computed efficiently by applying basic rules of symbolic calculus, including in higher spatial dimensions.

Then, we can express each Fourier coefficient of the approximate solution at the next time step as

$$[\hat{\mathbf{u}}^{n+1}]_\omega = \left[B_0^H E_{12}^H \exp[-\mathscr{T}_K(\omega)\Delta t] E_{12} B_0 \right]_{12} \tag{17}$$

where $E_{12} = \left[\mathbf{e}_1 \ \mathbf{e}_2 \right]$. The computation of (17) consists of computing the eigenvalues and eigenvectors of $\mathscr{T}_K(\omega)$ in order to obtain the nodes and weights for Gaussian quadrature, as described earlier.

This algorithm has local temporal accuracy $O(\Delta t^{2K-1})$ [14]. Furthermore, block KSS methods are more accurate than the original KSS methods described in [10,13], even though they have the same order of accuracy, because the solution $\mathbf{u}^n$ plays a greater role in the determination of the quadrature nodes. They are also more effective for problems with oscillatory or discontinuous coefficients [14].

3 Application to the Wave Equation

In this section, we review the application of Krylov subspace spectral methods to the wave equation

$$u_{tt} + Lu = 0 \quad \text{on } (0, 2\pi) \times (0, \infty), \tag{18}$$

with appropriate initial conditions, and periodic boundary conditions. A spectral representation of the operator L allows us to obtain a representation of the solution operator (the *propagator*) in terms of the sine and cosine families generated by L by a simple functional calculus. Introduce

$$R_1(t) = L^{-1/2} \sin(t\sqrt{L}) = \sum_{n=1}^{\infty} \frac{\sin(t\sqrt{\lambda_n})}{\sqrt{\lambda_n}} \langle \varphi_n^*, \cdot \rangle \varphi_n \,, \tag{19}$$

$$R_0(t) = \cos(t\sqrt{L}) = \sum_{n=1}^{\infty} \cos(t\sqrt{\lambda_n}) \langle \varphi_n^*, \cdot \rangle \varphi_n \,, \tag{20}$$

where $\lambda_1, \lambda_2, \ldots$ are the (positive) eigenvalues of L, and $\varphi_1, \varphi_2, \ldots$ are the corresponding eigenfunctions. Then the propagator of (18) can be written as

$$P(t) = \begin{bmatrix} R_0(t) & R_1(t) \\ -L\,R_1(t) & R_0(t) \end{bmatrix}. \tag{21}$$

The entries of this matrix, as functions of L, indicate which functions are the integrands in the Riemann–Stieltjes integrals used to compute the Fourier coefficients of the solution.

In [12, Theorem 6], it is shown that when the leading coefficient $p(x)$ is constant and the coefficient $q(x)$ is bandlimited, the 1-node KSS method, which has second-order local accuracy in time, is also unconditionally stable. In general, as shown in [12], the local temporal error is $O(\Delta t^{4K-2})$ when K block Gaussian nodes are used for each Fourier coefficient.

4 Application to Maxwell's Equations

In this section, we consider generalizations that must be made to block KSS methods for the wave equation in order to apply them to a non-self-adjoint system of coupled equations such as (3).

First, we consider the following initial-boundary value problem in one space dimension,

$$\frac{\partial^2 \mathbf{u}}{\partial t^2} + L\mathbf{u} = 0, \quad t > 0, \tag{22}$$

with appropriate initial conditions, and periodic boundary conditions, where $\mathbf{u} : [0, 2\pi] \times [0, \infty) \to \mathbb{R}^n$ for $n > 1$, and $L(x, D)$ is an $n \times n$ matrix where the (i, j) entry is an a differential operator $L_{ij}(x, D)$ of the form

$$L_{ij}(x, D)u(x) = \sum_{\mu=0}^{m_{ij}} a_\mu^{ij}(x) D^\mu u, \quad D = \frac{d}{dx}, \tag{23}$$

with spatially varying coefficients a_μ^{ij}, $\mu = 0, 1, \ldots, m_{ij}$.

Generalization of KSS methods to a system of the form (22) can proceed as follows. For $i, j = 1, \ldots, n$, let $\overline{L}_{ij}(D)$ be the constant-coefficient operator obtained by averaging the coefficients of $L_{ij}(x, D)$ over $[0, 2\pi]$. Then, for each wave number ω, we define $L(\omega)$ be the matrix with entries $\overline{L}_{ij}(\omega)$, i.e., the symbols of $\overline{L}_{ij}(D)$ evaluated at ω. Next, we compute the spectral decomposition of $L(\omega)$ for each ω. For $j = 1, \ldots, n$, let $\mathbf{q}_j(\omega)$ be the Schur vectors of $L(\omega)$. Then, we define our test and trial functions by $\boldsymbol{\phi}_{j,\omega}(x) = \mathbf{q}_j(\omega) \otimes e^{i\omega x}$.

For Maxwell's equations, the matrix A_N that discretizes the operator

$$A\hat{\mathbf{E}} = \frac{1}{\mu\varepsilon}\left(\Delta\hat{\mathbf{E}} + \mu^{-1}\operatorname{curl}\hat{\mathbf{E}} \times \nabla\mu\right)$$

is not symmetric, and for each coefficient of the solution, the resulting quadrature nodes λ_j, $j = 1, \ldots, 2K$, from (15) are now complex and must be obtained by a straightforward modification of block Lanczos iteration for unsymmetric matrices.

5 Numerical Results

We now apply a 2-node block KSS method to (3), with initial conditions

$$\hat{\mathbf{E}}(x, y, z, 0) = \mathbf{F}(x, y, z), \quad \frac{\partial \hat{\mathbf{E}}}{\partial t}(x, y, z, 0) = \mathbf{G}(x, y, z), \tag{24}$$

with periodic boundary conditions. The coefficients μ and ε are given by

$$\begin{aligned}\mu(x, y, z) = {} & 0.4077 + 0.0039\cos z + 0.0043\cos y - 0.0012\sin y \\ & +0.0018\cos(y + z) + 0.0027\cos(y - z) + 0.003\cos x \\ & +0.0013\cos(x - z) + 0.0012\sin(x - z) + 0.0017\cos(x + y) \\ & +0.0014\cos(x - y), \end{aligned} \tag{25}$$

$$\begin{aligned}\varepsilon(x, y, z) = {} & 0.4065 + 0.0025\cos z + 0.0042\cos y + 0.001\cos(y + z) \\ & +0.0017\cos x + 0.0011\cos(x - z) + 0.0018\cos(x + y) \\ & +0.002\cos(x - y). \end{aligned} \tag{26}$$

The components of $\mathbf{F}$ and $\mathbf{G}$ are generated in a similar fashion, except that the x- and z-components are zero.

Figure 1 demonstrates the convergence behavior using error estimates for solutions computed using $K = 2$ block quadrature nodes per coefficient in the basis described in Sect. 4. Since the exact solution is not available, the error estimate for each solution is obtained by taking the ℓ_2-norm of the relative difference between the y-component of the solution, and that of a solution computed using a smaller time step $\Delta t = 1/64$ and the maximum number of grid points.

At both spatial resolutions, the scheme exhibits approximately 6th-order accuracy in time as Δt decreases, except that for $N = 16$, the spatial error arising from truncation of Fourier series is significant enough that the overall error fails to decrease below the level achieved at $\Delta t = 1/8$. For $N = 32$, the solution is sufficiently resolved in space, and the order of overgence as $\Delta t \to 0$ is approximately 6.1.

We also note that increasing the resolution does not pose any difficulty from a stability point of view. Unlike explicit finite-difference schemes that are constrained by a CFL condition, KSS methods do not require a reduction in the time step to offset a reduction in the spatial step in order to maintain boundedness of the solution, because their domain of dependence includes the entire spatial domain for any Δt.

6 Conclusions

We have demonstrated that KSS methods can be applied to Maxwell's equations with smoothly varying coefficients. The order of temporal accuracy is the same as for the wave equation, even though Fourier coefficients are now represented by

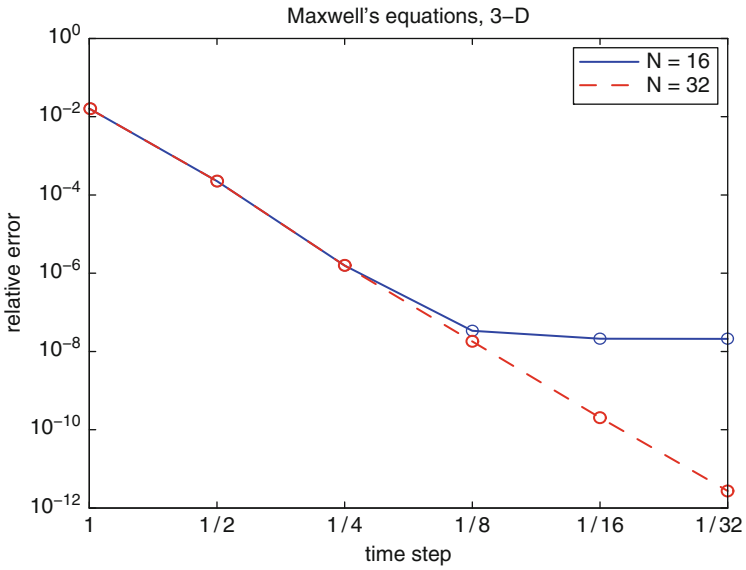


Fig. 1 Estimates of relative error in solutions of (3), (24) computed using a 2-node block KSS method on an N-point grid, with time step Δt, for various values of N and Δt

bilinear forms involving non-self-adjoint matrices, which are treated as Riemann–Stieltjes integrals over contours in the complex plane. Future work will extend the approach described in this paper to more realistic applications by using symbol modification to efficiently implement perfectly matched layers (see [2]), and various techniques (see [3, 16]) to effectively handle discontinuous coefficients.

References

1. Atkinson, K.: *An Introduction to Numerical Analysis, 2nd Ed.* Wiley, New York (1989)
2. Berenger, J.: A perfectly matched layer for the absorption of electromagnetic waves. *J. Comp. Phys.* **114** (1994) 185–200
3. Gelb, A., Tanner, J.: Robust reprojection methods for the resolution of the Gibbs phenomenon. *Appl. Comput. Harmon. Anal.* **20** (2006) 3–25
4. Golub, G. H., Gutknecht, M. H.: Modified moments for indefinite weight functions. *Numerische Mathematik* **57** (1989) 607–624s
5. Golub, G. H., Meurant, G.: Matrices, Moments and Quadrature. *Proceedings of the 15th Dundee Conference*, June–July 1993, Griffiths, D. F., Watson, G. A. (eds.), Longman Scientific & Technical (1994)
6. Golub, G. H., Underwood, R.: The block Lanczos method for computing eigenvalues. Rice J. (ed.), *Mathematical Software III*, 361–377 (1977)
7. Golub, G. H., Welsch, J.: Calculation of Gauss quadrature rules. *Math. Comp.* **23** (1969) 221–230
8. Guidotti, P., Lambers, J. V., Sølna, K.: Analysis of 1-D wave propagation in inhomogeneous media. *Numer. Funct. Anal. Optim.* **27** (2006) 25–55

9. Hochbruck, M., Lubich, C.: On Krylov subspace approximations to the matrix exponential operator. *SIAM J. Numer. Anal.* **34** (1996) 1911–1925
10. Lambers, J. V.: Krylov subspace spectral methods for variable-coefficient initial-boundary value problems. *Electron. Trans. Numer. Anal.* **20** (2005) 212–234
11. Lambers, J. V.: Practical implementation of Krylov subspace spectral methods. *J. Sci. Comput.* **32** (2007) 449–476
12. Lambers, J. V.: An explicit, stable, high-order spectral method for the wave equation based on block Gaussian quadrature. *IAENG J. Appl. Math.* **38** (2008) 333–348
13. Lambers, J. V.: Derivation of high-order spectral methods for time-dependent PDE using modified moments. *Electron. Trans. Numer. Anal.* **28** (2008) 114–135
14. Lambers, J. V.: Enhancement of Krylov subspace spectral methods by block Lanczos iteration. *Electron. Trans. Numer. Anal.* **31** (2008) 86–109
15. Lambers, J. V.: Krylov subspace spectral methods for the time-dependent Schrödinger equation with non-smooth potentials. *Numer. Algorithm.* **51** (2009) 239–280
16. Vallius, T., Honkanen, M.: Reformulation of the Fourier nodal method with adaptive spatial resolution: application to multilevel profiles. *Opt. Expr.* **10**(1) (2002) 24–34
17. Yee, K.: Numerical solution of initial boundary value problems involving Maxwell's equations in isotropic media. *Ant. Prop. IEEE Trans.* **14** (1966) 302–307

Numerical Simulation of Fluid–Structure Interaction in Human Phonation: Application

Martin Larsson and Bernhard Müller

Abstract Fluid-structure interaction in a simplified two-dimensional model of the larynx is considered in order to study human phonation. The flow is driven by an imposed pressure gradient across the glottis and interacts with the moving vocal folds in a self-sustained oscillation. The flow is computed by solving the 2D compressible Navier–Stokes equations using a high order finite difference method, which has been constructed to be strictly stable for linear hyperbolic and parabolic problems. The motion of the vocal folds is obtained by integrating the elastodynamic equations with a neo-Hookean constitutive model using a similar high order difference method as for the flow equations. Fluid and structure interact in a two-way coupling. In each time step at the fluid-structure interface, the structure provides the fluid with new no-slip boundary conditions and new grid velocities, and the fluid provides the structure with new traction boundary conditions.

1 Introduction

Fluid-structure interaction (FSI) occurs when a flexible structure interacts with a fluid. The fluid flow exerts a stress on the structure which causes it to deform, thereby generating a new geometry for the fluid flow. A direct consequence of FSI in the vocal tract is voice generation, where the motion of the soft tissue of the vocal folds interacts dynamically with the glottal airflow to produce sound. The self-sustained oscillations of the vocal folds can be explained by the Bernoulli principle which states that in the absence of gravity for inviscid incompressible steady flow, the velocity v, pressure p and density ρ are related by $p + \rho v^2/2 = \text{const}$. The vocal folds being closed in their equilibrium position, initially at rest, are forced apart by the increasing lung pressure. As the air starts flowing, the velocity in the glottis

M. Larsson (✉) and B. Müller
Norwegian University of Science and Technology (NTNU), Department of Energy and Process Engineering (EPT), 7491 Trondheim, Norway
e-mail: martin.larsson@ntnu.no, bernhard.muller@ntnu.no

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_61,

increases and thus the pressure must decrease according to the Bernoulli principle. The pressure drop together with restoring elastic forces results in a closure of the vocal folds and a build-up of pressure. This cycle then repeats itself, driven only by the lung pressure. The computational challenge in aeroelastic simulations lies in dealing with unsteady flows at high Reynolds numbers, large deformations, moving interfaces, fluid–structure interaction and intrinsically 3D motion [1].

In this paper, we employ a high order finite difference approach based on summation by parts (SBP) operators [2, 3, 14] to solve the compressible Navier–Stokes equations and the elastodynamic equations using a neo-Hookean model. Fluid and structure interact in a two-way coupling. The approach has been tested for a 2D model of the larynx and the vocal folds.

2 Governing Equations

2.1 Compressible Navier–Stokes Equations

The perturbation formulation is used to minimize cancellation errors when discretizing the Navier–Stokes equations for compressible low Mach number flow [9, 13]. The 2D compressible Navier–Stokes equations in conservative form can be expressed in perturbation form as [6, 10]

$$\mathbf{U}'_t + \mathbf{F}^{c\prime}_x + \mathbf{G}^{c\prime}_y = \mathbf{F}^{v\prime}_x + \mathbf{G}^{v\prime}_y, \tag{1}$$

where the vector $\mathbf{U}'$ denotes the perturbation of the conservative variables with respect to the stagnation values. $\mathbf{U}'$ and the inviscid (superscript c) and viscous (superscript v) flux vectors are e.g., defined in [6].

General moving geometries are treated by a time dependent coordinate transformation $\tau = t$, $\xi = \xi(t, x, y)$, $\eta = \eta(t, x, y)$. The transformed 2D conservative compressible Navier–Stokes equations in perturbation form read [6]

$$\hat{\mathbf{U}}'_\tau + \hat{\mathbf{F}}'_\xi + \hat{\mathbf{G}}'_\eta = 0, \tag{2}$$

where $\hat{\mathbf{U}}' = J^{-1}\mathbf{U}'$, $\hat{\mathbf{F}}' = J^{-1}(\xi_t \mathbf{U}' + \xi_x(\mathbf{F}^{c\prime} - \mathbf{F}^{v\prime}) + \xi_y(\mathbf{G}^{c\prime} - \mathbf{G}^{v\prime}))$ and $\hat{\mathbf{G}}' = J^{-1}(\eta_t \mathbf{U}' + \eta_x(\mathbf{F}^{c\prime} - \mathbf{F}^{v\prime}) + \eta_y(\mathbf{G}^{c\prime} - \mathbf{G}^{v\prime}))$.

No-slip adiabatic wall boundary conditions and the Navier–Stokes Characteristic Boundary Conditions (NSCBC) technique by Poinsot and Lele in [12] are employed at the outflow [7]. At the inflow, pressure, temperature and velocity in the y-direction are imposed as $p = p_{\text{atm}} + \Delta p$, $T = T_0 = 310\,\text{K}$, and $v = 0$, respectively.

2.2 Elastodynamic Equations

The governing equations for the motion of the structure are the Lagrangian field equations [11]

$$\frac{\partial S_{\alpha i}}{\partial X_\alpha} = \rho_0 \ddot{\phi}_i \tag{3}$$

where $\mathbf{X}$ are reference coordinates, $\mathbf{S}$ the nominal stress tensor, $\boldsymbol{\phi}$ the displacement vector, ρ_0 the density in the reference configuration and the dots stand for partial time derivative at fixed $\mathbf{X}$.

As a constitutive model to obtain the nominal stress as a function of the displacement, the compressible neo-Hookean law

$$\mathbf{S} = \left[\mu \mathbf{1} + \left(\frac{\lambda}{2}\ln(\det(\mathbf{C})) - \mu\right)\mathbf{C}^{-1}\right]\mathbf{F}^{\mathrm{T}} \tag{4}$$

was used cf. e.g., [18]. Here, μ and λ are the Lamé parameters, $\mathbf{F} = \mathbf{1} + \nabla_{\mathbf{X}}\boldsymbol{\phi}$ is the deformation gradient and $\mathbf{C} = \mathbf{F}^{\mathrm{T}}\mathbf{F}$ is the right Cauchy–Green strain tensor.

At boundaries where the structure is fixed in place, the displacement boundary condition $\boldsymbol{\phi}(\mathbf{X}, t) = \mathbf{0}$ is used, and on the fluid-structure interface the traction boundary condition $\mathbf{S}^{\mathrm{T}}\mathbf{N} = \mathbf{T}$ specifies the stress from the fluid on the structure boundary. If the stress tensor in the fluid is $\boldsymbol{\sigma}^{\mathrm{f}}$, then the force on the structure is [7, 8]

$$\mathbf{T} = \det(\mathbf{F})\boldsymbol{\sigma}^{\mathrm{f}}\mathbf{F}^{-\mathrm{T}}\mathbf{N}. \tag{5}$$

By introducing a coordinate transformation from the reference configuration to computational coordinates $\xi = \xi(X_\alpha), \eta = \eta(X_\alpha)$ and a variable $\boldsymbol{\psi} = \dot{\boldsymbol{\phi}}$ for the velocity, the Lagrangian field equations can be expressed as a first-order system in time on an equidistant Cartesian grid

$$\begin{cases} \dot{\boldsymbol{\psi}} = \frac{1}{J^{-1}\rho_0}\left[(\widehat{\mathbf{S}_1})_\xi + (\widehat{\mathbf{S}_2})_\eta\right] \\ \dot{\boldsymbol{\phi}} = \boldsymbol{\psi} \end{cases} \tag{6}$$

where ρ_0 is the material density in the reference configuration, $J^{-1} = |\partial(\xi, \eta)/\partial(X, Y)|$ is the Jacobian determinant of the coordinate transformation and $\widehat{\mathbf{S}_i} = J^{-1}\mathbf{S}^{\mathrm{T}}\nabla\xi_i$, $i = 1, 2$, are transformed flux vectors.

The traction boundary condition specifies the momentum flux over the fluid–structure boundary. It can be shown, cf. [7, 8], that

$$\widehat{\mathbf{S}_2} = J^{-1}\boldsymbol{\sigma}^{\mathrm{f}}\begin{pmatrix} F_{22} & -F_{21} \\ -F_{12} & F_{11} \end{pmatrix}\nabla\eta, \tag{7}$$

where $F_{i\alpha}$ are components of the deformation gradient $\mathbf{F}$, when the interface is at a line of constant η.

3 Fluid–Structure Interaction

3.1 Arbitrary Lagrangian–Eulerian (ALE) Formulation

The displacement of the structure interface determines the shape of the fluid domain and the structure velocity at the interface determines the internal grid point velocities in the fluid domain. The right and left boundaries of the fluid domain are the out- and inflow, respectively. The top and bottom parts of the fluid domain are bounded by the flexible vocal folds and the inner wall of the airpipe which is assumed to be rigid. As we do not assume symmetry, the motions of the two vocal folds are solved for individually. In our arbitrary Lagrangian–Eulerian (ALE) formulation, the positions and velocities of the grid points in the fluid domain are linearly interpolated from the positions and velocities of the structures at the interfaces. Figure 1 shows the given structure veolcities with bold arrows and the interpolated grid point velocities $\dot{x}, \dot{y}$ (thin arrows) for three grid lines.

To obtain the time derivative of J^{-1} as needed in (2), a geometric invariant [15] is used. This geometric conservation law states that $(J^{-1})_\tau + (J^{-1}\xi_t)_\xi + (J^{-1}\eta_t)_\eta = 0$. The time derivatives of the computational coordinates ξ, η can here be obtained from the grid point velocities $\dot{x}, \dot{y}$ as $\xi_t = -(\dot{x}\xi_x + \dot{y}\xi_y)$, $\eta_t = -(\dot{x}\eta_x + \dot{y}\eta_y)$ which can be seen by differentiating the transformation with respect to τ.

3.2 Description of Fluid–Structure Interaction Algorithm

First, we construct the fixed reference configuration for the structure and set the initial displacements and velocities to zero. The initial fluid domain is then uniquely

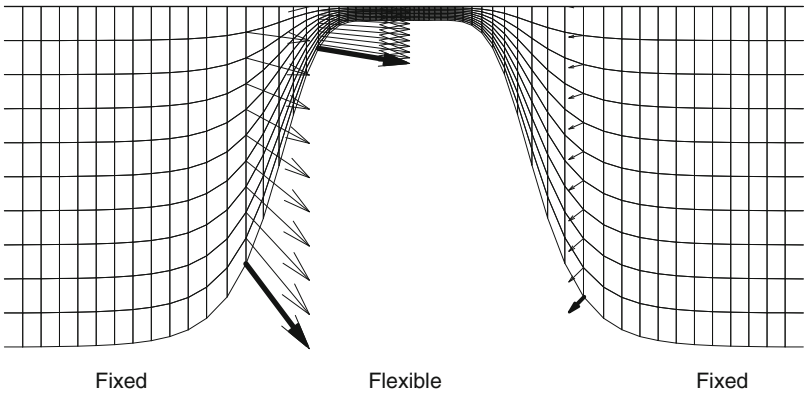


Fig. 1 The boundary of the fluid domain consists of fixed and flexible parts. The velocity at the boundary of the flexible part determines the internal grid point velocity. Only half domain shown

determined by the reference boundary of the structure. We then take one time step for the fluid with imposed pressure boundary conditions at the inflow, zero initial conditions for the perturbation variables $\mathbf{U}'$ and adiabatic no-slip conditions, i.e., $\mathbf{u} = 0$ and $\partial T/\partial n = 0$ on the wall. After the first fluid time step, the viscous fluid stress on the wall is calculated based on the new fluid velocities and pressures. These fluid stresses are passed on to the structure solver via the traction boundary condition (7). Using these boundary conditions, one time step is taken for the structure. The solution for the structure gives the velocities and displacements on the boundary, which in turn are used to generate the new fluid mesh and internal grid point velocities. This procedure is then repeated for each time step.

4 High Order Finite Difference Method

The summation by parts (SBP) operator Q is an approximation to the first spatial derivative. In the interior, Q corresponds to the standard 6th order explicit central operator, while Q is third order accurate at and near the boundaries. Through a special boundary treatment, SBP operators allow energy estimates for the discrete problems similar to the ones for the continuous problems that are approximated. Thus, SBP operators yield strictly stable schemes for general boundary conditions [7, 14]. The global order of accuracy of the present SBP operator Q is 4 [2, 3]. Second derivatives are approximated by applying the SBP operator Q twice. The classical fourth order explicit Runge–Kutta method is used for time integration. Spurious high wave number oscillations are suppressed by a sixth order explicit filter [10].

5 Results

The initial geometry for the vocal folds is here based on the geometry used in [17] for an oscillating glottis with a given time dependence. The initial shape of the vocal tract including the vocal fold is given as

$$r_w(x) = \frac{D_0 - D_{\min}}{4} \tanh s + \frac{D_0 + D_{\min}}{4}, \tag{8}$$

where r_w is the is the half height of the vocal tract, $D_0 = 5D_g$ is the height of the channel, $D_g = 4\,\text{mm}$ is the average glottis height, $D_{\min} = 2\,\text{mm}$ is the minimum glottis height, $s = b|x|/D_g - bD_g/|x|$, $c = 0.42$ and $b = 1.4$. For $-2D_g \le x \le 2D_g$, the function (8) describes the curved parts of the reference configuration for the top and bottom (with a minus sign) vocal folds.

5.1 Vocal Fold Material Parameters

The density in the reference configuraiton is $\rho_0 = 1,043\,\mathrm{kg\,m^{-3}}$, corresponding to the measured density of vocal fold tissue as reported by [4]. The Poisson ratio was chosen as $\nu = 0.47$ for the whole tissue, corresponding to a nearly incompressible material with $\nu = 0.5$ being the theoretical incompressible limit. A two-layer model for the vocal folds was used so that the shear modulus varied smoothly from $\mu_c = 3.5\,\mathrm{kPa}$ in the cover to $\mu_l = 4.4\,\mathrm{kPa}$ in the ligament. The Lamé parameter λ, as a function of space, was then obtained as $\lambda = 2\mu\nu/(1-2\nu)$. A compressible neo-Hookean material model was used, cf. [7].

5.2 Fluid Model

We used a Reynolds number of 3,000 based on the average glottis height $D_g = 0.004\,\mathrm{m}$ and an assumed average velocity in the glottis of $U_m = 40\,\mathrm{m\,s^{-1}}$. We used these particular values in order to be able to compare with previously published results by Zhao et al. [16, 17] and by ourselves [5, 6]. The Prandtl number was set to 1.0, and the Mach number was 0.2, based on the assumed average velocity and the speed of sound. We deliberately used a lower value for the speed of sound, $c_0 = 200\,\mathrm{m\,s^{-1}}$ in order to speed up the computations. The air density was $1.3\,\mathrm{kg\,m^{-3}}$ and the atmospheric pressure was $p_{\mathrm{atm}} = 101,325\,\mathrm{Pa}$. The equation of state was the perfect gas law, and we assumed a Newtonian fluid. At the inlet, we imposed a typical lung pressure during phonation with a small unsymmetric perturbance by setting the acoustic pressure to $p_{\mathrm{acoustic}} = p - p_{\mathrm{atm}} = (1 + 0.025\sin 2\pi\eta)2,736\,\mathrm{Pa}$, where $\eta = 0$ at the lower edge and $\eta = 1$ at the upper edge of the inflow boundary. The outlet pressure was set to atmospheric pressure, i.e., $p - p_{\mathrm{atm}} = 0\,\mathrm{Pa}$.

5.3 Numerical Simulation

Both fluid and structure used the same set of variables for nondimensionalization and the same time step was used for both fields so that the two solutions are always at the same time level. The structure grid consisted of 81×61 points for each vocal fold, i.e., for the upper and the lower vocal folds, and the fluid domain was 241×61 points. The time step was determined by the stability condition for the fluid, which was satisfied here by requiring $CFL \leq 1$. Since the fluid domain changes with time, the CFL condition puts a stricter constraint on the time step when the glottis is nearly closed. The solution was marched in time with given initial and boundary conditions to dimensional time $t = 20\,\mathrm{ms}$ or 416,948 time steps, implying an average dimensional time step of $\Delta t = 48\,\mathrm{ns}$.

Figure 2 shows results for the vorticity at certain time instants. Initially, the flow is symmetric with two start-up vortices followed by an elongated vortical structure

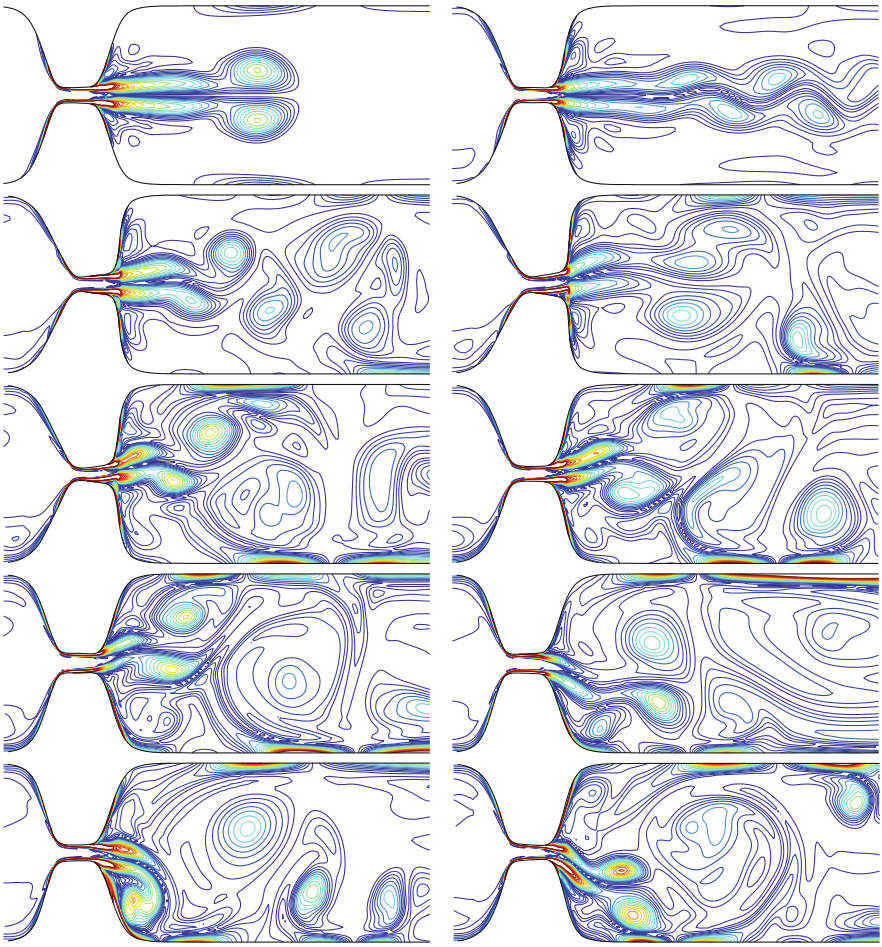


Fig. 2 Vorticity contours at 2 ms intervals. *Top left* subplot is the vorticity at $t = 2$ ms, *top right* is at $t = 4$ ms and so on up to $t = 20$ ms (*lower right*). There are 20 equally spaced contour lines between $\omega = 0\,\mathrm{s}^{-1}$ and $\omega = 5 \times 10^4\,\mathrm{s}^{-1}$ in each subplot

on each side of the centerline. After the start-up vortices leave the domain, the elongated structure becomes unstable and breaks up into smaller circular vortices. The observed frequency of the oscillation is about 80 Hz, which is close to the typical frequencies that occur in phonation, i.e., 100 Hz for men and 200 Hz for women.

6 Conclusions

Our 2D model for the vocal folds and the air flow in the vocal tract proves to be able to capture the self-sustained pressure-driven oscillations and vortex generation in the glottis. The simulated frequency of 80 Hz is close to 100 Hz, typical for men.

Acknowledgements The authors thank Bjørn Skallerud, Paul Leinan and Victorien Prot at the Department of Structural Engineering, NTNU for valuable discussions on the structure model and for Abaqus support. The current research has been funded by the Swedish Research Council under the project "Numerical Simulation of Respiratory Flow."

References

1. J.B. Grotberg and O.E. Jensen. Biofluid mechanics in flexible tubes. *Annu. Rev. Fluid Mech.*, 36:121–147, 2004
2. B. Gustafsson. *High order difference methods for time-dependent PDE*. Springer, Berlin, 2008
3. B. Gustafsson, H.-O. Kreiss, and J. Oliger. *Time dependent problems and difference methods*. Wiley, New York, 1995
4. E.J. Hunter, I.R. Titze, and F. Alipour. A three-dimensional model of vocal fold abduction/adduction. *J. Acoust. Soc. Am.*, 115(4):1747–1759, 2004
5. M. Larsson. *Numerical Simulation of Human Phonation*, Master Thesis, Uppsala University, Department of Information Technology, 2007
6. M. Larsson and B. Müller. Numerical simulation of confined pulsating jets in human phonation. *Comput. Fluid*, 38:1375–1383, 2009
7. M. Larsson and B. Müller. Numerical simulation of fluid-structure interaction in human phonation. In B. Skallerud and H.I. Andersson, editors, *MekIT 09 Fifth National Conference on Computational Mechanics*, pages 261–280, Tapir, Trondheim, 2009
8. M. Larsson and B. Müller. Numerical simulation of fluid–structure interaction in human phonation: Verification of structure part. In *Proceedings of ICOSAHOM 09 International Conference on Spectral and High Order Methods*, Trondheim, Norway, 2009
9. B. Müller. *Computation of compressible low Mach number flow*, Habilitation Thesis, ETH Zürich, 1996
10. B. Müller. High order numerical simulation of aeolian tones. *Comput. Fluids*, 37(4):450–462, 2008
11. R.W. Ogden. *Non-linear elastic deformations*. Ellis Horwood, Chichester, 1984
12. T.J. Poinsot and S.K. Lele. Boundary conditions for direct simulations of compressible viscous flows. *J. Comput. Phys.*, 101:104–129, 1992
13. J. Sesterhenn, B. Müller, and H. Thomann. On the cancellation problem in calculating compressible low Mach number flows. *J. Comput. Phys.*, 151:597–615, 1999
14. B. Strand. Summation by parts for finite difference approximations for d/dx. *J. Comput. Phys.*, 110:47–67, 1994
15. M.R. Visbal and D.V. Gaitonde. On the use of higher-order finite-difference schemes on curvilinear and deforming meshes. *J. Comput. Phys.*, 181:155–185, 2002
16. C. Zhang, W. Zhao, S.H. Frankel, and L. Mongeau. Computational aeroacoustics of phonation, Part II. *J. Acoust. Soc. Am.*, 112(5):2147–2154, 2002
17. W. Zhao, S.H. Frankel, and L. Mongeau. Computational aeroacoustics of phonation, Part I: Computational methods and sound generation mechanisms. *J. Acoust. Soc. Am.*, 112(5):2134–2146, 2002
18. O.C. Zienkiewicz and R.L. Taylor. *The finite element method for solid and structural mechanics, 5th edition*. Elsevier, Amsterdam, 2000

Error Estimation and Anisotropic Mesh Refinement for Aerodynamic Flow Simulations

Tobias Leicht and Ralf Hartmann

Abstract Aerodynamic flow fields are dominated by anisotropic features at both boundary layers and shocks. Solution-adaptive local mesh refinement can be improved considerably by respecting those anisotropic features. Two types of anisotropy indicators are presented. Whereas the first one is based on polynomial approximation properties and needs the evaluation of second and higher order derivatives of the solution the second one exploits the inter-element jumps arising in discontinuous Galerkin methods and can easily be used with higher order discretizations and even hp-refinement. Examples for sub-, trans- and supersonic flows combining these anisotropic indicators with reliable residual or adjoint based error estimation techniques demonstrate the potential and limitations of this approach.

1 Introduction

We consider steady-state laminar viscous flows governed by the compressible Navier–Stokes equations. In the inviscid limit these degenerate to the compressible Euler equations which are in many cases appropriate to describe flow fields in the presence of shocks.

Our discretization is based on a discontinuous Galerkin (DG) method with the symmetric interior penalty (SIPG) approach for viscous terms,[1] see [6, 14] and the references cited therein. The DG approach is a natural extension of the finite volume method predominantly used in aerodynamics to higher order and offers a great flexibility of the underlying meshes concerning both local mesh adaption with hanging

[1] However, results similar to those presented here have also been obtained with the second scheme of Bassi and Rebay [2].

T. Leicht (✉) and R. Hartmann
DLR (German Aerospace Center), Institute of Aerodynamics and Flow Technology, Center for Computer Applications in Aerospace Science and Engineering, 38108 Braunschweig, Germany
e-mail: tobias.leicht@dlr.de, ralf.hartmann@dlr.de

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_62,

nodes and variable order discretizations, combined in hp-algorithms. Being a finite element method a substantial error estimation framework is available.

In the following we will consider constant polynomial degree approximations on quadrilateral and hexahedral meshes with local mesh refinement. Such a refinement is often performed in an isotropic way by splitting all an element's edges and forming new children elements. However, flow phenomena may exhibit a strong directional behavior in boundary layers or interior layers like shocks. Highly stretched elements should be used for an efficient resolution of these features. Starting from a coarse initial mesh, such elements can be obtained by an anisotropic refinement which splits only some of an element's edges.

Considerable work has been devoted to anisotropic refinement for linear finite elements on simplex meshes where the information of an approximated Hessian-based mesh metric field is used within re-meshing algorithms, see [4, 5, 10, 16] for example. Here, the metric field approximates the interpolation error of the solution and is used to determine the local mesh density as well as the local element rotation and stretching in a re-meshing algorithm.

As opposed to a priori interpolation error estimates, a posteriori estimates based on an adjoint problem take into account error transportation and accumulation effects. Using these *goal-oriented* indicators to determine the local mesh density results in meshes which are specifically tailored to the accurate approximation of a target quantity like the aerodynamic lift or drag force. In [18] the directional information of the metric approach has been combined with a scaling based on adjoint-based error indicators, resulting in dual weighted metrics.

Another approach to anisotropy detection in the context of element subdivision is to use several trial refinements and selecting the case which reduces the error most effectively, see [13, 17]. However, such approaches seem unreasonably expensive, especially if they require solutions on globally refined meshes. Solving only local problems and including goal-oriented refinement has been considered in [9].

The purpose of this work is to employ anisotropy indicators which come computationally almost for free, i.e. no auxiliary problems shall be solved for obtaining anisotropic refinement information. Furthermore, these indicators shall be applicable to higher order DG discretizations and they shall be easily combined with different reliable error indicators.

We adopt the partitioned approach of using different indicators for selecting the elements to be refined and for choosing the anisotropic refinement case. In particular, we employ residual-based and adjoint-based indicators for goal-oriented refinement to select a certain fraction of all elements to be refined. In a second step, the discrete solution is analyzed using one of two different anisotropic indicators to decide upon a possibly anisotropic subdivision case. We note, that the presented ideas are only applicable to meshes with tensor-product elements (quadrilaterals in 2D, hexahedra in 3D), whereas other work based on metrics is often only applicable to simplicial meshes.

2 Error Estimation and Error Indicators

Adjoint-based Indicators Given a target functional $J(\mathbf{u})$, the error of the discrete solution $\mathbf{u}_h$ compared to the analytical solution $\mathbf{u}$ in terms of the target quantity can be approximated employing a linearized adjoint problem. This corresponds to the well known dual weighted residual (DWR) method by Becker and Rannacher [3], see [14] for an application in the current context. In the DG context the residual and thus this estimate can be decomposed into element-wise contributions which serve as local error indicators.

The total error estimate, i.e., the accumulated local contributions, can be used as a reliable estimate of the discretization error. As this estimates includes the sign of the error it can even be used to improve the computed target quantity value.

Residual-based Indicators Assuming that the solution of the adjoint problem is sufficiently smooth, an upper bound of the error in the target quantity can be derived, cf. [6]. The localized form of this estimate serves as residual-based error indicator. As the specific target quantity does not enter the definition of this indicator, it can be used to resolve all flow features but will in general be less efficient for any given target quantity.

3 Anisotropy Indicators

Jump Indicator Assuming that the analytical solution is continuous the presence of discontinuities in the discrete solution indicates local errors. We associate large jumps of the solution over element interfaces with large approximation errors orthogonal to the corresponding face. If the average jump K_i

$$K_i = \frac{\sum_j \left| \int_{f_i^j} (\phi^+ - \phi^-) \, \mathrm{d}x \right|}{\sum_j \int_{f_i^j} 1 \, \mathrm{d}x}, \quad i = 1, 2, 3, \tag{1}$$

over the two faces f_i^1 and f_i^2 orthogonal to the direction i on the tensor-product reference element is small compared to the maximal value encountered in any direction on the same element we do not refine the element along that direction. Here, $^+$ and $^-$ denote the traces of the function ϕ taken from within the current element and its neighbor, respectively.

If the analytical solution exhibits a discontinuity, e.g., at a shock, close to and almost parallel to a particular face, this indicator will also detect a large jump and refine the element parallel to that face. This is what is required to obtain an improved location of the discontinuity in the numerical solution, thus this behavior is desirable. The probability that the discontinuity exactly coincides with the face is vanishing for real applications.

Derivative Indicator After a transformation to the reference element to include scaling effects the local interpolation or projection error of a polynomial approximation of degree p to a sufficiently smooth function $\phi \in H^{p+1}$ is determined by the $(p+1)^{\text{th}}$ derivative tensor.

In the general case we compare the projected derivative along the coordinate axes of the reference element and do not refine the element along directions which feature a small derivative compared to other values on that element. For second order methods with $p = 1$ we evaluate the eigenvalues and eigenvectors of the derivative tensor (Hessian) and exclude directions from refinement if they are aligned with the eigenvectors of small eigenvalues.

The underlying smoothness assumption is especially critical for higher order methods and discontinuous solutions at shocks.

Application to Vector Valued Solution Functions Several strategies for extending the presented indicators from scalar valued solution functions ϕ to vector valued solution functions $\mathbf{u}$ have been investigated, see [14] for a comprehensive discussion.

For the jump indicator, we simply replace the jump of the scalar function ϕ in (1) by the jump of the l^2-norm of the vector-valued function $\mathbf{u}$. For the derivative indicator we differentiate between flow regimes: In inviscid cases we simply use the Mach number as a representative scalar variable, whereas in viscous cases the refinement indicator is evaluated separately for each component. We then default to isotropic refinement but select an anisotropic subdivision case if that is suggested by a sufficient number of individual indicators and if there is no contradiction in the predicted direction of anisotropy.

4 Numerical Examples

The basic performance of the proposed indicators will be analyzed using some two-dimensional computations. After that, a three-dimensional example will demonstrate the applicability to flows of increased complexity.

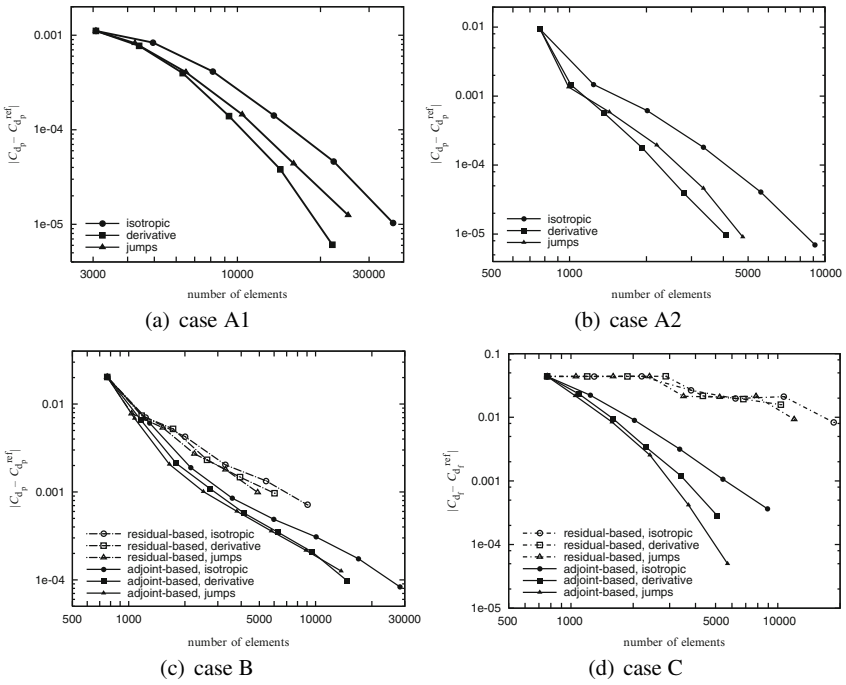
NACA0012 Airfoil Sub-, trans- and supersonic flows around the NACA0012 airfoil according to the flow conditions in Table 1 have been computed on sequences of refined meshes using both adjoint-based and residual-based error indicators as well as isotropic and anisotropic refinement.

Figure 1 plots the error in a selected target functional vs. the number of elements for the different refinement strategies. All reference values have been obtained by fine grid computations. The subsonic case A uses adjoint-based error estimation. Comparing the second order solution of case A1 with the third order solution of case A2 shows the increased accuracy of the method, as a significantly reduced number of elements produces results of the same accuracy in the higher order case. Apart from that the behavior is similar – the jump-based anisotropy indicator significantly reduces the number of elements required for a given accuracy and the

Table 1 Freestream conditions for NACA0012 test cases

Case	Mach number M	Angle of attack α	Reynolds number Re	Polynomial degree p	Target[a] $J(u)$
A1	0.5	0°	5,000	1	C_{d_p}
A2	0.5	0°	5,000	2	C_{d_p}
B	0.8	1.25°	*inviscid*	1	C_{d_p}
C	1.2	0°	1,000	1	C_{d_f}

[a] C_{d_p} and C_{d_f} denote the pressure and friction part of the drag coefficient $C_d = C_{d_p} + C_{d_f}$, respectively



(a) case A1 (b) case A2

(c) case B (d) case C

Fig. 1 Convergence of the error under different mesh refinement algorithms for the NACA0012 test cases

derivative-based indicator performs even better. As the solution is quite smooth in this subsonic case this can be expected.

Similar results can be seen for the transonic case B. Here the performance of the jump indicator is similar or even superior to the derivative indicator. This is probably due to the reduced smoothness of the solution at the shock which contradicts the assumptions of the derivative indicator. This effect is even stronger in the supersonic case C. In general, the simple jump-based criterion performs remarkably well in all flow regimes.

In transonic cases shocks are usually located in the vicinity of the airfoil and are of great importance for the computed aerodynamic forces. The residual-based

indicator resolves these prominent features and thus performs only slightly inferior to the goal-oriented error estimation, see Fig. 1c.

In contrast to that, the supersonic case C features a prominent detached bow shock in front of the airfoil. As the residual-based error estimation initially resolves mainly this feature whereas the boundary layer resolution is improved only later on we notice almost no reduction of the error. Goal-oriented refinement, however, yields significant error reductions already on the first adapted mesh, see Fig. 1d. This motivates the utilization of the adjoint problem in spite of its additional cost.

Laminar Delta Wing As a second more complex example we consider a laminar flow at Mach number $M = 0.3$, Reynolds number $Re = 4{,}000$ and an angle of attack $\alpha = 12.5°$ around a delta wing with sharp leading edge and a blunt trailing edge. Figure 2 illustrates the vortex dominated flow characteristics of this test case which has been considered in the EU project ADIGMA [12] and in [7], a similar case was treated earlier in [11].

In the following we consider the error of different approximations of the lift coefficient C_l. Similar results have been obtained for the drag coefficient C_d. We start by computing the lift from the second order flow solution on a sequence of globally refined meshes starting from a very coarse initial 3,264 elemental mesh for the half domain with symmetry boundary conditions. We then consider adaptive local mesh refinement starting from the results on the initial coarse mesh.

Figure 3 plots the error in the lift coefficient vs. the number of elements for various refinement strategies. Compared to global mesh refinement, lift coefficients of a

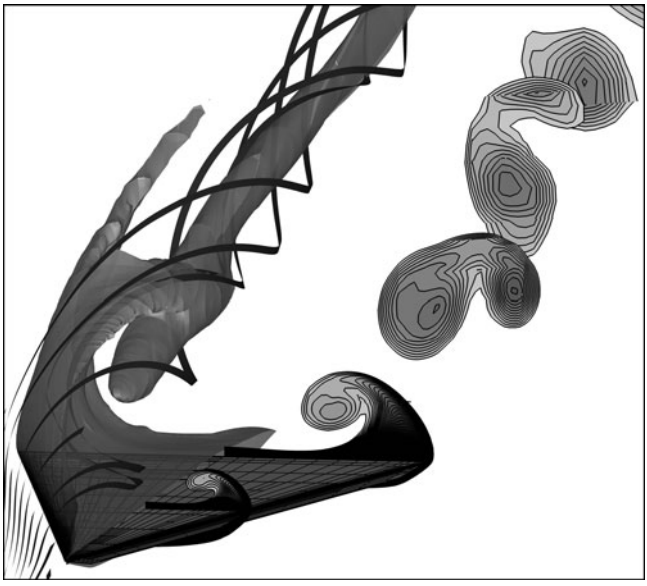


Fig. 2 Solution plot showing streamlines and a Mach number iso-surface over the left half of the delta wing immersed in a laminar flow at high angle of attack as well as Mach number slices over the right half

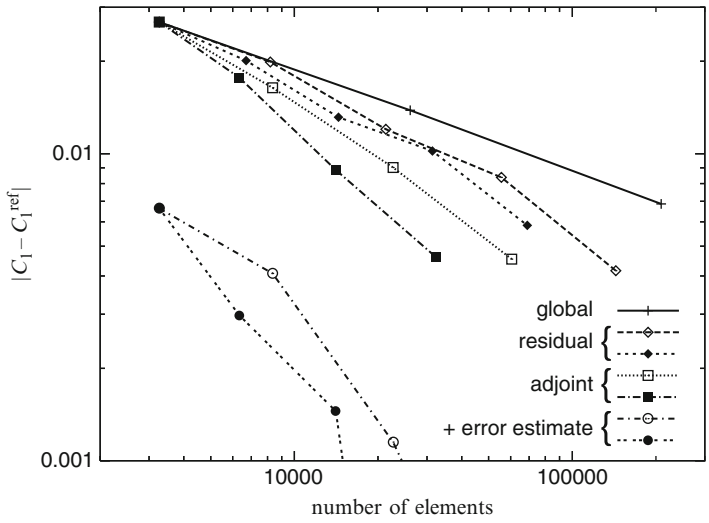


Fig. 3 Laminar delta wing: Error in the computed lift coefficient for sequences of locally refined meshes using different refinement indicators and isotropic (*open symbols*) as well as anisotropic refinement (*filled*)

specific accuracy are obtained with less elements for residual-based mesh refinement. We notice that the adjoint-based refinement procedure yields again better results.

Additionally, in case of adjoint-based mesh refinement Fig. 3 illustrates the errors of the enhanced lift coefficients obtained by adding the global error estimate to the computed lift coefficient. Already on the first adapted mesh the enhanced lift coefficient is more accurate than the unmodified values computed on the last adapted meshes.

Finally, we note that anisotropic mesh refinement using the jump indicator performs better than isotropic mesh refinement with an improvement by a factor of almost two on the final mesh in the adjoint-based case. In general, the gain improves for increasing accuracy requirements. Here, the anisotropy indicator works as a general aspect ratio optimizer.

5 Conclusion and Outlook

The presented anisotropy indicators have been successfully applied to a number of aerodynamic test cases. Especially the very simple jump indicator performs surprisingly well and is thus a good candidate for applications with increased complexity.

So far, only laminar flows with weak boundary layers have been considered. Employing a RANS approach with a suitable turbulence model much thinner and

thus stronger boundary layers dominate the flow field. Such cases are of more interest to the aerodynamicist. For theses applications different variables might be of different orders of magnitude, thus the question of how to treat systems of equations efficiently will arise again, perhaps a suitable scaling of the individual components might be necessary.

The presented approach, especially the jump indicator, will also be combined with an hp-adaptive algorithm. First experiments show promising results if the number of subdivided elements is not very small compared to the number of elements treated with an increased polynomial order.

Finally, splitting error estimation and anisotropy detection into two distinct indicators is reasonable in many cases but for the purpose of creating nearly optimal meshes for the approximation of a given target functional a combined approach respecting anisotropy in both the primal and dual solution would be ideal. Recently, Richter [15] proposed a unified approach in the context of continuous FEM and a reconstructed dual solution. Extending this approach to our application and to non-tensor-product basis functions will provide an interesting alternative to the proposed algorithm.

Acknowledgements The authors gratefully acknowledge the partial financial support of both the President's Initiative and Networking Fund of the Helmholtz Association of German Research Centres and the European project ADIGMA [12]. All computations have been performed with the flow solver PADGE [8] which employs a modified and extended version of the `deal.II` library [1].

References

1. Bangerth, W., Hartmann, R., Kanschat, G.: Deal II – A general purpose object oriented finite element library. ACM Trans. Math. Software **33**(4) (2007)
2. Bassi, F., Rebay, S., Mariotti, G., Pedinotti, S., Savini, M.: A high-order accurate discontinuous finite element method for inviscid and viscous turbomachinery flows. In: R. Decuypere, G. Dibelius (eds.) 2nd European Conference on Turbomachinery Fluid Dynamics and Thermodynamics, Antwerpen, Belgium, March 5–7, 1997, pp. 99–108. Technologisch Instituut (1997)
3. Becker, R., Rannacher, R.: An optimal control approach to a posteriori error estimation in finite element methods. Acta Numerica **10**, 1–102 (2001)
4. Formaggia, L., Micheletti, S., Perotto, S.: Anisotropic mesh adaptation in computational fluid dynamics: Application to the advection–diffusion-reaction and the Stokes problems. Appl. Numer. Math. **51**, 511–533 (2004)
5. Frey, P.J., Alauzet, F.: Anisotropic mesh adaptation for CFD computations. Comput. Meth. Appl. Mech. Eng. **194**, 5068–5082 (2005)
6. Hartmann, R., Houston, P.: Symmetric interior penalty DG methods for the compressible Navier–Stokes equations II: Goal–oriented a posteriori error estimation. Int. J. Numer. Anal. Model. **3**(2), 141–162 (2006)
7. Leicht, T., Hartmann, R.: Error estimation and anisotropic mesh refinement for 3d laminar aerodynamic flow simulations. *J. Comput. Phys.* **229**(19), 7344-7360 (2010)
8. Hartmann, R., Held, J., Leicht, T., Prill, F.: Discontinuous Galerkin methods for computational aerodynamics - 3D adaptive flow simulation with the DLR PADGE code. *Aerosp. Sci. Technol.* (2010), DOI 10.1016/j.ast.2010.04.002

9. Houston, P., Georgoulis, E.H., Hall, E.: Adaptivity and a posteriori error estimation for DG methods on anisotropic meshes. In: G. Lube, G. Rapin (eds.) Int. Conference on Boundary and Interior Layers, BAIL2006 (2006)
10. Huang, W.: Metric tensors for anisotropic mesh generation. J. Comput. Phys. **204**, 633–665 (2005)
11. Klaij, C.M., van der Vegt, J.J.W., van der Ven, H.: Space–time discontinuous Galerkin method for the compressible Navier–Stokes equations. J. Comput. Phys. **217**(2), 589–611 (2006)
12. Kroll, N.: ADGIMA – A European project on the development of adaptive higher-order variational methods for aerospace applications. 47th AIAA Aerospace Sciences Meeting (2009). AIAA 2009-176
13. Kurtz, J., Demkowicz, L.: A fully automatic hp-adaptivity for elliptic PDEs in three dimensions. Comput. Meth. Appl. Eng. **196**, 3534–3545 (2007)
14. Leicht, T., Hartmann, R.: Anisotropic mesh refinement for discontinuous Galerkin methods in two-dimensional aerodynamic flow simulations. Int. J. Numer. Meth. Fluid **56**(11), 2111–2138 (2008)
15. Richter, T.: A posteriori error estimation and anisotropy detection with the dual-weighted residual method. Int. J. Numer. Meth. Fluid (2009). DOI 10.1002/fld.2016
16. Sahni, O., Müller, J., Jansen, K.E., Shepard, M.S., Taylor, C.A.: Efficient anisotropic adaptive discretization of the cardiovascular system. Comput. Meth. Appl. Mech. Eng. **195**, 5634–5655 (2006)
17. Sun, S., Wheeler, M.F.: Anisotropic and dynamic mesh adaption for discontinuous Galerkin methods applied to reactive transport. Comput. Meth. Appl. Mech. Eng. **195**, 3382–3405 (2006)
18. Venditti, D.A., Darmofal, D.L.: Anisotropic grid adaption for functional outputs: Application to two–dimensional viscous flows. J. Comput. Phys. **187**, 22–46 (2003)

A MHD Problem on Unbounded Domains: Coupling of FEM and BEM

Wiebke Lemster and Gert Lube

Abstract We consider the MHD problem on $\mathbb{R}^3 = \Omega \cup \Omega_E$, where Ω is a bounded, conducting Lipschitz domain and Ω_E is an insulating region. After semidiscretization in time, we apply a finite element approach in Ω. A boundary element approach is used in Ω_E. We present results on the well-posedness of the continuous problem and for the semidiscrete coupled problems arising within each time step. Finally we show the quasi-optimality of a regularised FE discretisation within each time step.

1 Introduction

Magnetohydrodynamics (MHD) is the study of the flow of electrically conducting fluids in the presence of magnetic fields. In the so-called direct problem, the magnetic induction **B** and the electric field **E** are unknown. Some efforts have been made to treat this problem with Lagrange finite elements for bounded domains (cf. [2,6]). Our aim is to extend this to $\mathbb{R}^3$. In the bounded, simply connected conducting domain the model is time-dependent and nonlinear. For this reason we apply finite elements there. We regularise the system with a pressure stabilisation like method to get a saddle point problem. In the unbounded part, the model reduces to a Laplace equation. Therefore, we apply a boundary element technique for this domain. We use a symmetric coupling technique (cf. [4, 16]) to link both methods. For more informations of symmetric coupling of FEM and BEM see [4, 7, 9]. Properties of the used spaces and other helpful results can be found in [1, 5, 11, 12, 14, 15]. The paper is organised as follows: In Sects. 2 and 3 we derive a variational formulation and show the well-posedness. Section 4 deals with the time-discretised problem. Section 5 is related to the finite element approach. In Section 6 we give some final remarks.

W. Lemster (✉) and G. Lube
Department of Mathematics and Computer Science, University Göttingen, Germany
e-mail: lemster@math.uni-goettingen.de, lube@math.uni-goettingen.de

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_63,

2 Variational Formulation of the MHD Model

In this section we state the problem to find a vector potential for the interior solution and a scalar potential in the exterior domain. By $\|\cdot\|$ we denote the L^2-norm in the conducting domain Ω which is induced by the L^2 scalar product $(\cdot,\cdot)$. We seek the interior solution in the space

$$H^1(0,T;H) := \left\{\mathbf{B} \in L^2(0,T;H) \mid \mathbf{B}' \in L^2(0,T;H^*)\right\},$$
$$H := \left\{\mathbf{B} \in L^2(\Omega) \mid \nabla \times \mathbf{B} \in L^2(\Omega)\right\}.$$

Ω is divided into two disjoint parts, Ω_1 and Ω_2. We require $\overline{\Omega_2} \cap \partial\Omega = \emptyset$. We assume the known velocity $\mathbf{w}$ and the function f to be zero outside of Ω_2. The electric field is denoted by $\mathbf{E}$ the, the magnetic induction by $\mathbf{B}$, the magnetic permeability by $0 < \mu_1 \le \mu \le \mu_2$ and the electric conductivity by $0 \le \sigma_1 \le \sigma \le \sigma_2$. We can state the following problem:

$$\frac{\partial \mathbf{B}}{\partial t} = -\nabla \times \mathbf{E}, \tag{1}$$

$$\nabla \times \frac{1}{\mu}\mathbf{B} = \sigma\Big(\mathbf{E} + \mathbf{w} \times \mathbf{B} + \frac{Rf}{1 + s\,|\mathbf{B}|^2}\mathbf{B}\Big), \tag{2}$$

$$\nabla \cdot \mathbf{B} = 0, \tag{3}$$

$$\mathbf{B} = o\left(|x|^{-1}\right) \text{ for } |x| \to \infty. \tag{4}$$

We want to replace $\mathbf{B}$ by a potential ansatz. For the interior domain Ω we set $\mathbf{B} = \nabla \times \mathbf{u}$. From (1) we get $\mathbf{E} = -\frac{\partial \mathbf{u}}{\partial t} + \nabla\phi_c$. Since ϕ_c is only unique up to a constant we choose it such that $\int_\Omega \phi_c \, dx = 0$.

In the non-conducting region Ω_E it holds that $\sigma = 0$ and $\mu_0 := \mu$ is constant. Therefore (2) simplifies to $\nabla \times \mathbf{B} = 0$. We set $\mathbf{B} = \nabla\Phi$ in Ω_E. Hence equation (3) reduces to a Laplace equation. The fundamental solution for this equation is denoted by $U(x,y) := \frac{1}{4\pi}\frac{1}{|x-y|}$ and its normal derivative at the boundary by $T(x,y)$. We have the representation for Φ by the single-layer- and the double-layer-potential (cf. [8]).

$$\Phi(x) = -\int_{\partial\Omega} U(x,y)\frac{\partial}{\partial \mathbf{n}}\Phi(y)\, ds(y) + \int_{\partial\Omega} T(x,y)\Phi(y)\, ds(y), \quad x \in \Omega_E. \tag{5}$$

To normalise the vector potential $\mathbf{u}$ we use the spaces

$$H^1_*(\Omega) := \left\{q \in H^1(\Omega) \mid (q,1) = 0\right\},$$
$$H_* := \left\{\mathbf{v} \in H \mid (\mathbf{v}, \nabla q) = 0 \;\forall\; q \in H^1_*(\Omega)\right\}.$$

The transmission conditions $[\mathbf{B}\cdot\mathbf{n}] = 0$ and $[\mathbf{H}\times\mathbf{n}] = 0$ imply the interface conditions

$$(\frac{1}{\mu}\nabla\times\mathbf{u})\times\mathbf{n} = \left(\frac{1}{\mu_0}\nabla\Phi\right)\times\mathbf{n} \qquad \text{and} \qquad (\nabla\times\mathbf{u})\cdot\mathbf{n} = \frac{\partial\Phi}{\partial\mathbf{n}} \qquad \text{on } \partial\Omega.$$

This leads to the following problem:

Problem 1 *Find* $\mathbf{u} \in H^1(0,T;H_*)$ *such that for all* $\mathbf{v} \in H_*$ *and almost all* $t \in (0,T)$

$$\begin{aligned} 0 = \ & (\sigma\frac{\partial\mathbf{u}}{\partial t},\mathbf{v}) - \Big(\frac{1}{\mu}\nabla\times\mathbf{u},\nabla\times\mathbf{v}\Big) - \Big\langle\Big(\frac{1}{\mu}\nabla\times\mathbf{u}\Big)\times\mathbf{n},\mathbf{v}\Big\rangle_{\partial\Omega} \\ & -R\Big(\frac{\sigma f\nabla\times\mathbf{u}}{1+s\left|\nabla\times\mathbf{u}\right|^2},\mathbf{v}\Big) - (\sigma\mathbf{w}\times(\nabla\times\mathbf{u}),\mathbf{v}). \end{aligned} \tag{6}$$

In order to replace the boundary term in (6) by a formulation of the exterior problem, we use the Stokes formula on the boundary to get

$$\int_{\partial\Omega}\left[\left(\frac{1}{\mu}\nabla\times\mathbf{u}\right)\times\mathbf{n}\right]\cdot\mathbf{v}\,ds = \frac{1}{\mu_0}\int_{\partial\Omega}\Phi T_n\mathbf{v}\,ds$$

with $T_n v := (\nabla\times v)\cdot\mathbf{n}$. If we apply this result to (5) and take the boundary values, we derive the following Calderòn equations

$$2\,(VT_n\mathbf{u})(x) = -\Phi(x) + 2\,(K\Phi)(x), \quad 2\,(D\Phi)(x) = -T_n\mathbf{u}(x) - 2\,(K'T_n\mathbf{u})(x),$$

where the integral operators are given by (cf. [3])

$$\begin{aligned} (V\psi)(x) &:= \int_{\partial\Omega} U(x,y)\psi(y)\,ds(y), \quad (D\psi)(x) := -\frac{\partial}{\partial\mathbf{n}_x}\int_{\partial\Omega} T(x,y)\psi(y)\,ds(y), \\ \big(K'\psi\big)(x) &:= \int_{\partial\Omega}\frac{\partial}{\partial\mathbf{n}_x}U(x,y)\psi(y)\,ds(y), \quad (K\psi)(x) := \int_{\partial\Omega} T(x,y)\psi(y)\,ds(y). \end{aligned}$$

By defining the bilinear forms

$$\begin{aligned} \langle\mathscr{A}\mathbf{u},\mathbf{v}\rangle &:= (\mu^{-1}\nabla\times\mathbf{u},\nabla\times\mathbf{v}) - (\sigma\mathbf{w}\times(\nabla\times\mathbf{u}),\mathbf{v}), \\ \langle\mathscr{A}_t\mathbf{u},\mathbf{v}\rangle &:= (\sigma\frac{\partial\mathbf{u}}{\partial t},\mathbf{v}), \qquad \langle\mathscr{K}\phi,\mathbf{v}\rangle := \langle(\frac{1}{2}Id + K)\phi, T_n\mathbf{v}\rangle_{\partial\Omega}, \\ \langle\mathscr{D}\phi,\psi\rangle &:= \langle D\phi,\psi\rangle_{\partial\Omega} + \langle\phi,1\rangle_{\partial\Omega}\langle\psi,1\rangle_{\partial\Omega}, \qquad \langle\mathscr{V}\mathbf{u},\mathbf{v}\rangle := \langle VT_n\mathbf{u}, T_n\mathbf{v}\rangle_{\partial\Omega} \end{aligned}$$

and the forms

$$A_V(\mathbf{u},\phi,\mathbf{v}) := \frac{1}{\mu_0}\langle(VT_n\mathbf{u} - \frac{1}{2}\phi - K\phi), T_n\mathbf{v}\rangle_{\partial\Omega},$$

$$A_D(\mathbf{u},\phi,\psi) := \langle (D\Phi + \frac{1}{2}T_n\mathbf{u} + K'T_n\mathbf{u}),\psi\rangle_{\partial\Omega} + \langle \Phi,1\rangle_{\partial\Omega}\langle\psi,1\rangle_{\partial\Omega},$$
$$\langle \mathscr{A}_{nl}\mathbf{u},\mathbf{v}\rangle := -R\Big(\sigma\frac{f}{1+s\,|\nabla\times\mathbf{u}|^2}\nabla\times\mathbf{u},\mathbf{v}\Big),$$

we get the variational problem:

Problem 2 *Find* $(\mathbf{u},\Phi)\in H^1(0,T;H_*)\times L^2(0,T;H^{\frac{1}{2}}(\partial\Omega))$ *such that for all* $(\mathbf{v},\psi)\in H_*\times H^{\frac{1}{2}}(\partial\Omega)$ *and almost all* $t\in(0,T)$

$$\mathscr{A}_t\mathbf{u} + \mathscr{A}\mathbf{u} + \mathscr{A}_{nl}\mathbf{u} + \frac{1}{\mu_0}\mathscr{V}\mathbf{u} - \frac{1}{\mu_0}\mathscr{K}\Phi = 0, \tag{7}$$
$$\mathscr{D}\Phi + \mathscr{K}'\mathbf{u} = 0. \tag{8}$$

3 Continuous Problem

We sketch the proof of the existence and uniqueness of a solution for the continuous problem. All operators are bounded, $\mathscr{D}$ is invertible (cf. e.g.,[9]) and $\mathscr{A}+\mathscr{A}_{nl}$ fullfills a Garding inequality (cf. [2, 10]). Hence, the second equation can be transformed in such a way that we get from (7) an equation which only depends on $\mathbf{u}$.

Problem 3 *Find* $\mathbf{u}\in H^1(0,T;H_*)$ *such that for almost all* $t\in(0,T)$

$$S\mathbf{u} + \mathscr{A}_t\mathbf{u} + \mathscr{A}_{nl}\mathbf{u} := \Big(\mathscr{A} + \frac{1}{\mu_0}\mathscr{V} + \frac{1}{\mu_0}\mathscr{K}\mathscr{D}^{-1}\mathscr{K}'\Big)\mathbf{u} + \mathscr{A}_t\mathbf{u} + \mathscr{A}_{nl}\mathbf{u} = 0. \tag{9}$$

Therefore, we need an existence theorem for non-linear evolution problems (cf. [17] Theorem 30.A.).

Theorem 1. *Let* $V\subset X\subset V^*$ *be a Gelfand triple with dim* $V=\infty$. *Assume that the operators* $A := S+\mathscr{A}_{nl} : V\to V^*$ *fullfill the following conditions for* $p,q\in(1,\infty)$ *with* $\frac{1}{p}+\frac{1}{q}=1$ *and* $0<T<\infty$*:*

(a) $A(t)$ *is coercive for all* $t\in(0,T)$, *i.e., there exist constants* $M>0$ *and* $\Lambda\geq 0$ *such that*

$$\langle A(t)v,v\rangle \geq M\,\|v\|_V^p - \Lambda \qquad \forall\, v\in V \qquad \forall\, t\in(0,T).$$

(b) $A(t): V\to V^*$ *is monotone and hemicontinuous for all* $t\in(0,T)$.

(c) There exist a nonnegative function $K_1\in L^q(0,T)$ *and a constant* $K_2>0$ *such that for all* $v\in V$ *and* $t\in(0,T)$

$$\|A(t)v\|_{V^*}\leq K_1(t) + K_2\,\|v\|_V^{p-1}.$$

(d) The function $t \mapsto A(t)$ is weakly measurable, i.e., the function $t \mapsto \langle A(t)u, v\rangle_V$ is measurable for $t \in (0,T)$ and all $u, v \in V$.

Then there exists a unique solution $\mathbf{u} \in W^{1,p}(0,T;X)$ for $\mathbf{u}_0 \in X$ of

$$\mathbf{u}'(t) + A(t)\mathbf{u}(t) = 0, \qquad \mathbf{u}(0) = \mathbf{u}_0.$$

We have to modify Problem 3 in order to satisfy a coercivity condition. We multiply (9) with the function $e^{-\kappa t}$, where κ is the constant for the L^2-term in the Garding inequality which $S + \mathscr{A}_{nl}$ satisfies. The new scaled potential is denoted by $\tilde{\mathbf{u}} := e^{-\kappa t}\mathbf{u}$. S is linear, so we can replace $e^{-\kappa t}S\mathbf{u}$ by $S\tilde{\mathbf{u}}$. To get an equivalent problem, we have to scale the constant s in $\mathscr{A}_{nl}$ by $e^{2\kappa t}$. In addition, we have $e^{-\kappa t}\mathscr{A}_t\mathbf{u} = \mathscr{A}_t\tilde{\mathbf{u}} + \sigma\kappa\tilde{\mathbf{u}}$. We set $\mathbf{u} := \tilde{\mathbf{u}}$ and $S := S + \sigma\kappa\mathrm{Id}$.

Proof. We set $p = q = 2$, $V = H_r$ and $X = L^2(\Omega)$.

(a) To show the first condition, we first note

$$\left\langle \left(\mathscr{V} + \mathscr{K}\mathscr{D}^{-1}\mathscr{K}'\right)\mathbf{v}, \mathbf{v}\right\rangle \geq 0 \qquad \forall \mathbf{v} \in H.$$

The coercitivity follows with $\eta = \kappa$ and $c_k \leq \frac{1}{\mu_2}$ from

$$\begin{aligned}
\langle \mathscr{A}_{nl}(t)\mathbf{v}, \mathbf{v}\rangle + \langle S(t)\mathbf{v}, \mathbf{v}\rangle \;&\geq \frac{1}{\mu_2}\|\nabla\times\mathbf{v}\|^2 - \sigma_2\|\mathbf{u}\|_{L^\infty(\Omega)}\|\nabla\times\mathbf{v}\|\|\mathbf{v}\| \\
&\quad -\sigma_2|R|\,\|f\|_{L^\infty(\Omega)}\|\nabla\times\mathbf{v}\|\|\mathbf{v}\| + \sigma_1\eta\|\mathbf{v}\|^2 \\
&\geq \left(\frac{1}{\mu_2} - c_k\right)\|\nabla\times\mathbf{v}\|^2 - \frac{\kappa(c_k)}{2}\|\mathbf{v}\|^2 + \sigma_1\eta\|\mathbf{v}\|^2 \\
&\geq \min\left\{\frac{1}{2\mu_2}, \sigma_1\eta - \frac{\kappa(c_k)}{2}\right\}\|\mathbf{v}\|_H^2 .
\end{aligned}$$

(b) The monotonicity can be shown similarly. $\mathscr{A}_{nl}$ is Lipschitz continuous,

$$\left(\frac{f\Phi_1}{1+s|\Phi_1|^2} - \frac{f\Phi_2}{1+s|\Phi_2|^2}, \mathbf{v}\right) \leq 3\|f\|_{L^\infty(\Omega)}\|\Phi_1 - \Phi_2\|\|\mathbf{v}\| \qquad \forall \Phi_1, \Phi_2, \mathbf{v} \in H,$$

and S is linear and bounded. Therefore both operators are hemicontinuous (cf. [17] Fig. 27.1).

(c+d) The third condition follows from the boundedness of the two operators S and $\mathscr{A}_{nl}$. The operator A does not depend explicitly on time. □

By Gronwall inequality one can derive with

$$\Delta^2 := \sigma_2^2 \left(\|\mathbf{w}\|^2_{L^\infty(\Omega)} + R^2 \|f\|^2_{L^\infty(\Omega)} \right), \qquad L(t) := \tfrac{\mu_2}{\sigma_1} \int_0^t \Delta^2(s)\, ds$$

the following a-priori estimate for $t \in [0, T]$

$$\|\mathbf{u}(t)\| \le \|\mathbf{u}(0)\|\, e^{L(t)}, \qquad \|\nabla \times \mathbf{u}\| \le \sqrt{2}\mu_2 \Delta(t)\, \|\mathbf{u}(0)\|\, e^{L(t)}.$$

4 Semidiscrete Problem

We discretise Problem 3 in time by the implicit Euler scheme. An existence result is given. The proof relies on the main theorem of strongly monotone operators.

Theorem 2. *Suppose X is a real Hilbert space and the operator $A : X \to X^*$ is strongly monotone and Lipschitz continuous on X. Then, for each $b \in X^*$, the operator equation*

$$Au = b$$

has a unique solution $\mathbf{u} \in X$.

For simplicity we use an equidistant partition of the time interval $[0, T]$ into M parts with time step size $\tau = \frac{T}{M}$. Hence, we obtain the following semidiscrete problem:

Problem 4 *Find $(\mathbf{u}^n, \Phi^n) \in H_* \times H^{\frac{1}{2}}(\partial\Omega)$ such that for all $(\mathbf{v}, \psi) \in H_* \times H^{\frac{1}{2}}(\partial\Omega)$*

$$\begin{aligned} 0 = {} & \left(\sigma \frac{\mathbf{u}^n - \mathbf{u}^{n-1}}{\tau}, \mathbf{v}\right) + \left(\frac{1}{\mu} \nabla \times \mathbf{u}^n, \nabla \times \mathbf{v}\right) - R\left(\sigma \frac{f^n \nabla \times \mathbf{u}^n}{1 + s\,|\nabla \times \mathbf{u}^n|^2}, \mathbf{v}\right) \\ & - \left(\sigma \mathbf{w}^n \times (\nabla \times \mathbf{u}^n), \mathbf{v}\right) + A_V\left(\mathbf{u}^n, \Phi^n, \mathbf{v}\right), \\ 0 = {} & A_D\left(\mathbf{u}^n, \Phi^n, \psi\right). \end{aligned}$$

For an estimate of the semidiscrete solution we need the discrete Gronwall lemma.

Remark 3. Let $\{z_n\}_{n=1}^{N(\tau)}$ be a sequence of nonnegative real numbers which fullfill

$$z_n \le C_1 + \tau C_2 \sum_{i=0,\ldots,n-1} z_i \qquad \text{for } n = k, \ldots, N(\tau)$$

with C_1 and C_2 independent of τ. Let $z_i \le \tilde{z}/k$ for $i = 1, \ldots, k-1$. Then we obtain

$$z_n \le (\tau C_2 \tilde{z} + C_1)\,(1 + \tau C_2)^{n-k} \qquad \text{for } n = k, \ldots, N(\tau).$$

Hence, we get the main theorem of this section:

Theorem 4. *Problem 4 has a unique solution, if τ is chosen such that for $c_k < \frac{1}{\mu_2}$ it holds $c_1(\tau) := 1/\tau - \kappa^n/2 - (8R^2 \|f^n\|^2_{L^\infty(\Omega)})/(2c_k) > 0$. Furthermore we obtain for constant σ the following estimate:*

$$\max_{1\le n\le M} \|\mathbf{u}^n\|^2_{L^2(\Omega)} + \tau \sum_{n=1,\dots,M} \|\nabla\times\mathbf{u}^n\|^2_{L^2(\Omega)} \le C \left\|\mathbf{u}^0\right\|^2_{L^2(\Omega)}.$$

Proof. The strong monotonicity and the boundedness of $S^n + \mathscr{A}_{nl}$, defined by

$$\begin{aligned}\langle S^n\mathbf{v},\widehat{\mathbf{v}}\rangle &:= \frac{1}{\tau}(\mathbf{v},\tilde{\mathbf{v}}) + \Big(\frac{1}{\mu}\nabla\times\mathbf{v},\nabla\times\tilde{\mathbf{v}}\Big) + \Big\langle\frac{1}{\mu_0}\left(\mathscr{V}+\mathscr{K}\mathscr{D}^{-1}\mathscr{K}'\right)\mathbf{v},\tilde{\mathbf{v}}\Big\rangle \\ &\quad -(\sigma\mathbf{w}^n\times(\nabla\times\mathbf{v}),\tilde{\mathbf{v}}),\\ \langle\mathscr{A}^n_{nl}\mathbf{u},\mathbf{v}\rangle &:= -\Big(\frac{Rf^n}{1+s|\mathbf{u}|^2}\mathbf{u},\nabla\times\mathbf{v}\Big),\end{aligned}$$

can be proved by similar arguments as presented in the previous section. Therewith we can apply Theorem 2 to prove existence and uniqueness. □

5 Finite Element and Boundary Element Approach

By introducing a Lagrange multiplier $\tilde{p}^n$ Problem 4 can be regularised with an pressure stabilisation like term $\epsilon(\nabla\tilde{p}^n,\nabla q)$.

Problem 5 *Find $(\tilde{\mathbf{u}}^n,\tilde{p}^n,\tilde{\Phi}^n) \in H\times H^1(\Omega)\times H^{\frac{1}{2}}(\partial\Omega)$ such that for all test functions $(\mathbf{v},q,\psi)\in H\times H^1(\Omega)\times H^{\frac{1}{2}}(\partial\Omega)$*

$$\begin{aligned}\frac{\sigma}{\tau}\Big(\mathbf{u}^{n-1},\mathbf{v}\Big) &= \Big(\frac{\sigma}{\tau}\tilde{\mathbf{u}}^n,\mathbf{v}\Big) + \Big(\frac{1}{\mu}\nabla\times\tilde{\mathbf{u}}^n,\nabla\times\mathbf{v}\Big) + \langle\mathscr{A}^n_{nl}\tilde{\mathbf{u}}^n,\mathbf{v}\rangle\\ &\quad -(\sigma\mathbf{w}^n\times(\nabla\times\tilde{\mathbf{u}}^n),\mathbf{v}) + A_V(\tilde{\mathbf{u}}^n,\tilde{\Phi}^n,\mathbf{v}) + (\nabla\tilde{p}^n,\mathbf{v}),\\ (\tilde{\mathbf{u}}^n,\nabla q) &= \epsilon(\nabla\tilde{p}^n,\nabla q) + (\tilde{p}^n,1)(q,1),\\ 0 &= A_D\left(\tilde{\mathbf{u}}^n,\tilde{\Phi}^n,\psi\right).\end{aligned}$$

One can also introduce the Lagrange multiplier p^n in the original semi-discrete problem. For time steps like in Theorem 4 we get the error for the regularisation

$$c_1(\tau)\|\mathbf{u}^n-\tilde{\mathbf{u}}^n\|^2 + \|\nabla\times(\mathbf{u}^n-\tilde{\mathbf{u}}^n)\|^2 + \epsilon\mu_2\|\nabla(p^n-\tilde{p}^n)\|^2 \le \epsilon\mu_2\|\nabla p^n\|^2.$$

Consider the Galerkin discretisation of the coupled Problem 5. Let $X_h\subset H$ be the lowest order edge element space (see [13]) and $M_h\subset H^1(\Omega)$ the space of piecewise linear elements on a shape-regular tetrahedral mesh of Ω with mesh size h.

The space $W_h \subset H^{-\frac{1}{2}}(\partial\Omega)$ is defined by the traces on $\partial\Omega$ of piecewise linear nodal elements.

Problem 6 *Find* $(\tilde{\mathbf{u}}_h^n, \tilde{p}_h^n, \tilde{\Phi}_h^n) \in X_h \times M_h \times W_h$ *such that for all test functions* $(\mathbf{v}, q, \psi) \in X_h \times M_h \times W_h$

$$\begin{aligned}
\frac{\sigma}{\tau}\left(\mathbf{u}_h^{n-1}, \mathbf{v}\right) = & \left(\frac{\sigma}{\tau}\tilde{\mathbf{u}}_h^n, \mathbf{v}\right) + \left(\frac{1}{\mu}\nabla\times\tilde{\mathbf{u}}_h^n, \nabla\times\mathbf{v}\right) - \left(\sigma\mathbf{w}^n\times\left(\nabla\times\tilde{\mathbf{u}}_h^n\right), \mathbf{v}\right) \\
& + \left\langle \mathscr{A}_{nl}^n\tilde{\mathbf{u}}_h^n, \mathbf{v}\right\rangle + A_V(\tilde{\mathbf{u}}_h^n, \tilde{\Phi}_h^n, \mathbf{v}) + \left(\nabla\tilde{p}_h^n, \mathbf{v}\right), \\
\left(\tilde{\mathbf{u}}_h^n, \nabla q\right) = & \ \epsilon\left(\nabla\tilde{p}_h^n, \nabla q\right) + \left(\tilde{p}_h^n, 1\right)(q, 1), \\
0 = & \ A_D(\tilde{\mathbf{u}}_h^n, \tilde{\Phi}_h^n, \psi).
\end{aligned}$$

Lemma 1. *The error between the regularised semi-discrete problem and its Galerkin formulation can be estimated as follows*

$$\begin{aligned}
\left\|\tilde{\mathbf{u}}^n - \tilde{\mathbf{u}}_h^n\right\|_H \lesssim & \inf_{\mathbf{v}_h\in X_h}\|\mathbf{u}^n - \mathbf{v}_h\|_H + \inf_{q_h\in M_h}\|p - q_h\|_{H^1(\Omega)} \\
& + \inf_{\psi_h\in W_h}\|\Phi - \psi_h\|_{H^{\frac{1}{2}}(\partial\Omega)}.
\end{aligned}$$

Proof. Defining an operator $\widetilde{S}^n$ like S^n for the new problem leads to the additional term $\left\langle\mathscr{B}'\mathscr{C}^{-1}\mathscr{B}\mathbf{v}, \tilde{\mathbf{v}}\right\rangle$. The ellipticity of this operator and some other transformations lead to the statement (cf. [9, 10]). □

6 Conclusions

We presented a symmetric coupling approach for a nonlinear MHD model (cf. [2]). The continuous and the time discrete problems are well posed. A quasi-optimal estimate for the regularised problem is given. Part of our long-term objectives is to develop and implement an algorithm for this problem.

Acknowledgements We want to thank O. Steinbach for stimulating discussions on the symmetric coupling of FEM and BEM.

References

1. Alonso, A., Valli, A.: Some remarks on the characterization of the space of tangential traces of $H(\mathrm{rot};\Omega)$ and the construction of an extension operator. Manuscr. Math. **89**, 159–178 (1996)
2. Chan, K.H., Zhang, K., Zou, J.: Spherical interface dynamos: Mathematical theory, finite element approximation and application. Technical Report CUHK-2005-10 (331) http://www.math.cuhk.edu.hk/en/report/index.php
3. Costabel, M.: Boundary integral operators on Lipschitz domains: Elementary results. SIAM J. Math. Anal. **19**, 613–626 (1988)

4. Costabel, M.: Symmetric methods for the coupling of finite elements and boundary elements. In: Brebbia, C.A., Wendland, W.L., Kuhn, G. (eds.) Boundary Elements IX Vol.**1**, pp. 411–420. Springer (1987)
5. Girault, V., Raviart, P.-A.: Finite element methods for Navier-Stokes equations. Theory and algorithms. Springer, Berlin (1986)
6. Guermond, J.-L., Laguerre, R., Léorat, J., Nore, C.: An interior penalty Galerkin method for the MHD equations in heterogeneous domains. J. Comput. Phys. **221**, 349–369 (2007)
7. Hiptmair, R.: Symmetric coupling for eddy current problems. SIAM J. Numer. Anal. **40**, 41–65 (2002)
8. Kress, R.: Linear integral equations. 2nd ed. Springer, New York (1999)
9. Kuhn, M., Steinbach, O.:Symmetric coupling of finite and boundary elements for exterior magnetic field problems. Math. Methods Appl. Sci. **25**, 357–371 (2002)
10. Lemster, W.: A vector potential ansatz for a MHD problem. (in German) http://num.math.uni-goettingen.de/lemster/
11. Lions, J.L., Magenes, E.: Problèmes aux limites non homogenes et applications. Vol. 1. Dunod, Paris (1968)
12. Monk, P.: Finite element methods for Maxwell's equations. Oxford University Press, Oxford (2003)
13. Nedelec, J.-C.: Mixed finite elements in $\mathbb{R}^3$. Numer. Math. **35**, 315–341 (1980)
14. Picard, R.: An elementary proof for a compact imbedding result in generalized electromagnetic theory. Math. Z. **187**, 151–164 (1984)
15. Růžička, M.: Nonlinear functional analysis. An introduction. (Nichtlineare Funktionalanalysis. Eine Einführung.) Springer, Berlin (2004)
16. Steinbach, O.: Numerical approximation methods for elliptic boundary value problems. Finite and boundary elements. (Numerische Näherungsverfahren für elliptische Randwertprobleme. Finite Elemente und Randelemente.) Teubner, Stuttgart (2003)
17. Zeidler, E.: Nonlinear functional analysis and its applications II/B.Springer, New York (1990)

A Stable and High Order Interface Procedure for Conjugate Heat Transfer Problems

Jens Lindström and Jan Nordström

Abstract This paper analyzes stability and order of accuracy of a conjugate heat transfer problem in one space dimension. The energy method is used to derive boundary and interface conditions for the continuous problem and the resulting numerical scheme is proven stable. The scheme is implemented using 2nd-, 3rd- and 4th-order finite difference operators on Summation-By-Parts (SBP) form. The boundary and interface conditions are implemented weakly using the Simultaneous Appriximation Term (SAT). The rate of convergence is verified using the method of manufactured solutions.

1 Introduction

The coupling of fluid and heat equations is an area that has many interesting scientific and engineering applications. From the scientific side it is interesting to mathematically derive conditions to make the coupled system well posed and compare with actual physics. The applications for conjugate heat transfer ranges between cooling of turbine blades, electronic components, nuclear reactors or space-craft re-entry just to name a few. The particular application we are working towards

J. Lindström (✉)
Division of Scientific Computing, Department of Information Technology, Uppsala University, Box 337, SE-751 05 Uppsala, Sweden
e-mail: Jens.Lindstrom@it.uu.se

J. Nordström
Division of Scientific Computing, Department of Information Technology, Uppsala University, Box 337, SE-751 05 Uppsala, Sweden
School of Mechanical, Industrial and Aeronautical Engineering, University of the Witvatersrand, PO WITS 2050, Johannesburg, South Africa
and
Department of Aeronautics and Systems Integration, FOI, The Swedish Defense Research Agency, SE-164 90 Stockholm, Sweden
e-mail: Jan.Nordstrom@it.uu.se

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_64,

here is a microscale satellite cold gas propulsion system with heat sources that will be used for controlling the flow rate [4].

This paper is the first step in understanding the coupling procedure within our framework. The computational method that we are using is a finite difference method on Summation-By-Parts (SBP) form with the Simultaneous Approximation Term (SAT). This method has been developed in many papers [1, 2, 5, 9] and used for many difficult problems where it has proven to be robust [7, 11, 12].

2 The Continuous Problem

The equations we are studying in this paper are motivated by a gas flow in a long channel with heat sources. The channel is long compared to the height and hence the changes in the tangential direction are small in comparison to the changes in the normal direction. The equations are an incompletely parabolic system of equations for the flow and the scalar heat equation for the heat transfer,

$$w_t + Aw_x = \varepsilon B w_{xx}, \qquad -1 \le x \le 0, \tag{1}$$

and

$$T_t = kT_{xx}, \qquad 0 \le x \le 1, \tag{2}$$

where

$$w = \begin{bmatrix} \rho \\ u \\ \mathscr{T} \end{bmatrix}, \qquad A = \begin{bmatrix} a & b & 0 \\ b & a & c \\ 0 & c & a \end{bmatrix}, \qquad B = \begin{bmatrix} 0 & 0 & 0 \\ 0 & \alpha & 0 \\ 0 & 0 & \beta \end{bmatrix}. \tag{3}$$

We can view (1) as the Navier–Stokes equations linearized and symmetrized around a state with some mean velocity. In that case we would have

$$a = \bar{u}, \qquad b = \frac{\bar{c}}{\sqrt{\gamma}}, \qquad c = \bar{c}\sqrt{\frac{\gamma - 1}{\gamma}}, \qquad \alpha = \frac{\lambda + 2\mu}{\bar{\rho}}, \qquad \beta = \frac{\gamma\mu}{Pr\bar{\rho}}, \tag{4}$$

where $\bar{u}$, $\bar{\rho}$ and $\bar{c}$ is the mean velocity, density and speed of sound. γ is the ratio of specific heats, Pr the Prandtl number and λ and μ are the second and dynamic viscosities [12]. At this point the only assumption on the coefficients is that $a, \alpha, \beta > 0$ to keep the discussion general. Our objective is to couple (1) and (2) at $x = 0$ and investigate which interface conditions that will lead to an energy estimate.

2.1 Boundary and Interface Conditions

Since we are concerned with the interface between the equations in this report, the boundary conditions will not be analyzed. They have been derived using the energy

method and the result is stated below. At the left boundary $x = -1$ we have the semi characteristic boundary conditions

$$\frac{1}{\sqrt{2}d}(-\sqrt{2}c\rho + \sqrt{2}b\mathscr{T}) = f_1(t), \tag{5}$$

$$\frac{1}{\sqrt{2}d}(b\rho + du + c\mathscr{T}) = f_2(t) \tag{6}$$

$$\alpha d u_x - \beta c \mathscr{T}_x = f_3(t). \tag{7}$$

At the right boundary $x = 1$ we use a Dirichlet boundary condition on the temperature.

We will again use the energy method to derive the interface conditions, more in detail. Define the energy norm of w as

$$||w||^2 = \int_{\Omega} w^T w d\Omega. \tag{8}$$

By multiplying (1) with w^T, (2) with T, integrating them over their respective domain and adding them together we get (when ignoring boundary terms)

$$\frac{d}{dt}(||w||^2 + ||T||^2) = -w^T A w + 2\varepsilon w^T B w_x - 2kTT_x - 2\varepsilon \int_{-1}^{0} w_x^T B w_x dx - 2k \int_{0}^{1} T_x^2 dx. \tag{9}$$

where the boundary terms are evaluated at $x = 0$. Since we are considering the interface as a solid wall which separates the fluid from the solid and since we want a continuous heat transfer we impose

$$u = 0, \qquad \mathscr{T} = T. \tag{10}$$

Using the interface conditions (10), (9) reduces to

$$\frac{d}{dt}(||w||^2 + ||T||^2) = 2\mathscr{T}(\beta\varepsilon\mathscr{T}_x - kT_x) - 2\varepsilon \int_{-1}^{0} w_x^T B w_x dx - 2k \int_{0}^{1} T_x^2 dx \tag{11}$$

and we can easily see that if we impose

$$\beta\varepsilon\mathscr{T}_x - kT_x = 0 \tag{12}$$

as the final interface condition we get an energy estimate.

The Laplace transform technique has been applied to the coupled system and it has been proved that the above boundary and interface conditions constitute a minimal set of conditions. The energy method was used to derive the correct form

of the conditions and hence this coupled system is computationally meaningful. A formal proof of well-posedness would require additional work on the existence. Given the existence, uniqueness and bounded sensitivity follows by the minimal number of conditions and the energy estimate [8, 10].

3 The Semidiscrete Problem

Equation (1) is discretized on the single domain $[-1, 0]$ on a uniform grid of $M+1$ grid points. The vector $\mathbf{w} = [w_0, w_1, \ldots, w_M]^T = [\rho_0, u_0, \mathscr{T}_0, \rho_1, u_1, \mathscr{T}_1, \ldots, \rho_M, u_M, \mathscr{T}_M]^T$ is the discrete approximation of w. The derivatives are approximated by the operators on SBP form

$$\mathbf{w}_x \approx (D_1^L \otimes I_3)\mathbf{w} = (P_L^{-1} Q_L \otimes I_3)\mathbf{w} \tag{13}$$

$$\mathbf{w}_{xx} \approx (D_2^L \otimes I_3)\mathbf{w} = (P_L^{-1} Q_L \otimes I_3)^2\mathbf{w} \tag{14}$$

where P_L is a symmetric positive definite matrix and Q_L is an almost skew symmetric matrix satisfying $Q_L + Q_L^T = B_L = \mathrm{diag}(-1, 0, \ldots, 0, 1)$ [5, 9]. I_3 is the 3×3 identity matrix. Equation (2) is similarly discretized on a uniform grid of $N+1$ grid points.

Remark 1. The approximation (14) has the drawback that the computational stencil is wide. Compact formulations that uses minimal bandwidth does however exist [5].

Equations (1) and (2) can be discretized with the boundary and interface conditions using the SAT method as

$$\begin{aligned}
\mathbf{w}_t = & -(D_1^L \otimes A)\mathbf{w} + \varepsilon(D_2^L \otimes B)\mathbf{w} \\
& + (P_L^{-1} E_0^L \otimes \Sigma_1^0)(X^T w_0 - g_1^0) \\
& + (P_L^{-1} E_0^L \otimes \Sigma_3^0)(\alpha d (D_1^L u)_0 - \beta c (D_1^L \mathscr{T})_0 - g_3^0) \\
& + (P_L^{-1} (D_1^L)^T E_0^L \otimes \Sigma_5^0)(c u_0 + d\,\mathscr{T}_0 - g_5^0) \\
& + (P_L^{-1} E_M^L \otimes \Sigma_1^M)(w_M - g_1^M) \\
& + (P_L^{-1} E_M^L \otimes \Sigma_2^M)(w_M - g_1^M) \\
& + (P_L^{-1} E_M^L \otimes \Sigma_3^M)(\mathscr{T}_M - T_0) \\
& + (P_L^{-1} (D_1^L)^T E_M^L \otimes \Sigma_4^M)(\mathscr{T}_M - T_0) \\
& + (P_L^{-1} E_M^L \otimes \Sigma_5^M)(\beta\varepsilon (D_1^L \mathscr{T})_M - k (D_1^R T)_0) \\
& - \mathrm{DI}_L
\end{aligned} \tag{15}$$

$$
\begin{aligned}
\mathbf{T}_t &= kD_2^R\mathbf{T} \\
&+ \tau_1^0 P_R^{-1} E_0^R (T_0 - \mathscr{T}_M) \\
&+ \tau_2^0 P_R^{-1} (D_1^R)^T E_0^R (T_0 - \mathscr{T}_M) \\
&+ \tau_3^0 P_R^{-1} E_0^R (k(D_1^R T)_0 - \beta\varepsilon(D_1^L \mathscr{T})_M) \\
&+ \tau_1^N P_R^{-1} E_N^R (T_N - h_1^N) \\
&- \mathrm{DI}_R.
\end{aligned} \tag{16}
$$

The matrices $E_0^L = \mathrm{diag}(1, 0, \ldots, 0)$, $E_M^L = \mathrm{diag}(0, \ldots, 0, 1)$ and $E_{0,N}^R$ similarly defined, are used to select boundary elements. The 3×3 matrices $\Sigma_i^{0,M}$ and coefficients $\tau_j^{0,N}$ are called penalty matrices and penalty coefficients which have to be determined for stability [5,9]. The last term in (15) and (16) are artificial dissipation operators which reduces spurious oscillations. An extensive study of these operators can be found in [6]. These are undivided differece operators of the same order as the scheme for which an energy estimate can be obtained. Hence they do not cause stability problems or reduces the overall accuracy of the scheme. To keep the notation as simple as possible they will not be analyzed in the following estimates.

3.1 Stability Conditions at $x = 0$

We let all boundary data in (15) and (16) be zero and multiply $\mathbf{w}$ and $\mathbf{T}$ from the left with $(\mathbf{w}^\mathbf{T} \otimes I_3)$ and $\mathbf{T}^\mathbf{T}$ respectively. By using the SBP property of the operators we obtain

$$
\begin{aligned}
\frac{d}{dt}(||\mathbf{w}||_{P_L}^2 + ||\mathbf{T}||_{P_R}^2) &= -w_M^T A w_M + 2\varepsilon w_M^T B (D_1^L w)_M \\
&- 2\varepsilon (D_1^L \mathbf{w})^T (I_N \otimes B)(D_1^L \mathbf{w}) \\
&+ 2w_M^T \Sigma_1^M w_M + 2w_M^T \Sigma_2^M w_M \\
&+ 2w_M^T \Sigma_3^M (\mathscr{T}_M - T_0) + (D_1^L w)_N^T \Sigma_4^M (\mathscr{T}_M - T_0) \\
&+ 2w_M^T \Sigma_5^M (\beta\varepsilon (D_1^L w)_M - k(D_1^R T)_0) \\
&- 2kT_0 (D_1^R T)_0 - 2k(D_1^R \mathbf{T})^T P_R (D_1^R \mathbf{T}) \\
&+ 2\tau_1^0 T_0 (T_0 - \mathscr{T}_M) + 2\tau_2^0 (D_1^R T)_0 (T_0 - \mathscr{T}_M) \\
&+ 2\tau_3^0 T_0 (k(D_1^R T)_0 - \beta\varepsilon (D_1^L \mathscr{T})_M)
\end{aligned} \tag{17}
$$

where all outer boundary terms have been neglected. By definition [3] the scheme (17) will be stable if we can prove that $\frac{d}{dt}(||\mathbf{w}||_{P_L}^2 + ||\mathbf{T}||_{P_R}^2) \leq 0$. Hence we need to choose appropriate penalty matrices and coefficients such that this condition is satisfied.

We choose the penalty matrices as

$$\Sigma_1^M = \begin{bmatrix} 0 & \sigma_1^H & 0 \\ 0 & \sigma_2^H & 0 \\ 0 & \sigma_3^H & 0 \end{bmatrix}, \qquad \Sigma_2^M = \begin{bmatrix} 0 & 0 & 0 \\ 0 & \sigma_2^M & 0 \\ 0 & 0 & 0 \end{bmatrix} \tag{18}$$

$$\Sigma_3^M = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \sigma_3^M \end{bmatrix}, \quad \Sigma_4^M = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \sigma_4^M \end{bmatrix}, \quad \Sigma_5^M = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \sigma_5^M \end{bmatrix}, \tag{19}$$

where Σ_1^M is the penalty matrix for the hyperbolic part of u, Σ_2^M for the parabolic part of u and $\Sigma_{3,4,5}^M$ for the coupling terms. With these choices of penalty matrices, (17) can be expanded and all coefficients determined as

$$\sigma_1^H = \frac{b}{2}, \quad \sigma_2^H \leq 0, \quad \sigma_3^H = \frac{c}{2}, \quad \sigma_2^M \leq \frac{-\alpha\varepsilon}{4p_M^L}, \quad \sigma_3^M = \tau_1^0 \leq 0 \tag{20}$$

where $p_M^L > 0$ is such that $P_L^{(M,M)} - p_M^L \geq 0$. See [1,2]. Moreover, we have

$$s \in \mathbb{R}, \quad \sigma_4^M = -\beta\varepsilon(1+s), \quad \sigma_5^M = s, \quad \tau_2^0 = -ks, \quad \tau_3^0 = 1+s. \tag{21}$$

With these coefficients we have the energy estimate $\frac{d}{dt}(||\mathbf{w}||_{P_L}^2 + ||\mathbf{T}||_{P_R}^2) \leq 0$ and hence the interface treatment is stable. Details on the technique used to obtain the coefficients can be found in e.g., [1,2].

4 Numerical Results

An example of a solution, where the coefficients are chosen as

$$a = 0.5, \quad b = \frac{1}{\sqrt{\gamma}}, \quad c = \sqrt{\frac{\gamma-1}{\gamma}}, \quad \gamma = 1.4, \quad \alpha = \beta = 1, \quad \varepsilon = 0.1, \quad k = 1 \tag{22}$$

is given in Fig. 1. We start with zero initial data and at time $t = 0$ we let $\rho = 0$, $u = 0.5$ and $\mathscr{T} = 1$ at the left boundary while $T = 0$ at the right boundary. The values at the left boundary are transformed into data for the boundary conditions. The order of convergence is studied by the method of manufactured solutions. A small enough time step has been chosen in order to minimize the errors from the time discretization, which in this case is done by the classical 4th-order explicit Runge–Kutta method. We use the functions

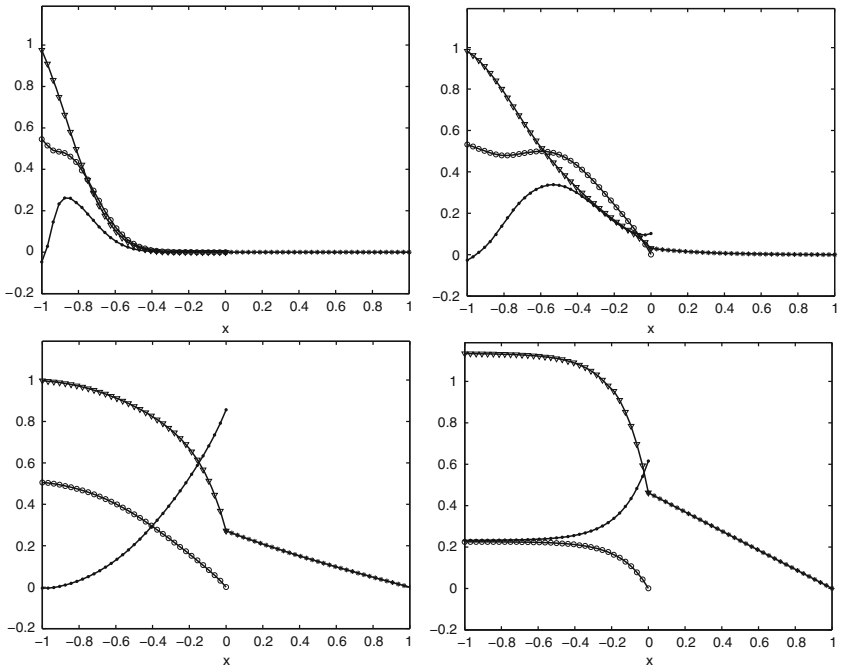


Fig. 1 ρ (*solid*), u (*circle*), $\mathscr{T}$ (*triangle*), T (*star*) for time $t = 0.15$, $t = 0.45$, $t = 1.25$ and $t = 5.0$. A sequence of solutions for different times using $M = N = 32$ grid points and 3rd-order operators. The last figure shows the steady-state solution

$$\rho = xe^{-\kappa t}, \quad u = \sin(x)e^{-\kappa t}, \quad \mathscr{T} = \frac{1}{\varepsilon}\sin(x)e^{-\kappa t}, \quad T = \frac{1}{k}\sin(x)e^{-\kappa t}, \quad \kappa = 0.1 \tag{23}$$

which inserted into (1) and (2) gives a modified system of equations with additional forcing functions. The functions (23) have been chosen since they satisfy the interface conditions in a non-trivial way. Using (23) we create exact initial and time dependent boundary conditions while no data is created at the interface. The rate of convergence is obtained as

$$q_j^i = \log_{10}\left(\frac{||u_{j-1}^i - v_{j-1}^i||}{||u_j^i - v_j^i||}\right) / \log_{10}\left(\frac{h_j}{h_{j-1}}\right) \tag{24}$$

where q_j^i denotes the convergence rate for either of the variables $i = \rho, u, \mathscr{T}, T$ at mesh refinement level j. u_j^i is the exact analytic solution for either of the variables i at mesh refinement level j and v_j^i is the discrete solution. The ratio h_j / h_{j-1} is the ratio between the number of grid points at each refinement level. The results can be seen in Table 1.

Table 1 Order of convergence. The order of convergence agree with the theoretically expected results

$M = N$	2nd-order	3rd-order	4th-order	2nd-order	3rd-order	4th-order
	ρ	ρ	ρ	u	u	u
32	0.2895	2.0197	2.5359	1.6700	2.7938	3.7233
64	1.0769	3.0137	3.7153	2.0652	3.3314	3.9939
128	1.7340	3.5255	4.1774	2.1487	3.1518	4.3242
256	2.0922	3.3945	4.1646	2.1047	3.0587	4.1851
512	2.2167	3.1591	4.1140	2.0588	3.0229	4.0531
	$\mathscr{T}$	$\mathscr{T}$	$\mathscr{T}$	T	T	T
32	0.9780	2.7634	3.8021	2.3601	3.1699	4.0291
64	1.7613	2.7542	3.3286	2.1627	3.2639	3.9000
128	2.0164	2.9310	3.5881	2.0824	3.1205	3.9133
256	2.0277	2.9789	3.7798	2.0420	3.0492	3.9476
512	2.0212	2.9928	3.8895	2.0213	3.0226	3.9711

5 Summary and Conclusions

An incompletely parabolic system of equations has been coupled with the heat equation in one space dimension. The energy method has been used to derive boundary and interface conditions and the resulting numerical scheme has been proven stable using finite differences on SBP form and the SAT boundary and interface treatment. The rate of convergence is verified to be 2nd-, 3rd- and 4th-order by using the method of manufactured solutions. This is consistent with the theoretically expected results.

References

1. Carpenter, M.H., Nordström, J., Gottlieb, D.: A Stable and Conservative Interface Treatment of Arbitrary Spatial Accuracy. J. Comp. Phys. **148**(2), 341–365 (1999)
2. Gong, J., Nordström, J.: *Stable, Accurate and Efficient Interface Procedures for Viscous Problems.* Tech. Rep., Uppsala University, Department of Information Techology, Division of Scientific Computing (2006)
3. Kreiss, H.O., Gustafsson, B., Oliger, J.: *Time Dependent Problems and Difference Methods.* Wiley, New York (1995)
4. Lindström, J., Bejhed, J., Nordström, J.: Measurements and Numerical Modelling of Orice ow in Microchannels In the 41st AIAA Thermophysics Conference, AIAA Paper No. 2009-4098, San Antonio, USA, 22–25 June (2009)
5. Mattsson, K., Nordström, J.: Summation by Parts Operators for nite Difference Approximations of Second Derivatives. J. Comp. Phys. **199**(2), 503–540 (2004)
6. Mattsson, K., Svärd, M., Nordström, J.: Stable and Accurate Artificial Dissipation. J. Sci. Comput. **21**(1), 57–79 (2004)
7. Mattsson, K., Svärd, M., Carpenter, M.H., Nordström, J.: High-order Accurate Computations for Unsteady Aerodynamics. Comput. Fluids **36**(3), 636–649 (2007)
8. Renardy, M., Rogers, R.C.: *An Introduction to Partial Differential Equations.* Springer, Berlin (1993)

9. Strand, B.: Summation by Parts for Finite Difference Approximations for d/dx. J. Comp. Phys. **110**(1), 47–67 (1994)
10. Svärd, M., Nordström, J.: Well-posed Boundary Conditions for the Navier–Stokes Equations. J. Comp. Phys. **43**(3), 1231–1255 (2005)
11. Svärd, M., Nordström, J.: A Stable High-order Finite Difference Scheme for the Compressible Navier–Stokes Equations: No-slip Wall Boundary Conditions. J. Comp. Phys. **227**(10), 4805–4824 (2008)
12. Svärd, M., Carpenter, M.H., Nordström, J.: A Stable High-order Finite Difference Scheme for the Compressible Navier–Stokes Equations, Far-field Boundary Conditions. J. Comp. Phys. **225**(1), 1020–1038 (2007)

Local Time-Space Mesh Refinement for Finite Difference Simulation of Waves

Vadim Lisitsa, Galina Reshetova, and Vladimir Tcheverda

Abstract This paper presents a new approach to a local time-space mesh refinement for a finite difference simulation of waves. The approach is based on the approximation of wave equation at the interface where two grids are coupled. As no interpolation or projection techniques are used the finite difference scheme preserves second order of convergence. We proved that this approach is low-reflecting and the artificial reflections are about 10^{-4} of incident wave. We also proved that if successive refinement is applied, i.e., temporal and spatial steps are refined at a different interfaces, the approach is stable.

1 Introduction

The main motivation of the presented research is the development of a finite difference algorithm to simulate elastic wave propagation in multi-scale media. As an example of such a media one can keep in mind fractured carbonated reservoirs of oil and gas. Models of this type possess rather strong scattered waves and may also affect the macroscopic velocity model. Typical discretizations used for simulation of seismic waves propagation are about 10 points per minimal wavelength, so that a grid step is about 100 m. On the other hand one needs for the steps of about 0.1 m to match micro-heterogeneity as fractures, cracks, caverns etc. Note that this micro-heterogeneities are usually located within a layer (subdomain) of the size of 100 m. So, it is natural to use a fine grid at the subdomain while a coarse grid is to be applied elsewhere. Applying refinement of spatial steps it is necessary to implement temporal steps refinement to avoid dispersion and reduce computational time.

V. Lisitsa (✉) and V. Tcheverda
Institute of Petroleum Geology and Geophysics of SB RAS, 3 Koptug pr., Novosibirsk, Russia
e-mail: lisitsavv@ipgg.nsc.ru, tcheverdava@ipgg.nsc.ru

G. Reshetova
Institute of Computational Mathematics and Mathematical Geophysics of SB RAS, 6 Lavrentev pr., Novosibirsk, Russia
e-mail: kgv@sscc.nsc.ru

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_65,

Thus, a suitable approach to a spatio-temporal mesh refinement for finite difference simulation of wave propagation problem is needed.

Several approaches to refine a grid for hyperbolic problems have been done by now. The simplest and most well-known one is that based on the interpolation of the solution with respect to time at the refinement interface [5]. This approach possesses rather low reflections but as shown in [2] that it may cause instability. A completely different technique was proposed in [3] and studied in [4]. This algorithm is based on energy conservation, hence it is stable. On the other hand it is rather complicated and hard to implement especially in 2 and 3D. In addition it possesses high artificial reflections up to 0.1 of incident wave. It means that this approach is not suitable to simulate scattered waves which are about 10^{-3}–10^{-2} of incident wave.

In this paper we propose a new approach to grid refinement. First of all the method is based on the approximation of wave equation at the interface, see Sect. 3. Thus, it preserves second order of convergence which is not the case for interpolation type methods, [2]. As it is shown in Sect. 3.1 the artificial reflections caused by the refinement are about 10^{-4} for typical discretizations. In the Sect. 3.2 we proved that the approach is stable if the refinement of a grid is performed in turns. It means that the temporal and spatial steps are refined at a different interfaces.

2 Preliminary

2.1 Statement of the Problem

Consider a 1D scalar wave equation written as first order system:

$$\frac{1}{c^2}\frac{\partial p}{\partial t} - \frac{\partial u}{\partial x} = 0,\ \frac{\partial u}{\partial t} - \frac{\partial p}{\partial x} = 0\,, \tag{1}$$

with $(t, x) \in D = \{t \geq 0, x \in \mathbf{R}\}$. The initial conditions are considered to be zero. The source introduced as right-hand side will be specified below. We assume the velocity c to be equal to 1, otherwise the problem can be unscaled.

In order to resolve the problem numerically the following finite difference scheme on staggered grids is implemented:

$$\frac{p_j^n - p_j^{n-1}}{\tau} = \frac{u_{j+1/2}^{n-1/2} - u_{j-1/2}^{n-1/2}}{h},\ \frac{u_{j+1/2}^{n+1/2} - u_{j+1/2}^{n-1/2}}{\tau} = \frac{p_{j+1}^n - p_j^n}{h} \tag{2}$$

where τ and h are the temporal and spatial grid steps respectively.

To investigate the properties of a scheme for wave propagation it is natural to introduce new parameters, see [1] for details:

- Number of grid points per wavelength (ppw) $N = \frac{2\pi}{kh}$, where k is the spatial frequency;
- Courant ratio $\alpha = \frac{\tau}{h} \leq 1$.

2.2 Theoretical Reflection Coefficients

Velocity of the numerical solution depends on the discretization, implementation of different grids possesses artificial reflections. The phase velocity of a finite difference solution computed by (2) is dispersive and can be represented as:

$$c_j = \frac{N_j}{\alpha_j \pi} \arcsin\left(\alpha_j \sin\left(\frac{\pi}{N_j}\right)\right),$$

where N_j is a number of grid points per wavelength on jth grid, and α_j is a Courant ratio, see [1] for the details. On this base one may estimate the artificial reflections by the formula

$$R = \frac{c_2 - c_1}{c_1 + c_2},$$

where c_j is a phase velocity on a jth grid, $j = 1, 2$. The particular coefficient corresponds to the case of right-going wave, that is the wave propagating from the region with velocity c_1 to that with the velocity c_2. This formula allows us to estimate the reflections on the base of dispersion analysis of the scheme and regardless of particular conjugation conditions used at the interface. We studied the estimations in dependence on a type of refinement (spatial, temporal, simultaneous), for the spatial discretisations – $N = 20$ to 100 ppw, Courant ratios $\alpha = 0.1$ to 1 and refinement ratios K and L denoting the ratios of coarse grid steps to fine grid steps with respect to time and space respectively. We proved that artificial reflections caused by the velocity dispersion are about 10^{-4}. Later we will call an approach low-reflected if the artificial reflections possessed by the algorithm are about 10^{-4}, i.e., close to those estimated on the basis of velocity difference.

3 Description of the Approach

Assume a coarse grid is defined for $x < 0$, and a fine one – for $x > 0$. It means that the following finite difference scheme is used:

$$\begin{aligned}
&\frac{p_j^n - p_j^{n-1}}{\tau} = \frac{u_{j+1/2}^{n-1/2} - u_{j-1/2}^{n-1/2}}{h}, && j < 0,\\
&\frac{u_{j+1/2}^{n+1/2} - u_{j+1/2}^{n-1/2}}{\tau} = \frac{p_{j+1}^n - p_j^n}{h}, && j < 0,\\
&\frac{p_{j+l/L}^{n+k/K} - p_{j+l/L}^{n+(k-1)/K}}{\tau/K} = \frac{u_{j+(2l-1)/2L}^{n+(2k-1)/2K} - u_{j+(2l-3)/2L}^{n+(2k-1)/2K}}{h/L}, && j > 0,\\
&\frac{u_{j+(2l-1)/2L}^{n+(2k-1)/2K} - u_{j+(2l-1)/2L}^{n+(2k-3)/2K}}{\tau/K} = \frac{p_{j+l/L}^{n+(k-1)/K} - p_{j+(l-1)/L}^{n+(k-1)/K}}{h/L}, && j > 0.
\end{aligned} \tag{3}$$

We denote the refinement ratios with respect to time and space as K and L respectively. Parameters k and l defining fractional points at a fine grid are $k = 0, ..., K-1$ and $l = 0, ..., L-1$. Case $K = 1$ means that no refinement with respect to time is applied, $L = 1$ – no refinement with respect to space.

As it follows from (3) one can compute solution at all points except the interface $x = 0$ or $j = 0$. To construct finite difference conjugation conditions at the interface we suggest to use an algorithm based on approximation of original system of equations together with approximation of second order equation. To describe the algorithm let us consider the two cases separately.

1. Updating solution at the instants from n to $n + 1/2$. In terms of superscripts introduced in (3) it means that we seek for the solution $p_0^{n+k/K}$ for $k = 1, ..., (K-1)/2$. The formulae based on the approximation of the first order equation $\frac{\partial p}{\partial t} = \frac{\partial u}{\partial x}$ are not applicable in this case because they require variable u to be defined at fractional instants at the coarse grid. On the other hand one may exclude variable u from the system (1) to obtain a second order equation with respect to pressure p and construct corresponding finite difference scheme to approximate it. Below we provide the equation and the scheme:

$$\frac{\partial^2 p}{\partial t^2} = \frac{\partial^2 p}{\partial x^2}; \quad \frac{p_J^{N+1} - 2p_J^N + p_J^{N-1}}{(\delta t)^2} = \frac{p_{J+1}^N - 2p_J^N + p_{J-1}^N}{(\delta x)^2}.$$

   We write indices in capital letters and use different notations for the grid steps to point out that the finite difference scheme approximates the equation with the second order on any equidistant grid not necessarily connected with that used in (3).

   To adopt the scheme for the second order equation to compute solution $p_0^{n+k/K}$, for $k = 1, ..., (K-1)/2$ one needs to rewrite it as follows:

$$\frac{p_0^{n+k/K} - 2p_0^n + p_0^{n-k/K}}{(\tau k/K)^2} = \frac{p_1^n - 2p_0^n + p_{-1}^n}{h^2}. \tag{4}$$

   The obtained formula involves three values of pressure, defined at instant n, i.e., p_0^n, $p_{\pm 1}^n$ and the value $p_0^{n-k/K}$ defined at the interface but at the instant below n. Values $p_{\pm 1}^n$ can be computed by the formulae (3), the rule to calculate p_0^n and $p_0^{n-k/K}$ are being formulated below. Graphical representation of the points involved in the computations of solution at the interface at instants from integer to half-integer are provided in Fig. 1 left.
2. Updating the solution at the interface at instants $n + 1/2$ to $n + 1$. One may apply approximation of the first order equation to compute the solution in this case. As variables $p_0^{n+k/K}$ for $k = 1, \ldots, (N-1)/2$ are supposed to have been computed by means of (4) one is able to use them to construct finite difference approximation:

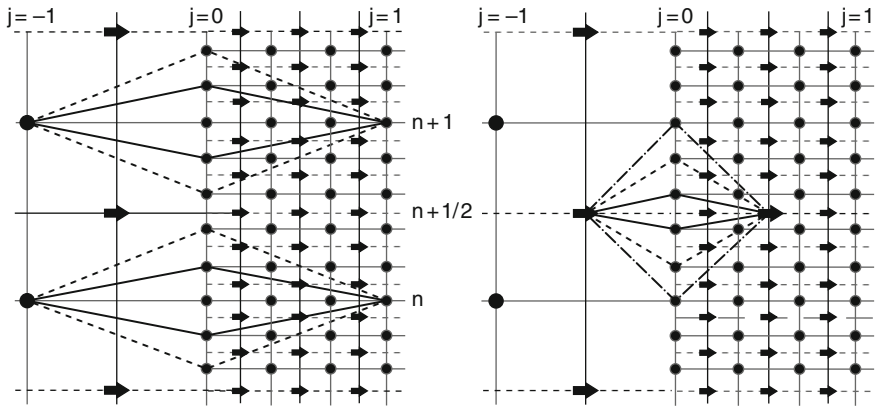


Fig. 1 Points involved in computations to update a solution at the interface. Approximation of second order equation at the *left* (to update solution from integer to half-integer instants), approximation of the first order equation at the *right* (to compute solution form half-integer to integer instants)

$$\frac{p_0^{n+1/2+(2k+1)/2K} - p_0^{n+1/2-(2k+1)/2K}}{(2k+1)\tau/K} = \frac{u_{1/2}^{n+1/2} - u_{-1/2}^{n+1/2}}{h}, \qquad (5)$$

for $k = 0, ..., (K-1)/2$. Note, that variables $u_{\pm 1/2}^{n+1/2}$ can be computed by the scheme (3).

Note that each scheme (3), (4), and (5) was constructed on a symmetric stencil even though different temporal steps were used for each k in formulae (4) and (5). It means that the constructed finite difference scheme approximates (1) with the second order. In addition the described approach may be applied for different refinement ratios L and K of spatial and temporal steps. In particular the scheme (3), (4), and (5) is valid for $K = 1$, i.e., for spatial refinement only, and for $L = 1$ – temporal refinement only. Moreover, if $K = L = 1$ the scheme turns to the scheme (2) which is free from refinement.

3.1 Reflectivity

In order to study reflection coefficients caused by grid refinement one needs to seek for a plane wave solution assuming that having approached the interface a plane wave gives rise to a reflected wave and a transmitted wave. We are not going to go into the ins and outs of the construction of reflection and transmission coefficients for grid refinement as it can be found in [2,4]. So, we applied the technique from [2] to our approach. We proved that the reflection coefficients converges to zero with the second order as grid steps tend to zero. In addition we considered frequency

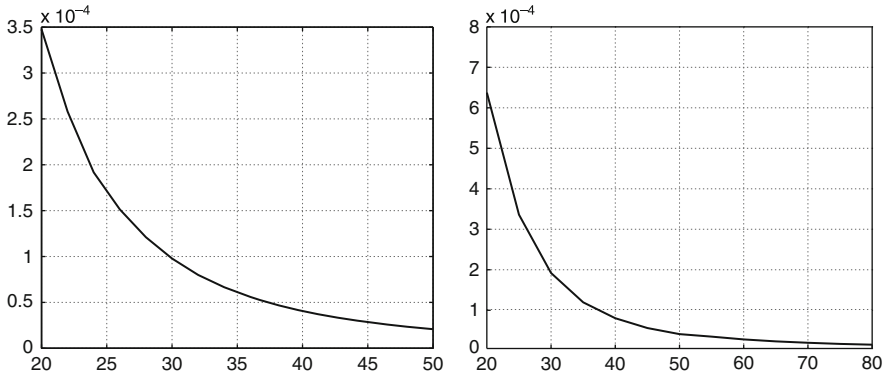


Fig. 2 Reflection coefficients for temporal (*left*) and spatial (*right*) mesh refinement in dependence on ppw for coarse grid. Courant ratio was fixed

dependent reflection coefficients and applied inverse Fourier transform to the source wavelet multiplied by the reflection coefficients to derive the reflections in physical space. So, the obtained reflections were of about 10^{-4} of incident wave.

To verify the results a series of the numerical simulations were done. We generated a wave propagating from a coarse grid to fine one and vice versa and measured the reflections appearing as if the wave passed the interface. After that the reflections coefficient was constructed as a ratio of maximum of reflected wave to that of incident wave. Figure 2 represents the reflection coefficients depending on a number of grid points per wavelength for temporal mesh refinement $K = 3$ (left) and spatial refinement $L = 3$ (right). The Courant ratio was 0.9. So, one can see that the reflections are about 10^{-4}, and decay with the second order as the ppw increases.

3.2 Stability

There are two different types of instability associated with local mesh refinement. The first is caused by ill-possedness of conjugation conditions. It means that a particular condition may possess exponentially growing modes. This type of instability is easy to determine and avoid on the basis of Kreiss-Sakamoto-Lopatinski determinant as the theoretical background is rather well developed, see [6–8] and others. We proved analytically, on the basis of Kreiss–Sakamoto–Lopatinski determinant, that the approach proposed in this paper does not possess this type of instability. It means that the problem (3)–(5) is stable if stated in infinite domain with constant velocity model.

On the other hand, if the domain is bounded or if there are discontinuities of the velocity model grid refinement may cause the other type of instability associated with multiple wave passing though the interface. It is discussed in details in [2] for the method based on interpolation with respect to time. To study this type of

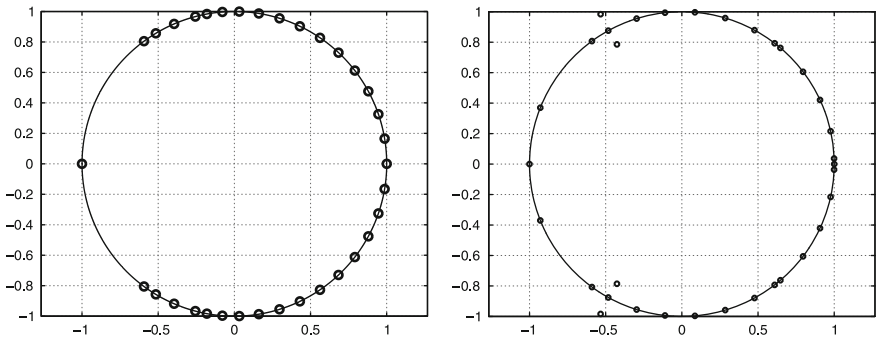


Fig. 3 Eigenvalues of the transition operator for successive refinement (*left*) and simultaneous refinement (*right*) on a complex plane

instability we applied to the spectral necessary stability condition of finite difference scheme approximating an initial-boundary value problem:

Theorem 1. *If a finite difference scheme is stable all the eigenvalues* λ *of the operator mapping data from one instant to the next one satisfy the inequality:* $|\lambda| \leq 1$.

So, we considered the scheme (3) given at the interval $j = -J, \ldots, J$ with zero boundary conditions. The interface of grid refinement was $j = 0$. According to the stability condition any simultaneous mesh refinement of a presented type, i.e., $K \neq 1$ and $L \neq 1$ is unstable. Figure 3 represents the eigenvalues of the operator for the case $K = L = 3$. One can see that there are eigenvalues outside the unit circle.

On the contrary, if only temporal or only spatial steps are refined, i.e., if $K = 1$ or $L = 1$, the eigenvalues belong to the unit circle, for all Courant ratios less than one, see Fig. 3. The same behavior is observed for the successive refinement, i.e., the refinement of time steps is performed at the interface $j = j_t$ and spatial steps are refined at the interface $j = j_s$, where $j_t < j_s$. So, the approach to successive local mesh refinement based on the approximation of wave equation at the interface satisfies the necessary stability criteria of finite difference scheme.

In order to study stability of the algorithm we simulated a wave propagation within interval $[0, 10\lambda]$, where λ is a wavelength. Time interval was 10^6 wave periods. The interface of refinement was 5λ in case of simultaneous refinement. In case of successive refinement we had two interfaces 5λ and 5.5λ for temporal and spatial step refinement respectively. In case of simultaneous refinement energy growth took place at rather low instants of about 10 wave periods. At the same time no energy growth was observed for successive refinement.

4 Conclusions

In this paper we provided a new algorithm of local spacio-temporal mesh refinement on the basis of approximation of the wave equation. This approach is free from interpolation or projection techniques thus it preserves the second order of convergence that was proved analytically and confirmed by the numerical experiments. Presented approach was designed to be applicable to simulations of scattered waves. Hence, one of the main requirement was low level of artificial reflections caused by the refinement. According to the theoretical study and numerical experiments reflection coefficients possessed by the approach are about 10^{-4} that is comparable with that caused by numerical dispersion of the scheme. We also proved that the finite difference scheme with successively refined grid (refinement of temporal and spatial steps are separated from each other) satisfies the necessary spectral stability criterion, while simultaneous one is unstable. This fact was also confirmed by a series of numerical simulations.

Acknowledgements This research was done under financial support RFBR grants 08-05-00265, 10-01-92604, 10-05-00233, 10-05-00337.

References

1. Cohen, G. (ed.): Metodes numeriques d'ordre eleve pour les ondes en regime transitoire. INRIA (1994) In French
2. Collino, F., Fouquet, T., Joly, P.: Analyse numerique d'une methode de raffinement de maillage espace-temps pour l'equation des ondes. INRIA Rapprot de recherche **3474** (1998)
3. Collino, F., Fouquet, T., Joly, P.: A conservative space-time mesh refinement method for the 1-D wave equation. Part I: Construction. Numer. Math. **95**, 197–221 (2003)
4. Collino, F., Fouquet, T., Joly, P.: A conservative space-time mesh refinement method for the 1-D wave equation. Part II: Analysis. Numer. Math. **95**, 223–251 (2003)
5. Kim, S.I., Hoefer, W.J.R.: A local mesh refinement algorithm for the time-domain finite-difference method to solve Maxvell's equations. IEEE Trans. Microw. Theor. Tech. **38**, 812–815 (1990)
6. Kreiss, H.O.: Initial boundary value problems for hyperbolic systems. Comm. Pure Appl. Math. **23**, 277–298 (1970)
7. Sakamoto, R.: Mixed problems for hyperbolic equations. I. Energy inequalities. J. Math. Kyoto Univ. **10**, 349–373 (1970)
8. Sakamoto, R.: Mixed problems for hyperbolic equations. II. Existence theorems with zero initial datas and energy inequalities with initial datas. J. Math. Kyoto Univ. **10**, 403–417 (1970)

Formulation of a Staggered Two-Dimensional Lagrangian Scheme by Means of Cell-Centered Approximate Riemann Solver

R. Loubère, P.-H. Maire, and P. Váchal

Abstract In this work we develop a general framework to derive and analyze staggered numerical scheme devoted to solve hydrodynamics equations.

1 Introduction

In this work we develop a general framework to derive and analyze staggered numerical scheme devoted to solve hydrodynamics equations. This framework creates a link between well-known staggered Lagrangian schemes (see [1] and references therein) and new cell-centered ones [3]. After the governing equations and notation are set, the framework is presented. Discretization following physical principles is obtained through the use of fundamental objects: subcell forces. We will show that the Geometric Conservation Law [3] is compatible with the trajectory equation. Momentum and total energy conservation will also be satisfied. Moreover, the definition of total energy and its conservation will uniquely imply the discretization of internal energy equation. Finally, in order to ensure isentropic consistency subcell forces will be equipped with a viscous part that fulfills the second law of thermodynamics. This implies the form of the viscous force as a multiplication of a positive definite matrix and the difference between cell-centered and nodal velocity. This matrix is the corner stone to construct a Lagrangian staggered scheme. Within this

R. Loubère (✉)
Université de Toulouse, UPS, IMT CNRS, France
e-mail: raphael.loubere@math.univ-toulouse.fr

P.-H. Maire
UMR CELIA, CNRS, Université Bordeaux 1, CEA, France
e-mail: maire@celia.u-bordeaux1.fr

P. Váchal
CVUT in Prague Czech Republic
e-mail: vachal@galileo.fjfi.cvut.cz

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_66,

general framework we completely describe the derivation of a Lagrangian staggered scheme that is finally tested on classical 2D test cases.

2 Governing Equations and Notation

In Lagrangian framework, 2D gas dynamics equations write

$$\rho \frac{d}{dt}\left(\frac{1}{\rho}\right) - \nabla \cdot \boldsymbol{U} = 0, \qquad \rho \frac{d}{dt}\boldsymbol{U} + \boldsymbol{\nabla} P = \boldsymbol{0}, \qquad \rho \frac{d}{dt} E + \nabla \cdot (P\boldsymbol{U}) = 0, \quad (1)$$

where ρ is the density, $\boldsymbol{U}$ the velocity and E the total energy. The first equation expresses the volume conservation equation, whereas the second and third ones are the momentum-total energy conservation equations. Volume conservation equation is often referred to as the Geometrical Conservation Law (GCL).

The previous system is equipped with a thermodynamics closure (equation of state) $P = P(\rho, \varepsilon)$ where the specific internal energy is given by $\varepsilon = E - \frac{U^2}{2}$. Note that for smooth solutions energy equation can be rewritten as

$$\rho \frac{d}{dt}\varepsilon + P \nabla \cdot \boldsymbol{U} = 0, \quad (2)$$

and, substituting volume equation yields $\rho \frac{d}{dt}\varepsilon + P\rho \frac{d}{dt}\left(\frac{1}{\rho}\right) = 0$. Recalling Gibbs relation for temperature T and specific entropy S: $T dS = d\varepsilon + P d\left(\frac{1}{\rho}\right)$, and the second law of thermodynamics, namely $T \frac{dS}{dt} \geq 0$, implies that for non smooth flows the following relation holds:

$$\rho \frac{d}{dt}\varepsilon + P \nabla \cdot \boldsymbol{U} \geq 0. \quad (3)$$

As a consequence, internal energy equation can be viewed as an entropy evolution equation as $\rho \frac{d}{dt}\varepsilon + P\rho \frac{d}{dt}\left(\frac{1}{\rho}\right) \geq 0$. The previous system (1) can therefore be rewritten as a non-conservative system by replacing the energy equation by (3).
The last equations are the trajectory equations

$$\frac{d\boldsymbol{X}}{dt} = \boldsymbol{U}(\boldsymbol{X}(t), t), \quad \boldsymbol{X}(0) = \boldsymbol{x}, \quad (4)$$

expressing the Lagrangian motion of any point initially located at position $\boldsymbol{x}$.

We use a staggered discretization (see Fig. 1). Position and velocity are defined at grid points while thermodynamical variables are located at cell centers. An

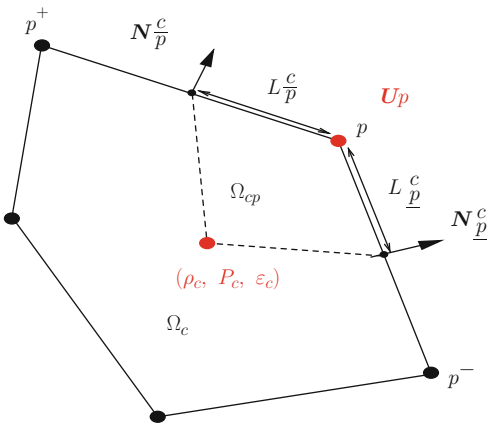


Fig. 1 Notation used in the derivation of the framework. Position and velocity are defined at grid points while thermodynamical variables are located at cell centers. A polygonal cell Ω_c is subdivided into subcells Ω_{cp}. Half-edge lengths: $L_{\underline{p}}^{c}$, $L_{\overline{p}}^{c}$, unit normal vectors $\boldsymbol{N}_{\underline{p}}^{c}$, $\boldsymbol{N}_{\overline{p}}^{c}$

unstructured grid consisting of a collection of non-overlapping polygons is considered. Each polygonal cell is assigned a unique index c and denoted Ω_c. Each vertex/point of the mesh is assigned a unique index p and we denote $\mathscr{C}(p)$ the set of cells sharing a particular vertex p. A polygonal cell is subdivided into a set of subcells; each being uniquely defined by a pair of indices c and p and denoted Ω_{cp}. This subcell is constructed by connecting the cell center of Ω_c to the mid-points of cell edges impinging on point p. The union of subcells Ω_{cp} that share a particular vertex p allows to define the (dual) vertex-centered cell Ω_p related to point p with $\Omega_p = \bigcup_{c\in\mathscr{C}(p)} \Omega_{cp}$. This defines the primary grid $\bigcup_c \Omega_c$ and the dual grid $\bigcup_p \Omega_p$. Primary cells Ω_c and dual cells Ω_p volumes are functions of time t. *We make the fundamental assumption that the subcells are Lagrangian volumes*. Namely the subcell mass m_{cp} is constant in time; knowing initial density field $\rho^0(\boldsymbol{X})$ one introduces the initial mean density in cell c as $\rho_c^0 = \int_{\Omega_c(0)} \rho^0(\boldsymbol{X}) d\boldsymbol{X}/V_c^0$, where V_c^0 is the volume of cell Ω_c at time $t = 0$. Subcell mass is defined as $m_{cp} = \rho_c^0 V_{cp}^0$ where V_{cp}^0 are the initial volume of subcell Ω_{cp}. By summation of Lagrangian subcell masses one defines Lagrangian cell/point masses as $m_c = \sum_{p\in\mathscr{P}(c)} m_{cp}$, and $m_p = \sum_{c\in\mathscr{C}(p)} m_{cp}$, where $\mathscr{P}(c)$ is the set of counterclockwise ordered vertices of cell c.

3 Framework and Discretization

Since velocity is defined at point p, the GCL is satisfied and the volume equation writes

$$\frac{d}{dt} V_c = \sum_{p\in\mathscr{P}(c)} \left(L_{\underline{p}}^{c} \boldsymbol{N}_{\underline{p}}^{c} + L_{\overline{p}}^{c} \boldsymbol{N}_{\overline{p}}^{c} \right) \cdot \boldsymbol{U}_p, \tag{5}$$

where V_c is the volume of cell c and $\boldsymbol{U}_p$ corresponds to the point velocity, so that trajectory equation (4) yields $\frac{d}{dt}\boldsymbol{X}_p = \boldsymbol{U}_p$. The previous discretization is obtained

by time differentiation of $V_c(t)$, it allows to define the discrete divergence operator over cell c as

$$(\nabla \cdot \boldsymbol{U})_c = \frac{1}{V_c} \sum_{p \in \mathscr{P}(c)} L_{pc} \boldsymbol{N}_{pc} \cdot \boldsymbol{U}_p, \tag{6}$$

where $\boldsymbol{N}_{pc}$ is the unit corner vector defined by $L_{pc}\boldsymbol{N}_{pc} = L^c_{\underline{p}}\boldsymbol{N}^c_{\underline{p}} + L^c_{\overline{p}}\boldsymbol{N}^c_{\overline{p}}$. Equation (5) is compatible with the discrete version of the trajectory equation (4), that is to say $\frac{d}{dt}\boldsymbol{X}_p = \boldsymbol{U}_p, \quad \boldsymbol{X}_p(0) = \boldsymbol{x}_p$, where $\boldsymbol{U}_p \equiv \boldsymbol{U}_p(t) = \boldsymbol{U}_p(\boldsymbol{X}_p(t), t)$ is the point velocity.

Let $J(\boldsymbol{x}, t)$ be the Jacobian of the transformation $\boldsymbol{x} \mapsto \boldsymbol{X}(\boldsymbol{x}, t)$, defined through the trajectory equation (4). Integrating the momentum equation over dual cell $\Omega_p(0)$ and using $J\,\mathrm{d}\boldsymbol{x} = \mathrm{d}\boldsymbol{X}$ leads to

$$\frac{d}{dt}\int_{\Omega_p(0)} \rho^0(\boldsymbol{x})\boldsymbol{U}(\boldsymbol{x}, t)\,\mathrm{d}\boldsymbol{x} + \int_{\Omega_p(t)} \nabla P\,\mathrm{d}\boldsymbol{X} = \boldsymbol{0}. \tag{7}$$

Finally, using Green formula one replaces the second integral in previous equation by $\int_{\partial\Omega_p(t)} P\boldsymbol{N}\,\mathrm{d}l$. Moreover, recalling that $m_p = \int_{\Omega_p(0)} \rho^0(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$, we introduce a finite volume interpretation of $\boldsymbol{U}_p(t)$ as

$$\boldsymbol{U}_p(t) = \frac{1}{m_p}\int_{\Omega_p(0)} \rho^0(\boldsymbol{x})\boldsymbol{U}(\boldsymbol{x}, t)\,\mathrm{d}\boldsymbol{x} = \frac{1}{m_p}\int_{\Omega_p(t)} \rho(\boldsymbol{X}, t)\boldsymbol{U}(\boldsymbol{X}, t)\,\mathrm{d}\boldsymbol{X}. \tag{8}$$

Finally momentum equation is semi-discretized in space over the dual cell Ω_p as

$$m_p \frac{d}{dt}\boldsymbol{U}_p + \sum_{c \in \mathscr{C}(p)} \boldsymbol{F}_{cp} = \boldsymbol{0}, \tag{9}$$

where $\boldsymbol{F}_{cp}$ is a fundamental object called *subcell force* from cell c that acts on point p and is defined as

$$\boldsymbol{F}_{cp} = \int_{\partial\Omega_p(t)\cap\Omega_c(t)} P\boldsymbol{N}\,\mathrm{d}l, \tag{10}$$

so that $\sum_{c\in\mathscr{C}(p)} \boldsymbol{F}_{cp} = \int_{\partial\Omega_p} P\boldsymbol{N}\,\mathrm{d}l$. Total momentum conservation (away from boundary conditions) is ensured provided subcell forces fulfill $\sum_{p\in\mathscr{P}(c)} \boldsymbol{F}_{cp} = \boldsymbol{0}$. The proof is left to the reader.

Contrary to cell-centered approach [3], total energy equation is not discretized. Here, we derive a semi discrete internal energy equation that ensures total energy conservation. Away from boundary conditions, we introduce total kinetic at time $t > 0$ as a sum over the dual cells $\mathscr{K}(t) = \sum_p \frac{1}{2} m_p \boldsymbol{U}_p^2(t)$ and internal energy as $\mathscr{E}(t) = \sum_c m_c \varepsilon_c(t)$, where ε_c is the cell averaged internal energy. Conservation of total energy writes $\frac{d}{dt}\mathscr{K} + \frac{d}{dt}\mathscr{E} = 0$. As cell/point masses are Lagrangian objects one gets

$$\sum_c m_c \frac{d}{dt}\varepsilon_c + \sum_p m_p \frac{d}{dt}\boldsymbol{U}_p \cdot \boldsymbol{U}_p = \sum_c m_c \frac{d}{dt}\varepsilon_c - \sum_p \sum_{c\in\mathscr{C}(p)} \boldsymbol{F}_{cp}\cdot\boldsymbol{U}_p = 0,$$

by shifting sums and using discrete momentum equation (9) one gets

$$\sum_c \left(m_c \frac{d}{dt}\varepsilon_c - \sum_{p\in\mathscr{P}(c)} \boldsymbol{F}_{cp}\cdot\boldsymbol{U}_p \right) = 0. \tag{11}$$

A sufficient condition for total energy conservation is gained by requiring the previous equation to hold in each cell c:

$$m_c \frac{d}{dt}\varepsilon_c - \sum_{p\in\mathscr{P}(c)} \boldsymbol{F}_{cp}\cdot\boldsymbol{U}_p = 0. \tag{12}$$

Once the subcell force is known then momentum and internal energy can be updated using (9) and (12). As previously seen, the subcell force formalism ensures total energy conservation.

Using Gibbs formula, the time rate of change of entropy in cell c writes $m_c T_c \frac{d}{dt}S_c = m_c \left[\frac{d}{dt}\varepsilon_c + P_c \frac{d}{dt}\left(\frac{1}{\rho_c}\right)\right]$. Substituting time rate of change of internal energy (12) and volume leads to

$$m_c T_c \frac{d}{dt}S_c = \sum_{p\in\mathscr{P}(c)} \left(\boldsymbol{F}_{cp} + P_c L_{cp}\boldsymbol{N}_{cp}\right)\cdot\boldsymbol{U}_p. \tag{13}$$

For smooth flow entropy must be conserved. This requirement leads to the following decomposition of subcell force: $\boldsymbol{F}_{cp} = -P_c L_{cp}\boldsymbol{N}_{cp} + \boldsymbol{F}_{cp}^{\text{viscous}}$, the substitution of which in (12) yields

$$m_c \left[\frac{d}{dt}\varepsilon_c + P_c \frac{d}{dt}\left(\frac{1}{\rho_c}\right)\right] = \sum_{p\in\mathscr{C}(p)} \boldsymbol{F}_{cp}^{\text{viscous}}\cdot\boldsymbol{U}_p. \tag{14}$$

In order to satisfy second law of thermodynamics one must require that subcell viscous forces satisfy $\sum_{p\in\mathscr{P}(c)} \boldsymbol{F}_{cp}^{\text{viscous}}\cdot\boldsymbol{U}_p \geq 0$, and viscous forces must vanish for smooth flows (e.g., rarefaction, isentropic compression). As previously shown momentum conservation requires $\sum_{p\in\mathscr{P}(c)} \boldsymbol{F}_{cp} = \boldsymbol{0}$. Knowing the geometrical relation $\sum_{p\in\mathscr{P}(c)} L_{cp}\boldsymbol{N}_{cp} = \boldsymbol{0}$ we get $\sum_{p\in\mathscr{P}(c)} \boldsymbol{F}_{cp}^{\text{viscous}} = \boldsymbol{0}$.

The computation of subcell force is performed through a cell-centered approximation of the multidimensional Riemann problem. We introduce two pressures at the cell center called $\Pi_{\underline{c}}^p, \Pi_{\overline{c}}^p$ which are related to the unit outward normals $\boldsymbol{N}_{\underline{c}}^p$ and $\boldsymbol{N}_{\overline{c}}^p$ respectively. The subcell force is then defined as

$$\boldsymbol{F}_{cp} = L_{\underline{c}}^{p} \Pi_{\underline{c}}^{p} \boldsymbol{N}_{\underline{c}}^{p} + L_{\overline{c}}^{p} \Pi_{\overline{c}}^{p} \boldsymbol{N}_{\overline{c}}^{p}. \tag{15}$$

The pressures are obtained by means of the half-Riemann problems

$$P_c - \Pi_{\underline{c}}^{p} = Z_{\underline{c}}^{p} \left(\boldsymbol{U}_c - \boldsymbol{U}_p \right) \cdot \boldsymbol{N}_{\underline{c}}^{p}, \qquad P_c - \Pi_{\overline{c}}^{p} = Z_{\overline{c}}^{p} \left(\boldsymbol{U}_c - \boldsymbol{U}_p \right) \cdot \boldsymbol{N}_{\overline{c}}^{p}, \tag{16}$$

where $Z_{\underline{c}}^{p}$, $Z_{\overline{c}}^{p}$ denote the swept mass fluxes, and $\boldsymbol{U}_c$ is the cell-centered velocity which remains to be defined. Using (16) the subcell force is rewritten

$$\boldsymbol{F}_{cp} = \left(L_{\underline{c}}^{p} \boldsymbol{N}_{\underline{c}}^{p} + L_{\overline{c}}^{p} \boldsymbol{N}_{\overline{c}}^{p} \right) P_c - \mathsf{M}_{cp} \left(\boldsymbol{U}_c - \boldsymbol{U}_p \right), \tag{17}$$

where $\mathsf{M}_{cp} = Z_{\underline{c}}^{p} L_{\underline{c}}^{p} (\boldsymbol{N}_{\underline{c}}^{p} \otimes \boldsymbol{N}_{\underline{c}}^{p}) + Z_{\overline{c}}^{p} L_{\overline{c}}^{p} (\boldsymbol{N}_{\overline{c}}^{p} \otimes \boldsymbol{N}_{\overline{c}}^{p})$, is a 2×2 symmetric positive definite matrix. As a consequence the viscous part of the force writes $\boldsymbol{F}_{cp}^{\text{viscous}} = -\mathsf{M}_{cp} \left(\boldsymbol{U}_c - \boldsymbol{U}_p \right)$. Recalling that $L_{\underline{c}}^{p} \boldsymbol{N}_{\underline{c}}^{p} + L_{\overline{c}}^{p} \boldsymbol{N}_{\overline{c}}^{p} = -L_{cp} \boldsymbol{N}_{cp}$ we observe that the first part of the subcell force is compatible with the entropy conservation. The cell center velocity is determined through the use of momentum conservation as

$$\sum_{p \in \mathscr{P}(c)} \boldsymbol{F}_{cp} = \boldsymbol{0} \Longleftrightarrow \boldsymbol{U}_c = \mathsf{M}_c^{-1} \sum_{p \in \mathscr{P}(c)} \mathsf{M}_{cp} \boldsymbol{U}_p,$$

where $\mathsf{M}_c = \sum_{p \in \mathscr{P}(c)} \mathsf{M}_{cp}$. The previous equation is a 2×2 non-linear system which can be solved utilizing an iterative algorithm. The non-linearity comes from the swept mass fluxes, that, following Dukowicz [2, 3], one approximates as

$$Z_{\underline{c}}^{p} = \rho_c \left[\sigma_c + \Gamma_c \mid (\boldsymbol{U}_c - \boldsymbol{U}_p) \cdot \boldsymbol{N}_{\underline{c}}^{p} \mid \right], \qquad Z_{\overline{c}}^{p} = \rho_c \left[\sigma_c + \Gamma_c \mid (\boldsymbol{U}_c - \boldsymbol{U}_p) \cdot \boldsymbol{N}_{\overline{c}}^{p} \mid \right]. \tag{18}$$

Here, σ_c is the isentropic sound speed and Γ_c a material dependent coefficient, which for a γ gas law is defined by $\Gamma_c = \begin{cases} \frac{\gamma+1}{2} & \text{if } (\nabla \cdot \boldsymbol{U})_{cp} < 0, \\ 0 & \text{if } (\nabla \cdot \boldsymbol{U})_{cp} \geq 0, \end{cases}$ where $(\nabla \cdot \boldsymbol{U})_{cp} = -\frac{1}{V_{cp}} L_{cp} \boldsymbol{N}_{cp} \cdot (\boldsymbol{U}_c - \boldsymbol{U}_p)$, is the sub-cell contribution to the velocity divergence. In case of rarefaction we recover the acoustic approximation whereas in case of shock wave we get a two-shock approximation.

Once $\boldsymbol{U}_c$ is known, the subcell force is obtained with (17). Entropy inequality is fulfilled as the entropy production writes $\sum_{p \in \mathscr{C}(p)} -\mathsf{M}_{cp} \left(\boldsymbol{U}_c - \boldsymbol{U}_p \right) \cdot \left(\boldsymbol{U}_c - \boldsymbol{U}_p \right) \geq 0$.

4 Numerics

The first test problem is the classical 1D Sod shock tube. Figure 2 shows the cell-centered density (markers) *vs* the exact solution (line). The next test is the 2D Noh problem (see [4]). A cold gas with unit density is given an initial inward radial

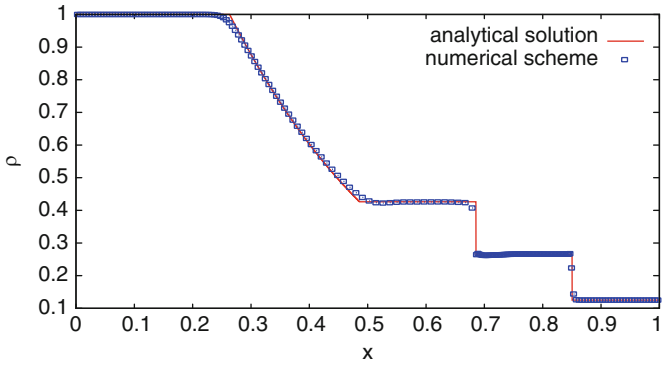


Fig. 2 1D Sod shock tube problem at $t = 0.2$ with 200 cells (*markers*) vs. the exact solution (*line*). On [0, 1]: Left state $(\rho_L, u_L, p_L) = (1, 0, 1)$, right state $(0.125, 0, 0.1)$, perfect gas $\gamma = 7/5$ discontinuity at $\boldsymbol{X} = 0.5$. Symmetry boundary conditions

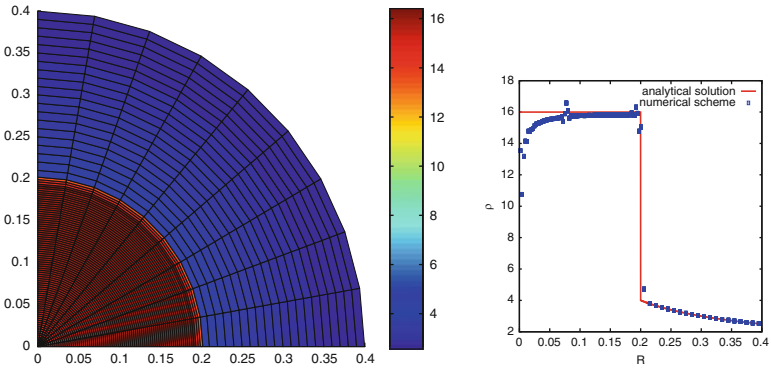


Fig. 3 Noh problem for a polar grid at $t = 0.6$. *Left*: Density map and final mesh. *Right*: Density as a function of radius for all cells vs. the exact solution (*line*)

velocity of magnitude 1. Then, a diverging cylindrical shock wave which propagates at speed $1/3$ is generated. A 100×9 polar grid is used. Figure 3 presents the mesh and density map (left) and the density as a function of cell radius (right) for all cells. Then, a non-conformal initial grid consisting of 1, 700 cells (triangles, quadrangles, pentagons) is considered (it can be guessed from Fig. 4-left). A zoom on the final grid at $t = 0.6$ is presented in Fig. 4 (left panel) and the density as function of radius in the right panel. No special treatment is required to correctly perform on such a mesh. Finally we present numerical results for the Sedov blast wave problem [5], which describes the evolution of a blast wave in a point symmetric explosion (total energy is concentrated at the origin with $E_{\text{total}} = 0.244816$). An ideal gas with $\gamma = 1.4$ initially at rest with a density equal to 1 is considered. The final time is 1. Two meshes are considered; a 30×30 Cartesian mesh and a polygonal mesh (Voronoï tessellation). Figure 5 shows the final mesh and density map (left) and the

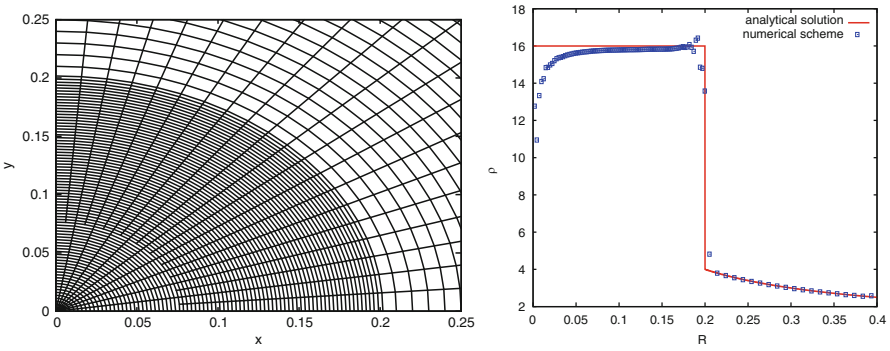


Fig. 4 Noh problem for a non-conformal mesh. *Left*: Zoom on the final mesh. *Right*: density as a function of radius for all cells *vs* the exact solution (*line*)

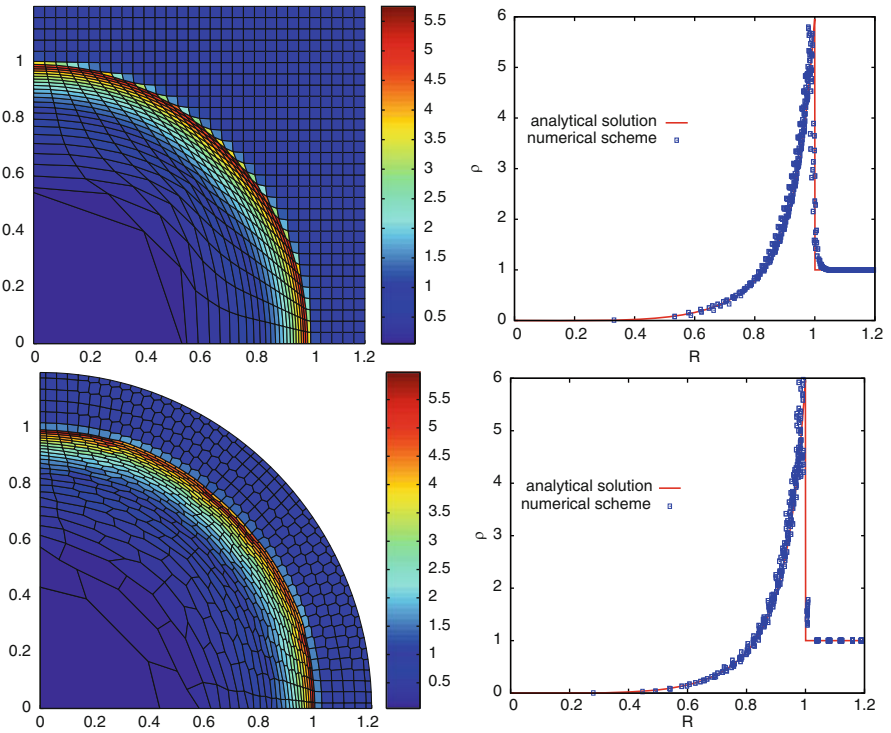


Fig. 5 Sedov problem at $t = 1$. *Top*: Density and mesh (*left*) and density as function of radius (*right*) for a 30×30 Cartesian grid. *Bottom*: Density and mesh (*left*) and density as function of radius (*right*) for a polygonal grid (Voronoï tessellation)

density as function of cell radius (right) demonstrating the capability of the method to handle polygonal grids.

5 Conclusion

We have presented a general abstract framework to develop staggered schemes for Lagrangian hydrodynamics equations. This framework is based on fundamental objects: Lagrangian subcell mass and subcell force. We provided an example of such a scheme. We presented several test cases showing the efficiency and pertinence of the proposed approach. Within this framework one expects to better analyze the similarity with the more recent cell-centered Lagrangian schemes, see [3] as well as with classical staggered schemes, see [1] and references therein.

References

1. A.L. Bauer, D.E. Burton, E.J. Caramana, R. Loubere, M.J. Shashkov, P.P. Whalen, The internal consistency, stability, and accuracy of the discrete, compatible formulation of Lagrangian hydrodynamics, J. Comp. Phys. **218**(2), 572–593 (2006)
2. J.K. Dukowicz, A general, non-iterative Riemann solver for Godunov's method, J. Comp. Phys. **61**(1), 119–137 (1985)
3. P.-H. Maire, A high-order cell-centered Lagrangian scheme for two-dimensional compressible fluid flows on unstructured mesh, J. Comp. Phys. **228**(7), 2391–2425 (2009)
4. W.F. Noh, Errors for calculations of strong shocks using artificial viscosity and an artificial heat flux, J. Comp. Phys. **72**, 78–120 (1987)
5. L.I. Sedov, *Similarity and Dimensional Methods in Mechanics*, Academic Press, New York (1959)

Optimal Control for River Pollution Remediation

Aurea Martínez, Lino J. Alvarez-Vázquez, Miguel E. Vázquez-Méndez, and Miguel A. Vilar

Abstract The main goal of this work is to use mathematical modelling and numerical optimization to obtain the optimal purification of a polluted section of a river. The most common strategy consists of the injection of clear water from a reservoir in a nearby point. In this process, the main problem consists, once the injection point is selected by geophysical reasons, of finding the minimum quantity of water which is needed to be injected into the river in order to purify it up to a fixed level: this will be the aim of this paper. We formulate this problem as a hyperbolic optimal control problem with control constraints, and deal with its numerical resolution, where a finite elements/finite differences discretization is used, an optimization algorithm is proposed, and computational results are provided.

1 Introduction

Nowadays, contamination levels in rivers usually exceed desirable thresholds given by legislative rules. In order to palliate these high levels of contamination, one of the most used techniques is based on injecting – from a reservoir close to the river – a great amount of clear water. In this process of increasing the river flow by controlled releases from a reservoir, the main problem consists – once the injection point is selected by geophysical reasons – of finding the minimum quantity of injected water which is needed to purify the river section up to a fixed level.

A. Martínez (✉) and L.J. Alvarez-Vázquez
Departamento de Matemática Aplicada II, E.T.S.I. Telecomunicación, Universidad de Vigo, 36310 Vigo, Spain
e-mail: lino@dma.uvigo.es, aurea@dma.uvigo.es

M.E. Vázquez-Méndez and M.A. Vilar
Departamento de Matemática Aplicada, E.P.S., Universidad de Santiago de Compostela, 27002 Lugo, Spain
e-mail: miguelernesto.vazquez@usc.es, miguel.vilar@usc.es

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_67,

In this work, we give the mathematical formulation of this real-world problem (a control problem arisen in the management of a reservoir for the remediation of a polluted river section), and concentrate on its numerical resolution. By using mathematical modeling, the problem is formulated as a hyperbolic optimal control problem with control constraints. Technological reasons demand a time discretization on the control and, by using the method of characteristics, we obtain a semi-discretized problem. Next, we make a finite element/finite difference space discretization of the semi-discrete state system, stating an algorithm to solve the nonlinear resultant problem, and obtain a fully discretized problem. Despite the existence of a number of well-known explicit shallow water solvers, we have preferred – because it fits better our control purposes – using our own solver, whose properties have been theoretically analyzed in [4], and that has shown to be quite accurate in problems previously studied by the authors [2]. Finally, we propose a gradient-free method (the Nelder–Mead algorithm) to solve it, and present numerical results showing the efficiency of the complete algorithm. As previously reported by the authors in [1], Nelder–Mead algorithm is fully competitive (even better in some cases) against other standard gradient-type techniques (adjoint methods, interior-point algorithms...) when dealing with several environmental control problems of geometric nature.

2 The Mathematical Problem

We consider a river L meters in length, with O tributaries (located at points $e_1, \ldots, e_O$) flowing into the river, V wastewater discharges (located at points $v_1, \ldots, v_V$) coming from purifying plants, and one point p where clear water is discharged from a nearby reservoir (a diagram of a realistic example can be seen in Fig. 1).

We are interested in controlling pollution in the river section corresponding to $[p, r]$ (with $p < r \leq L$) over a time interval of T seconds. So, for $(x, t) \in [0, L] \times [0, T]$ we denote by $A(x, t)$ the area of the river section occupied by water (wet section); denote by $u(x, t)$ the average velocity in the wet section; denote by $q(x, t)$ the flow rate across the section (that is, $q(x, t) = A(x, t)u(x, t)$); and denote by $c(x, t)$ the quantity of a generic pollutant in the wet section. The evolution of A, q and c is given by the following hyperbolic initial-boundary value problem in $(0, L) \times (0, T)$:

$$\frac{\partial A}{\partial t} + \frac{\partial q}{\partial x} = Q\delta(x - p) + \underbrace{\sum_{j=1}^{O} q_j \delta(x - e_j) + \sum_{k=1}^{V} p_k \delta(x - v_k)}_{=g_1(x,t)}, \tag{1}$$

$$\frac{\partial q}{\partial t} + \frac{\partial}{\partial x}\left(\frac{q^2}{A}\right) + gA\frac{\partial \eta}{\partial x} = QW\cos(\gamma)\delta(x-p) + \underbrace{\sum_{j=1}^{O} q_j U_j \cos(\alpha_j)\delta(x-e_j) + \sum_{k=1}^{V} p_k V_k \cos(\beta_k)\delta(x-v_k) + S_f}_{=g_2(x,t)}, \tag{2}$$

$$\frac{\partial c}{\partial t} + \frac{\partial}{\partial x}\left(\frac{qc}{A}\right) + kc = \underbrace{\sum_{j=1}^{O} n_j \delta(x-e_j) + \sum_{k=1}^{V} m_k \delta(x-v_k)}_{=g_3(x,t)}, \tag{3}$$

$$A(L,t) = A_L(t), \qquad q(0,t) = q_0(t), \qquad c(0,t) = c_0(t), \tag{4}$$

$$A(x,0) = A^0(x), \qquad q(x,0) = q^0(x), \qquad c(x,0) = c^0(x), \tag{5}$$

where $\delta(x-b)$ denotes de Dirac measure at $b \in [0,L]$; for $j = 1,\ldots,O$, $e_j \in (0,L)$ is the point where the mouth of the jth tributary is located, $q_j(t)$ is the corresponding flow rate, $U_j(t)$ is its velocity, α_j is the angle between the jth tributary and the main river, and $n_j(t)$ is its mass pollutant flow rate; for $k = 1,\ldots,V$, $v_k \in (0,L)$ is the point where the kth wastewater discharge is located, $p_k(t)$ is the corresponding flow rate, $V_k(t)$ is its velocity, β_k is the angle between the kth discharge and the river, and $m_k(t)$ is its mass pollutant flow rate; $p \in (0,L)$ is the point where clear water is discharged, $Q(t)$ is the corresponding flow rate (which will be our control), $W(t)$ is its velocity, and γ is the corresponding angle (it is worthwhile remarking here that, since we are injecting clear water, this term does not appear in the second member of the pollutant equation); g stands for the gravity acceleration; S_f denotes the bottom friction stress, which can be given, for instance, by the Chézy law; $\eta(x,t) = H(x,t) + b(x)$ is the height of water with respect to a fixed reference level, where $H(x,t)$ represents the height of the water column and $b(x)$ geometrically describes the river bottom; and $k(x,t)$ is the loss rate for pollutant.

At first sight we can detect four unknowns in state system (1)–(5): A, q, η and c. Nevertheless, it is obvious that, if river geometry is known, A can be derived from η. In effect, for each $x \in [0,L]$, the geometry of the river gives us a smooth, strictly increasing and positive function $S(.,x)$ verifying $S(0,x) = 0$ and $S(H(x,t),x) = A(x,t)$ in $[0,L]\times[0,T]$. (Specific characterizations of S for usual geometries can be found, for instance, in [4].) So, since we are dealing with a system of balance laws whose conservative variables are A, q and c, if we write η in terms of A, the non-conservative unknown η can be suppressed in (2), which reads now:

$$\frac{\partial q}{\partial t} + \frac{\partial}{\partial x}\left(\frac{q^2}{A}\right) + g\left(\frac{\partial \zeta}{\partial x} - F(A,x) + Ab'(x)\right) = QW\cos(\gamma)\delta(x-p) + g_2, \tag{6}$$

where the detailed characterization of ζ and F can be found in [4].

On the other hand, by technological reasons we are led to consider only the positive fluxes in the set of admissible controls $U_{ad} = \{Q \in L^2(0,T) : 0 \leq Q \leq Q_{max}\}$, since we are just injecting (not extracting) a bounded quantity of clear water.

Finally, in order to formulate the control problem, we consider as the cost functional the total amount of clear water injected through the point p together with a measure (in the region of the river $[p,r]$) of contaminant concentration remaining higher than a fixed threshold c_{max}. So, we define the cost function:

$$J(Q) = \frac{\varepsilon}{2}\int_0^T Q(t)^2\,dt + \frac{1}{2}\int_0^T \int_p^r (c(x,t) - c_{max})_+^2\,dx\,dt \tag{7}$$

where ε is a weight parameter emphasizing the role of the total amount of injected water, and $(c - c_{max})_+$ denotes the positive part of $c - c_{max}$, that is, $(c - c_{max})_+ = \max\{c - c_{max}, 0\}$.

Thus, the problem of the optimal water injection for the purification of a polluted section in a river (denoted by $(\mathscr{P})$) consists of finding the control flux $Q \in U_{ad}$ of injected clear water in such a way that, verifying the state system (1)–(5), minimizes the cost function J given by (7). Thus, the problem can be written in short as:

$$(\mathscr{P}) \qquad \min_{Q \in U_{ad}} J(Q)$$

Questions regarding existence of solution for $(\mathscr{P})$ are still an open problem. In [3] a formal optimality condition was derived by means of adjoint state techniques. Here, we center our attention into numerical resolution of control problem $(\mathscr{P})$.

3 Numerical Discretization

Taking into account technological reasons (flow control mechanisms cannot act upon water flow in a continuous way, but discontinuously at short time periods) we look for admissible controls Q into piecewise-constant functions. So, for the time interval $[0,T]$ we choose a number $K \in \mathbf{N}$, consider the time step $\Delta\tau = T/K$, and define the discrete times $\tau_m = m\Delta\tau$, for $m = 0,\ldots,K$. Thus, a function $Q \in U_{ad}$ which is constant at each subinterval determined by the grid $\{\tau_0,\ldots,\tau_K\}$ is completely fixed by the set of values $Q^{\Delta\tau} = (Q^0,\ldots,Q^{K-1}) \in [0, Q_{max}]^K \subset \mathbf{R}^K$, where $Q^m = Q(\tau_m)$, $m = 0,\ldots,K-1$.

This discretization leads to a time-discretization of the cost function J and the state system (1)–(5). To solve this state system we have used our own finite element solver: For $N \in \mathbf{N}$ given (preferentially a multiple of K), we define $\Delta t = T/N$ and take $t_n = n\Delta t$, for $n = 0,\ldots,N$. Equations (1) and (3) are to be discretized in an implicit way. However, for (2) we are to use the method of characteristics, that stems from considering the equality:

$$\frac{D(Vq)}{Dt}(x,t) = \frac{\partial q}{\partial t}(x,t) + \frac{\partial (uq)}{\partial x}(x,t), \tag{8}$$

for the total derivative $\frac{D(Vq)}{Dt}(x,t) = \frac{\partial}{\partial \tau}[V(x,t;\tau)q(X(x,t;\tau),\tau)]_{|\tau=t}$, where $X(x,t;\tau)$ is the characteristic line (providing the position at time τ of the particle that occupied the position x at time t), and $V(x,t;\tau)$ is the evolution of the element of volume.

Then, if we denote $X^n(x) = X(x,t_{n+1};t_n)$ and $V^n(x) = V(x,t_{n+1};t_n)$, the state system (1)–(5) can be approximated by the following semi-discrete system:

For $n = 0,\ldots,N-1$ find functions $A^{n+1}(x)$, $q^{n+1}(x)$, $c^{n+1}(x)$ in $(0,L)$ such that:

$$A^{n+1} = A^n(x) + \Delta t\left(Q(t_{n+1})\delta(x-p) + g_1(x,t_{n+1}) - \frac{\partial q^{n+1}}{\partial x}(x)\right), \tag{9}$$

$$\frac{q^{n+1}(x)}{\Delta t} + g\left(\frac{\partial \zeta^{n+1}}{\partial x}(x) - F(A^{n+1}(x),x) + A^{n+1}(x)b'(x)\right)$$
$$= \frac{q^n(X^n(x))V^n(x)}{\Delta t} + Q(t_{n+1})W(t_{n+1})\cos(\gamma)\delta(x-p) + g_2(x,t_{n+1}), \tag{10}$$

$$\frac{c^{n+1}(x)-c^n(x)}{\Delta t} + \frac{\partial}{\partial x}\left(\frac{q^{n+1}c^{n+1}}{A^{n+1}}(x)\right) + k(x,t_{n+1})c^{n+1}(x) = g_3(x,t_{n+1}). \tag{11}$$

The admissible set U_{ad} is approached by $U_{ad}^{\Delta\tau}$, the set of controls in U_{ad} which are piecewice-constant for the grid $\{\tau_0,\ldots,\tau_K\}$. Then, for any given control $Q^{\Delta\tau} \in U_{ad}^{\Delta\tau}$, we consider the following discrete approximation of the cost function J:

$$J^{\Delta t}(Q^{\Delta\tau}) = \frac{\varepsilon}{2}\Delta\tau\sum_{m=0}^{K-1}\frac{(Q^m)^2 + (Q^{m+1})^2}{2}$$
$$+\frac{1}{2}\Delta t\sum_{n=0}^{N-1}\int_p^r \frac{(c^n(x)-c_{max})_+^2 + (c^{n+1}(x)-c_{max})_+^2}{2}\,dx \tag{12}$$

Thus, the problem $(\mathscr{P})$, can be approached by the semi-discretized problem:

$$(\mathscr{P}^{\Delta t}) \qquad \min_{Q^{\Delta\tau}\in U_{ad}^{\Delta\tau}} J^{\Delta t}(Q^{\Delta\tau})$$

To solve problem $(\mathscr{P}^{\Delta t})$, we need to resolve the semi-discrete system (9)–(11) for $n = 0,\ldots,N-1$. Since variable c^{n+1} is uncoupled with (9)–(10), we can proceed to solve it sequentially.

First, we compute A^{n+1} and q^{n+1}: For a standard variational formulation of (10), we choose $\Lambda_h = \{x_0 = 0, x_1, \ldots, x_M = L\}$ a partition of interval $[0,L]$ in M subintervals $I_k = [x_{k-1}, x_k]$, $k = 1,\ldots,M$, such that there exist

$P, R \in \{1, \ldots, M\}$ verifying $x_P = p$, $x_R = r$. We also consider the space $\mathscr{V}_h = \{q_h \in C([0, L]) : q_{h|_{I_k}} \in P_1, \ \forall k = 1, \ldots, M\}$. Then, we take approximations $A_h^0, q_h^0 \in \mathscr{V}_h$ and, for $n = 0, \ldots, N-1$, we look for $A_h^{n+1}, q_h^{n+1} \in \mathscr{V}_h$ verifying:

$$A_h^{n+1} = A_h^n + \Delta t \left(Q(t_{n+1})\delta(x-p) + g_1(x, t_{n+1}) - \frac{\partial q_h^{n+1}}{\partial x}(x) \right), \tag{13}$$

$$\begin{aligned} &\int_0^L \frac{q_h^{n+1}(x)}{\Delta t} z_h(x)\, dx - g \int_0^L \zeta_h^{n+1}(x) \frac{\partial z}{\partial x}(x)\, dx - g \int_0^L F\left(A_h^{n+1}(x), x\right) z_h(x)\, dx \\ &\quad + g \int_0^L A_h^{n+1}(x) b'(x)\, z_h(x)\, dx = \int_0^L \frac{q_h^n(x)(X_h^n(x))V_h^n(x)}{\Delta t} z_h(x)\, dx \\ &\quad + \int_0^L \left(Q(t_{n+1})W(t_{n+1})\cos(\gamma)\delta(x-p) + g_2(x, t_{n+1})\right) z_h(x)\, dx, \quad \forall z_h \in \mathscr{V}_h, \end{aligned} \tag{14}$$

$$q_h^{n+1}(0) = q_0(t_{n+1}), \tag{15}$$

where X_h^n and V_h^n are, respectively, numerical approximations of X^n and V^n. We solve this nonlinear discretized system doing an implicit discretization of the operator F and using the trapezoidal rule for the integrals.

Second, we compute c^{n+1}: Equation (11) can be now solved by using an implicit upwind finite difference scheme. In order to do it, because of the Dirac measures characterizing the sources, we consider the following approximation $\delta_{hk}(b)$ of $\delta(x_k - b)$: for each $k = 0, \ldots, M$, we define $\delta_{hk} : [0, L] \to [0, +\infty)$ by:

$$\delta_{hk}(b) = \begin{cases} (b - x_{k-1})/(x_k - x_{k-1})^2 & \text{if } b \in [x_{k-1}, x_k], \\ (x_{k+1} - b)/(x_{k+1} - x_k)^2 & \text{if } b \in [x_k, x_{k+1}], \\ 0 & \text{otherwise.} \end{cases}$$

So, taking $\{c_0^i = c_0(t_i), \ i = 0, \ldots, N\}$ and $\{c_j^0 = c^0(x_j), \ j = 0, \ldots, M\}$ as data, for each $n = 0, \ldots, N-1$, and for each $k = 1, \ldots, M$, we compute c_k^{n+1} from:

$$\begin{aligned} &\frac{c_k^{n+1} - c_k^n}{\Delta t} + \frac{\frac{q_h^{n+1}(x_k)}{A_h^{n+1}(x_k)} c_k^{n+1} - \frac{q_h^{n+1}(x_{k-1})}{A_h^{n+1}(x_{k-1})} c_{k-1}^{n+1}}{x_k - x_{k-1}} + k(x_k, t_{n+1}) c_k^{n+1} \\ &\qquad = \sum_{j=1}^{O} n_j(t_{n+1})\delta_{hk}(e_j) + \sum_{i=1}^{V} m_i(t_{n+1})\delta_{hk}(v_i). \end{aligned}$$

Finally, for each $n = 0, \ldots, N-1$, we approach $c^{n+1}(x)$ by the unique continuous function $c_h^{n+1}(x) \in \mathscr{V}_h$ verifying $c_h^{n+1}(x_k) = c_k^{n+1}$, for all $k = 0, \ldots, M$.

Thus, the semi-discrete problem $(\mathscr{P}^{\Delta t})$ is finally approached by:

$$(\mathscr{P}_h^{\Delta t}) \qquad \min_{Q^{\Delta\tau} \in U_{ad}^{\Delta\tau}} J_h^{\Delta t}(Q^{\Delta\tau})$$

where

$$J_h^{\Delta t}(Q^{\Delta\tau}) = \frac{\varepsilon}{2}\Delta\tau \sum_{m=0}^{K-1} \frac{(Q^m)^2 + (Q^{m+1})^2}{2} + \frac{1}{2}\Delta t \sum_{n=0}^{N-1} \sum_{k=P}^{R-1} \int_{x_k}^{x_{k+1}} \frac{(c_h^n(x) - c_{max})_+^2 + (c_h^{n+1}(x) - c_{max})_+^2}{2}\, dx \tag{16}$$

In order to solve the minimization problem $(\mathscr{P}_h^{\Delta t})$ we propose the use of a derivative-free algorithm (the Nelder–Mead algorithm), but previously to do this, we need to change our discretized problem $(\mathscr{P}_h^{\Delta t})$ into an unconstrained optimization problem by introducing a penalty function involving the constraints appearing in the definition of the set of admissible controls U_{ad}, that is, $Q \geq 0$ and $Q - Q_{max} \leq 0$.

Thus, we define the penalty function $\widetilde{J}$ in the following way:

$$\widetilde{J}(Q^{\Delta\tau}) = J_h^{\Delta t}(Q^{\Delta\tau}) + \beta \sum_{m=0}^{K-1} \max\{-Q^m, Q^m - Q_{max}, 0\} \tag{17}$$

where the parameter $\beta > 0$ determines the relative contribution of the objective function and the penalty terms. Function $\widetilde{J}$ is an exact penalty function in the sense that, for sufficiently large β, the solutions of our constrained problem $(\mathscr{P}_h^{\Delta t})$ are equivalent to the minimizers of function $\widetilde{J}$ in $\mathbf{R}^K$.

For computing a minimum of this penalty function $\widetilde{J}$ we use a direct search algorithm: the Nelder–Mead simplex method [6]. This is a gradient-free method, which merely compares function values; the values of the objective function being taken from a set of sample points (simplex) are used to continue the sampling. Although the Nelder–Mead algorithm is not guaranteed to converge in the general case, it presents good convergence properties in low dimensions, which is our case. Moreover, to prevent stagnation at non-optimal points, we use a modification proposed by Kelley [5]: when stagnation is detected, we modify the simplex by an oriented restart, replacing it by a new smaller simplex.

4 Numerical Example

We present here numerical results obtained by using above method to determine the optimal inflow flux in a river which is $L = 2{,}000\,\text{m}$ in length, and where we consider $O = 3$ tributaries, $V = 2$ domestic wastewater discharges, and one clear water discharge from a reservoir (diagram and data can be seen in Fig. 1).

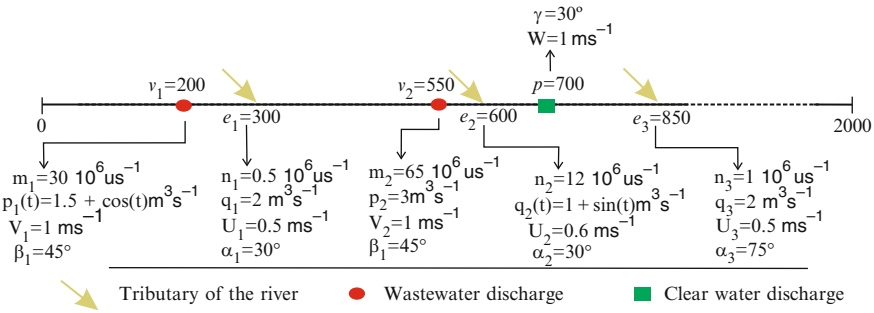


Fig. 1 Diagram of the river and data for the numerical example

We consider a parabolic river bed with a non-constant bottom in such a way that:

$$A = S(H,x) = \frac{4\sqrt{H^3}}{3}, \qquad b(x) = \begin{cases} \dfrac{500-x}{200} & \text{if } 0 \le x \le 500, \\ 0 & \text{if } 500 \le x \le 2000. \end{cases}$$

Both initial and boundary conditions were taken as constant, particularly, $A_L(t) = A^0(x) = \frac{4\sqrt{125}}{3}\,\mathrm{m}^2$, $q_0(t) = q^0(x) = 1\,\mathrm{m}^3\mathrm{s}^{-1}$ and $c_0(t) = c^0(x) = 0\,\mathrm{um}^{-1}$. Time interval for controlling pollution in the river section $[700, 1000]$ was $T = 3{,}600\,\mathrm{s}$, and the pollutant loss rate was considered constant ($k(x,t) = 10^{-4}\,\mathrm{s}^{-1}$).

Out of the several numerical experiences developed by the authors, we present here only one example corresponding to the case of $K = 4$ time subintervals. We have chosen the threshold $c_{max} = 4.5\,\mathrm{um}^{-1}$, the bound $Q_{max} = 25\,\mathrm{m}^3\mathrm{s}^{-1}$, and the weight parameters $\varepsilon = 10^{-4}$ and $\beta = 10^4$. For the time discretization we have taken $N = 6{,}000$ (that is, a time step of $\Delta t = 0.6\,\mathrm{s}$), and for the space discretization we have tried a regular partition of $[0, L]$ in $M = 2{,}000$ subintervals (consequently, the clear water inflow point $p = 700\,\mathrm{m}$ corresponds to the node $x_P = x_{700}$).

Then, applying the Nelder–Mead algorithm, we have passed, after 123 function evaluations, from an initial random cost $\widetilde{J} = 0.180$ to the minimum cost $\widetilde{J} = 0.069$, corresponding to the optimal flow rate $Q^0 = 9.98\,\mathrm{m}^3\mathrm{s}^{-1}$, $Q^1 = 6.90\,\mathrm{m}^3\mathrm{s}^{-1}$, $Q^2 = 5.44\,\mathrm{m}^3\mathrm{s}^{-1}$, $Q^3 = 3.99\,\mathrm{m}^3\mathrm{s}^{-1}$. In Fig. 2 we can observe the differences between no injection of water (uncontrolled case) and the optimal injection of water in point $p = 700\,\mathrm{m}$ (controlled case): In the first case pollutant c remains over c_{max} at the three shown times $t = 1{,}000,\ 2{,}000,\ 3{,}000\,\mathrm{s}$; we can also see how, at those same times, c turns everywhere under threshold c_{max} from injected point p, when the optimal discharge of clear water is considered. Finally, we must note that c is not necessarily lower than c_{max} when considering large values of the weight parameter $\varepsilon \gg 1$ since, in that case, we are mostly concerned about reducing the amount of injected water.

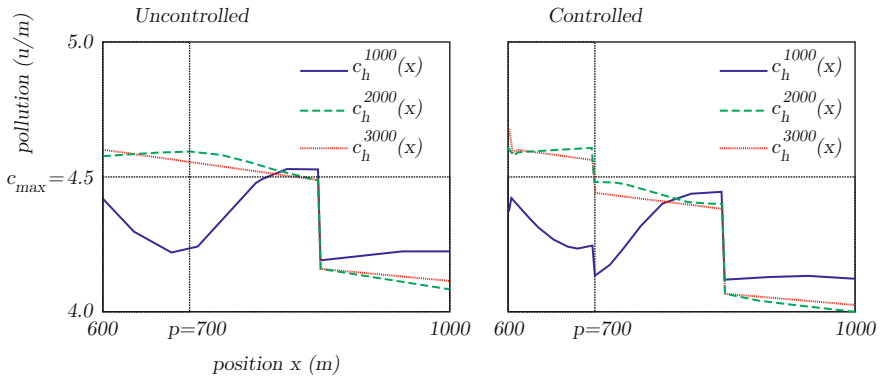


Fig. 2 Uncontrolled (*left*) and controlled (*right*) quantity of pollutant at three significant times ($t = 1{,}000,\ 2{,}000,\ 3{,}000$) around points $p = 700$ and $e_3 = 850$

Acknowledgements The authors are grateful for financial support to Projects MTM2009-07749 from MICIIN (Spain), and INCITE09-291-083-PR from Xunta de Galicia.

References

1. Alvarez-Vázquez, L.J., Martínez, A., Rodríguez, C., Vázquez-Méndez, M.E.: Numerical optimization for the location of wastewater outfalls. Comput. Optim. Appl. **22**, 399–417 (2002)
2. Alvarez-Vázquez, L.J., Martínez, A., Vázquez-Méndez, M.E., Vilar, M.A.: Optimal location of sampling points for river pollution control. Math. Comput. Simul. **71**, 149–160 (2006)
3. Alvarez-Vázquez, L.J., Martínez, A., Vázquez-Méndez, M.E., Vilar, M.A.: An application of optimal control theory to river remediation. Appl. Numer. Math. **59**, 845–858 (2009)
4. Bermúdez, A., Muñoz-Sola, R., Rodríguez, C., Vilar, M.A.: Theoretical and numerical study of an implicit discretization of a 1D inviscid model for river flows. Math. Mod. Meth. Appl. Sci. **16**, 375–395 (2006)
5. Kelley, C.T.: Detection and remediation of stagnation in the Nelder–Mead algorithm using a sufficient decrease condition. SIAM J. Optim. **10**, 43–55 (1999)
6. Nelder, J.A., Mead, R.: A simplex method for function minimization. Comput. J. **7**, 308–313 (1965)

An Anisotropic Micro-Sphere Approach Applied to the Modelling of Soft Biological Tissues

A. Menzel, T. Waffenschmidt, and V. Alastrué

Abstract A three-dimensional model for the simulation of anisotropic soft biological tissues is discussed. The underlying constitutive equations account for large strain deformations and are based on a hyper-elastic form. As various soft biological tissues are nearly incompressible, we adopt the classical volumetric-isochoric split of the strain energy density. While its isotropic part is chosen to take a standard neo-Hookean form, its anisotropic part is determined by means of the so-called micro-sphere model. In this regard, physically sound one-dimensional constitutive models – as for instance the worm-like chain model – can be used and straightforwardly be extended to the three-dimensional case. As a key aspect, the micro-sphere model is extended to further capture remodelling. Such deformation-induced anisotropy is introduced by setting up evolution equations for the integration directions used to perform numerical integrations on the unit-sphere. The particular model proposed captures orthotropic material behaviour and additionally accounts for saturation effects combined with a visco-elasticity-type time-dependent anisotropy evolution.

A. Menzel
Institute of Mechanics, Department of Mechanical Engineering, TU Dortmund, Leonhard-Euler-Str. 5, 44227 Dortmund, Germany
e-mail: andreas.menzel@udo.edu
and
Division of Solid Mechanics, Lund University, P.O. Box 118, SE-22100 Lund, Sweden
e-mail: andreas.menzel@solid.lth.se

T. Waffenschmidt (✉)
Institute of Mechanics, Department of Mechanical Engineering, TU Dortmund, Leonhard-Euler-Str. 5, 44227 Dortmund, Germany
e-mail: tobias.waffenschmidt@udo.edu

V. Alastrué
Group of Structural Mechanics and Materials Modelling, Aragón Institute of Engineering Research (I3A), University of Zaragoza, María de Luna, 3 E-50018 Zaragoza, Spain
e-mail: victorav@unizar.es

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_68,

1 Introduction

Apart from biological and chemical effects affecting the behaviour of soft biological materials such as ligaments, tendons, muscles, and skin – to name just a few examples – biological tissues in general possess a pronounced composite-type multi-scale structure together with strongly anisotropic mechanical properties. The local mechanical response of these tissues is typically determined by elastin and collagen fibre-bundels.

In this regard, adaptation of these fibre networks influenced by mechanical loading is a main biomechanical phenomenon occurring in hard as well as soft biological tissues. In general, adaptation processes can include changes in mass and internal structure, whereas this paper exclusively focuses on the latter, which we denote as remodelling – the related processes often being denoted as fibre reorientation or rather turnover.

The computational remodelling approach proposed in the following is partly motivated by the investigations reported in [4], where a fibroblast-populated collagen lattice was tested. As a result, macroscopically tension-type mechanical loads cause the initially unstructured collagen fibre network to reorient with the local dominant stretch direction and thus showing transversely isotropic characteristics. However, as many biological tissues – for example arteries – show fibre alignment with more than one single direction, we here extend the remodelling formulation proposed in [6] for transversal isotropy to orthotropic material behaviour.

The paper is organised as follows: Section 2 briefly reviews essential kinematic relations, based on which key aspects of the micro-sphere model are outlined in Sect. 3. Section 4 constitutes the main part of this contribution, wherein the remodelling formulation is introduced. In Sect. 5 a numerical example is discussed, before the paper closes with a short summary in Sect. 6.

2 Essential Kinematics

Let $\mathbf{x} = \varphi(\mathbf{X}, t) : \mathscr{B}_0 \times \mathscr{T} \to \mathscr{B}_t$ describe the motion of a body mapping position vectors $\mathbf{X} \in \mathscr{B}_0$ from the material configuration to their spatial counterpart $\mathbf{x} \in \mathscr{B}_t$. The local deformation is characterised by the common deformation gradient tensor $\mathbf{F} = \nabla_X \varphi$ with the Jacobian $J = \det(\mathbf{F}) > 0$ and the corresponding right Cauchy-Green strain tensor $\mathbf{C} = \mathbf{F}^{\mathrm{t}} \cdot \mathbf{F}$, while their isochoric counterparts are represented as $\overline{\mathbf{F}} = J^{-\frac{1}{3}}\,\mathbf{F}$ and $\overline{\mathbf{C}} = \overline{\mathbf{F}}^{\mathrm{t}} \cdot \overline{\mathbf{F}}$.

In view of the computational micro-sphere-scheme used later on, additional kinematic relations referring to the underlying unit-sphere $\mathbb{U}^2$ are introduced. In this regard, an affine stretch in the direction of a referential unit-vector $\mathbf{r} \in \mathbb{U}^2$ can be determined by the macroscopic deformation tensor $\overline{\mathbf{C}}$ via $\overline{\lambda} = \sqrt{\mathbf{r} \cdot \overline{\mathbf{C}} \cdot \mathbf{r}}$. However, it is well-known that the affinity assumption is not in agreement with experimental observations for cross-linked polymer-type materials. For this reason, according to [7], we make use of a non-affine stretch taking the form

$$\lambda = \left[\frac{1}{4\,\pi} \int_{\mathbb{U}^2} \overline{\lambda}^{p}\, A \right]^{1/p} . \tag{1}$$

Obviously the non-affine stretch can be interpreted as an averaged stretch over the unit-sphere, with p defining a non-affine stretch parameter.

Moreover, inspired by [7], a single collagen chain is additionally constrained by another micro-kinematic variable, namely the contraction ν of the cross-section of a micro-tube that contains the related chain. Therefore, analogous to relation (1), one could also introduce a non-affine area stretch

$$\nu = \left[\frac{1}{4\,\pi} \int_{\mathbb{U}^2} \bar{\nu}^{q}\, A \right]^{1/q} , \tag{2}$$

where q denotes a non-affine tube parameter. At this stage, however, we restrict ourselves to account only for the non-affine stretch-contributions λ.

3 Hyper-Elastic Micro-Sphere Model

Apart from the remodelling approach discussed later on, we make use of a hyper-elastic form of the strain energy. In this regard, we adopt the well-established volumetric-isochoric split and decompose the isochoric part into an isotropic and an anisotropic contribution, namely

$$\Psi(\mathbf{C}, \mathbf{r}_i) = \Psi^{\text{vol}}(J) + \Psi^{\text{iso}}(\overline{\mathbf{C}}) + \Psi^{\text{ani}}(\lambda(\overline{\mathbf{C}}, r_{\mathbf{i}})); \tag{3}$$

see [1, 2] in view of an affine anisotropic part. Due to the almost incompressible response of soft biological tissues, we assume a nearly-incompressible neo-Hooke model to account for the volumetric and isotropic isochoric part of the strain energy, i.e.

$$\Psi^{\text{vol}} = \frac{1}{4}\, D\, [\, J - 1\,]^2 \quad \text{and} \quad \Psi^{\text{iso}} = \mu\, [\, \mathbf{I} : \overline{\mathbf{C}} - 3\,], \tag{4}$$

with D defining a penalty parameter, μ being a material parameter and $\mathbf{I}$ representing the second-order identity tensor.

According to the highly anisotropic material properties of the type of biological tissue we are interested in, the strain energy function (3) is assumed to depend not only on the right Cauchy–Green strain tensor $\mathbf{C}$ but also on a finite number of referential direction vectors or rather integration directions $\mathbf{r}_i$ defined on the micro-sphere $\mathbb{U}^2$. In this regard, a one-dimensional constitutive equation is applied for every integration direction $\mathbf{r}_i$. To be specific, we make use of the micro-mechanically motivated worm-like chain model, which takes the representation

$$\psi^{\text{ani}}(\tilde{\lambda}) = \frac{K\,\theta\,L}{4\,A}\left[\,2\,\frac{\bar{r}^2}{L^2} + \frac{1}{1-\frac{\bar{r}}{L}} - \frac{\bar{r}}{L} - \frac{\ln(\tilde{\lambda}^4 r_0^2)}{4\,r_0\,L}\left[\,4\,\frac{r_0}{L} + \frac{1}{[1-\frac{r_0}{L}]^2} - 1\,\right] - \psi_c\,\right], \tag{5}$$

for $\tilde{\lambda} \geq 1$ while $\psi^{\text{ani}}(\tilde{\lambda}) = 0$ is assumed for compression, i.e. $\tilde{\lambda} < 1$. Herein the Boltzmann constant is denoted by $K = 1.38 \times 10^{-23}\,\text{JK}^{-1}$, θ is the absolute temperature, and A is established as persistent contour length. While r_0 characterises the length of the chain for the undeformed state, the actual representative chain length follows from $\bar{r} = \tilde{\lambda}\,r_0 \in [0, L)$.

The extension of this one-dimensional constitutive law (5) to the three-dimensional macroscopic level is performed by means of the micro-sphere formulation. Characteristic for this approach is a finite number of unit vectors $\mathbf{r}_i$ to be considered for the numerical integration over the unit sphere $\mathbb{U}^2$, which yields the total anisotropic contribution Ψ^{ani} to be computed by means of the fibre-related strain energy ψ^{ani} via

$$\Psi^{\text{ani}}(\lambda(\overline{\mathbf{C}}, \mathbf{r}_i)) = \frac{1}{4\,\pi}\int_{\mathbb{U}^2} \psi^{\text{ani}}(\lambda(\overline{\mathbf{C}}, \mathbf{r}_i))\;A. \tag{6}$$

4 A Remodelling Formulation for Orthotropic Material Behaviour

The key aspect of this contribution consists in incorporating remodelling-phenomena by setting up deformation-driven evolution equations for the integration directions $\mathbf{r}_i$, which means that these are not constant but evolve in time.

To be specific we directly relate the integration directions – now taking the interpretation as internal variables – to the numerical framework, i.e. the integration of (6), which, algorithmically, leads to a summation over a finite number of integration directions

$$\Psi^{\text{ani}}(\lambda(\overline{\mathbf{C}}, \mathbf{r}_i)) \cong \sum_{i=1}^{m} w_i\;\psi^{\text{ani}}(\lambda(\overline{\mathbf{C}}, \mathbf{r}_i)) = \psi^{\text{ani}}(\lambda(\overline{\mathbf{C}}, \mathbf{r}_i)), \tag{7}$$

with w_i denoting integration factors, which depend on the particular integration scheme.

Since various biological tissues show fibre alignment with more than one single direction, we subsequently propose a remodelling formulation reflecting macroscopically orthotropic behaviour. An analogous approach for the transversely isotropic case has recently been discussed in detail, see [6].

In view of the reorientation criterion, a crucial point consists in the identification of the deformation-dependent mean directions $\mathbf{l}_{1,2}$, which on the one hand should determine the alignment of the integration directions and on the other hand is here assumed to reflect extremal states of strain energy. In this context, one could align

the integration directions $\mathbf{r}_i$ with respect to the principal stretch directions or alternatively such that the directions, according to which the integration unit-vector are aligned with, share identical angles with the principal stretch directions. As a special case of the latter approach, we make use of two particular directions reported in [3]: the so-called limiting directions, which can be calculated via the relation

$$\mathbf{l}_{1,2} = \frac{\sqrt{\lambda_1^{\overline{C}}}\,\mathbf{n}_2^{\overline{C}} \pm \sqrt{\lambda_2^{\overline{C}}}\,\mathbf{n}_1^{\overline{C}}}{\sqrt{\lambda_1^{\overline{C}} + \lambda_2^{\overline{C}}}} \quad \text{for} \quad \lambda_1^{\overline{C}} > \lambda_2^{\overline{C}} > 1 \tag{8}$$

using the principal values $\lambda_{1,2}^{\overline{C}}$ and principal directions $\mathbf{n}_{1,2}^{\overline{C}}$ of the isochoric right Cauchy–Green strain tensor $\overline{\mathbf{C}}$. Practically speaking, these limiting directions suffer the maximum shear in the considered plane of tension.

As a result, the evolution of $\mathbf{r}_i$ is motivated by its alignment with the limiting directions $\mathbf{l}_{1,2}$ – see Fig. 1 – as reflected by

$$\dot{\mathbf{r}}_i = f\ \text{sign}\,(\mathbf{r}_i \cdot \mathbf{l})\ [\mathbf{l} - [\mathbf{r}_i \cdot \mathbf{l}]\,\mathbf{r}_i] \quad \text{so that} \quad \dot{\mathbf{r}}_i \cdot \mathbf{r}_i = 0, \tag{9}$$

where the integration direction $\mathbf{r}_i$ aligns either with $\mathbf{l}_2$ in case of $\mathbf{r}_i$ being closer to $\mathbf{l}_2$ or with $\mathbf{l}_1$ else, i.e.

$$\mathbf{l} = \begin{cases} \mathbf{l}_2 & \text{if} \quad |\mathbf{r}_i \cdot \mathbf{l}_1| \leq |\mathbf{r}_i \cdot \mathbf{l}_2| \\ \mathbf{l}_1 & \text{else} \end{cases} . \tag{10}$$

Unlike the approach used in [6] and due to the present assumption of orthotropy with two mean directions, in this case two second-order generalised structural tensors are introduced as

$$\mathbf{A}^{1,2} = \sum_{i=1}^{m} w_i\ \mathbf{r}_i \otimes \mathbf{r}_i = \sum_{j=1}^{3} A_j^{1,2}\ \mathbf{n}_j^{1,2} \otimes \mathbf{n}_j^{1,2} \qquad \forall \qquad \mathbf{r}_i \to \mathbf{l}_{1,2}, \tag{11}$$

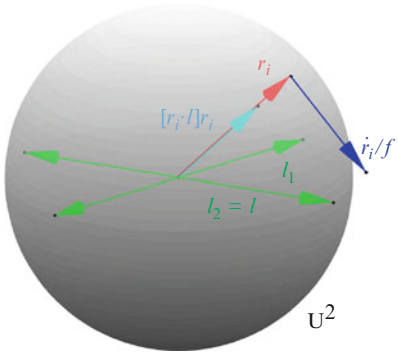


Fig. 1 Graphical illustration of the evolution equation (9): construction of the direction of the rate of $\mathbf{r}_i$ according to an alignment with respect to the closest limiting direction $\mathbf{l}_{1,2}$

where $A_1^{1,2} \geq A_2^{1,2} \geq A_3^{1,2} \geq 0$ with $\sum_j A_j^{1,2} = 1/2$. To give an example, $\mathbf{A}^1$ will be calculated for those integration directions $\mathbf{r}_i$, which – due to the deformation – align with the first mean direction $\mathbf{l}_1$. In order to later on visualise local anisotropic material properties the orientation-distribution-type function $\rho^A = \rho^{A^1} \cup \rho^{A^2}$ is introduced with $\rho^{A^{1,2}} = \mathbf{e} \cdot \mathbf{A}^{1,2} \cdot \mathbf{e}$ and $\mathbf{e} \in \mathbb{U}^2$ denoting a unit-vector.

The remaining task consists in particularising the factor f occurring in the evolution equation (9). As the adaptation of biological tissues is usually bounded by certain biological limits, the evolution equation for $\mathbf{r}_i$ should account for saturation effects. On the one hand, the evolution saturates for $\mathbf{r}_i$ aligning with one of the limiting directions $\mathbf{l}_{1,2}$. On the other hand, we restrict the maximum degree of anisotropy by assuming $\mathbf{r}_i$ to evolve as long as the difference between the largest and smallest eigenvalue of $\mathbf{A}^{1,2}$ remain smaller than a pre-defined limit value A_Δ. In addition, a relaxation parameter t^* is incorporated and we also set $\dot{\mathbf{r}}_i$ to zero in case the difference of the related fibre stretches remain smaller than a certain threshold $\overline{\lambda}_c$. In summary, the proportionality factor f introduced in (9) is assumed as

$$f = \begin{cases} \dfrac{A_\Delta - [\,A_1^{1,2} - A_3^{1,2}\,]}{t^*\, A_\Delta} & \text{if} \quad \lambda_2^{\overline{C}} > 1 \quad \text{and} \quad \lambda_1^{\overline{C}} - \lambda_2^{\overline{C}} > \overline{\lambda}_c \\ 0 & \text{else} \end{cases} . \tag{12}$$

5 Numerical Example and Results

The model is now investigated for homogeneous biaxial tension with the corresponding deformation gradient $\mathbf{F} = \lambda_1^U\, \mathbf{e}_1 \otimes \mathbf{e}_1 + \lambda_2^U\, \mathbf{e}_2 \otimes \mathbf{e}_2 + \left[\lambda_1^U \lambda_2^U\right]^{-1} \mathbf{e}_3 \otimes \mathbf{e}_3$. Material parameters are chosen by analogy with data identified for a media arterial layer, see [1], namely $\mu = 1.268\,\text{kPa}$, $B = 1.019\,\text{kPa}$, $r_0 = 1.045\,\text{mm}$, $L = 1.477\,\text{mm}$, $A_\Delta = 0.5$, $t^* = 2\,\text{s}$ and $\overline{\lambda}_c = 0$. The particular loading history considered is based on linearly increasing the representative loading parameters λ_1^U and λ_2^U within a time period of 20 time steps and then fixing its value for a time period of 380 steps; see Fig. 2a.

Special emphasis is thereby placed on the evolution of deformation-induced anisotropy, which is illustrated by means of $\mathbf{A}^{1,2}$ in terms of the odf-type function ρ^A and via the difference between its respective maximal and minimal principal values, $A_1^{1,2} - A_3^{1,2}$.

Figure 2b shows the saturation behaviour of the anisotropy evolution by means of visualising the degree of anisotropy $A_1^{1,2} - A_3^{1,2}$. Obviously it takes place in a viscous manner as the loading is fixed after 20 steps and the graph of $A_1^{1,2} - A_3^{1,2}$ continues to increase.

The anisotropy evolution is additionally visualised in a more descriptive odf-type manner by Fig. 2c. We observe for the different states of deformation, that the anisotropy evolves in time as the odf deviates from a spherical distribution.

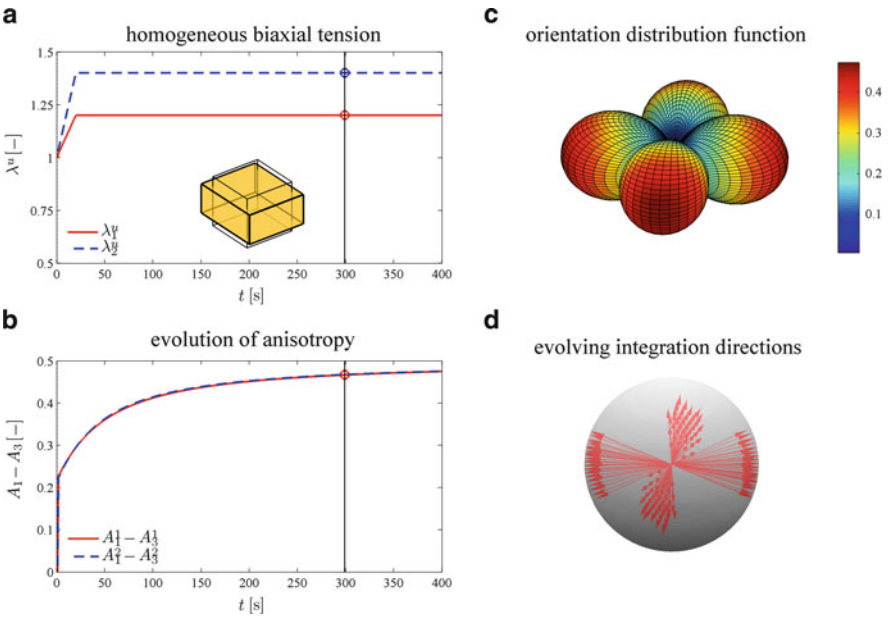


Fig. 2 Biaxial tension: (a) applied loading history; (b) evolution of $A_1^{1,2} - A_3^{1,2}$; (c) odf-type function ρ_{300}^{A} at load step 300; (d) integration directions at load step 300

The entire information can be obtained by directly displaying the integration directions in Fig. 2d. We see that these directions indeed align according to two limiting directions. Note that in this case we used 162 integration directions for the integration on the unit sphere as discussed in, e.g. [5].

6 Summary

In this work, remodelling is understood as a process that renders the internal substructure of the material to adapt to the local loading conditions. Such an alignment of fibres is often also denoted as reorientation or, from the biological point of view, as turnover. The model developed directly combines this remodelling with the computational micro-sphere approach. To be specific, the respective directions introduced to perform the numerical integration over the unit-sphere are reoriented. As a result, the formulation accounts for deformation-induced anisotropy evolution.

In order to capture orthotropic material behaviour, the evolution equation describing the reorientation, was assumed to align the integration directions with respect to two particular mean line elements – the so-called limiting directions. Saturation effects are on the one hand naturally included by a stopping remodelling process as soon as a direction is aligned with the particular limiting direction. On the other

hand, an additional saturation value has been introduced to be able to further limit the maximal degree of anisotropy of the tissue.

The numerical example investigated showed the basic algorithmic applicability of the modelling framework and captured the fundamental reorientation and remodelling effects observed for soft biological tissues. Moreover – even though not shown here – the formulation can be applied to the simulation of general boundary value problems as based on, for instance, iterative finite element approaches.

References

1. V. Alastrué, M.A. Martínez, M. Doblaré, and A. Menzel. Anisotropic micro-sphere-based finite elasticity applied to blood vessel modelling. J. Mech. Phys. Solids, 57:178–203 (doi:10.1016/j.jmps.2008.09.005), 2009
2. V. Alastrué, M.A. Martínez, A. Menzel and M. Doblaré. On the use of non-linear transformations for the evaluation of anisotropic rotationally symmetric directional integrals. Application to the stress analysis in fibred soft tissues. Int. J. Numer. Meth. Eng., 79(4):474–504 (doi: 10.1002/nme.257), 2009
3. P. Boulanger, M. Hayes. On finite shear. Arch. Ration. Mech. Anal. 151, 125–185, 2000
4. M. Eastwood, V.C. Mudera, D.A. McGrouther, and R.A. Brown. Effect of precise mechanical loading on fibroblast populated collagen lattices: morphological changes. Cell Motil. Cytoskeleton, 40:1321, 1998
5. I. Kurzhöfer. Mehrskalen-Modellierung polykristalliner Ferroelektrika basierend auf diskreten Orientierungsverteilungsfunktionen, Universität Duisburg-Essen, Institut für Mechanik, Bericht Nr. 4, 2007
6. A. Menzel, T. Waffenschmidt. A micro-sphere-based remodelling formulation for anisotropic biological tissues. Phil. Trans. R. Soc. A, 367(1902):3499–3523 (doi: 10.1098/rsta.2009.0103), 2009
7. C. Miehe, S. Göktepe, and F. Lulei. A micro-macro approach to rubber-like materials – Part I: the non-affine micro-sphere model of rubber elasticity. J. Mech. Phys. Solids, 52:2617–2660, 2004

Anisotropic Adaptation via a Zienkiewicz–Zhu Error Estimator for 2D Elliptic Problems

S. Micheletti and S. Perotto

Abstract We propose a Zienkiewicz–Zhu a posteriori error estimator in 2D, which shares the computational advantages typical of the original estimator. The novelty is the inclusion of the geometrical features of the computational mesh, useful for an anisotropic mesh adaptation. The adapted triangulations are shown numerically to be quasi-optimal with respect to the error-vs-number of elements behavior.

1 Motivations

Among the various a posteriori error estimation techniques available in the literature, one of the most popular in practice is the one proposed by Zienkiewicz and Zhu [11, 12]. The idea behind this estimator is quite simple: for example, consider the finite element approximation u_h to the solution u of an advection–diffusion-reaction (ADR) equation. Since the gradient ∇u_h is less accurate than the solution, we recover an improved gradient, say $\nabla^* u_h$, by suitably fitting ∇u_h over some patches of elements. The discrepancy $\|\nabla^* u_h - \nabla u_h\|_{L^2(\Omega)}$ then identifies an estimator for the $H^1(\Omega)$-seminorm of the discretization error $u - u_h$.

The popularity of this methodology can be attributed to various factors: the method is independent of the problem, of the governing equations and of most details of the finite element formulation (except for the finite element space), it is cheap to compute and easy to implement, and works very well in practice.

On the other hand, ADR problems often exhibit strong directional features (e.g., internal or boundary layers). In these cases the effectiveness of the finite element approximation benefits from a suitable anisotropic computational mesh, fitting size, shape and orientation of its triangles to the directional features of the solution at hand [1, 5, 9].

S. Micheletti and S. Perotto (✉)
MOX, Dipartimento di Matematica "F. Brioschi", Politecnico di Milano, Via Bonardi 9, I-20133 Milano, Italy
e-mail: stefano.micheletti, simona.perotto@polimi.it

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_69,

In this paper we propose some gradient recovery techniques suited to define an anisotropic counterpart of the Zienkiewicz–Zhu estimator. The novelty is the inclusion of the geometrical information of the mesh triangles, maintaining the above good properties of the standard Zienkiewicz–Zhu estimator. Despite the somewhat heuristic nature of the proposed estimator, the overall anisotropic adaptation procedure turns out to be effective in practice. The adapted meshes, built through a metric-based optimisation algorithm, are shown numerically to be quasi-optimal with respect to the error-vs-number of elements behavior.

2 Recovery Procedures

According to a Zienkiewicz–Zhu approach, we distinguish between two steps: first we furnish a procedure for obtaining an approximate recovered gradient; second, we employ this recovered gradient for a posteriori error purposes.

To fix ideas, we consider the standard ADR problem completed with homogeneous Dirichlet boundary conditions, i.e., find $u \in V$, such that

$$\int_\Omega \mu \nabla u \cdot \nabla v \, d\mathbf{x} + \int_\Omega \mathbf{b} \cdot \nabla u \, v \, d\mathbf{x} + \int_\Omega \gamma \, u \, v \, d\mathbf{x} = \int_\Omega f \, v \, d\mathbf{x} \quad \forall v \in V, \tag{1}$$

with Ω a polygonal domain in $\mathbb{R}^2$, $\mu > 0$, $\mathbf{b} \in [W^{1,\infty}(\Omega)]^2$, $\gamma \in L^\infty(\Omega)$, and where $V = H_0^1(\Omega)$, standard notation being adopted for the Sobolev spaces and their norms. Proper assumptions are enforced to guarantee the well-posedness of (1).

Let $\mathcal{T}_h = \{K\}$ be a conforming partition of Ω consisting of triangles and u_h be the Galerkin affine finite element approximation to (1), possibly involving stabilization.

We now provide a family of recovery procedures to improve the discrete gradient ∇u_h, using information only related to u_h. Several approaches are available in the literature for this purpose (see, e.g., [8, 10, 11]). We propose here a recovered gradient, denoted by $P^r_{\Delta_K}(\nabla u_h)$, which has degree r over the patch $\Delta_K = \{T \in \mathcal{T}_h : T \cap K \neq \emptyset\}$. We seek $P^r_{\Delta_K}(\nabla u_h) \in [\mathbb{P}_r]^2$ such that

$$\int_{\Delta_K} (\nabla u_h - P^r_{\Delta_K}(\nabla u_h)) \cdot \mathbf{w} \, d\mathbf{x} = 0 \quad \forall \mathbf{w} \in [\mathbb{P}_r]^2, \tag{2}$$

with $\mathbb{P}_r = \text{span}\{x_1^i \, x_2^j \mid i + j \leq r\}$. The recovered gradient $P^r_{\Delta_K}(\nabla u_h)$ is strictly associated with K, and not to the elements comprising Δ_K (i.e., for any $T \in \Delta_K$, with $T \neq K$, $P^r_{\Delta_T}(\nabla u_h)$ is, in general, different from $P^r_{\Delta_K}(\nabla u_h)$). In the particular case $r = 0$, we can write out the formula for the recovered gradient, given by

$$P^0_{\Delta_K}(\nabla u_h) = \frac{1}{|\Delta_K|} \sum_{T \in \Delta_K} |T| \, \nabla u_h|_T,$$

namely, we compute the area-weighted average over the patch Δ_K of the gradients of the discrete solution.

3 The Anisotropic Estimator

To devise the estimator proposed in this work, we embed the recovery procedures above in a convenient anisotropic setting. This leads to a Zienkiewicz–Zhu-like estimator, automatically including the anisotropic information of the mesh elements. The same potentiality is not so evident in the case of the standard Zienkiewicz–Zhu estimator [12].

3.1 Anisotropic Source

We employ the anisotropic setting in [4]. The size, shape and orientation of each element K of $\mathscr{T}_h$ are characterized by the affine map $T_K : \widehat{K} \to K$, where $\widehat{K}$ is the equilateral reference triangle centred at the origin, with coordinates $(-\sqrt{3}/2, -1/2)$, $(\sqrt{3}/2, -1/2)$, $(0, 1)$ and edge length $\sqrt{3}$. It holds $\mathbf{x} = T_K(\widehat{\mathbf{x}}) = M_K\,\widehat{\mathbf{x}} + \mathbf{t}_K$, with $M_K \in \mathbb{R}^{2\times 2}$ the Jacobian and $\mathbf{t}_K \in \mathbb{R}^2$ the shift vector. Matrix M_K is factorized as $M_K = B_K\,Z_K$ via the polar decomposition, 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.

Through T_K the unit circle circumscribing $\widehat{K}$ is changed into an ellipse circumscribing K: the unit vectors $\{\mathbf{r}_{i,K}\}$ define the corresponding principal directions, whereas the quantities $\{\lambda_{i,K}\}$ measure the length of the ellipse semi-axes. Without loss of generality, we assume $\lambda_{1,K} \geq \lambda_{2,K} > 0$ so that the stretching factor, $s_K = \lambda_{1,K}/\lambda_{2,K}$, satisfies $s_K \geq 1$, for any $K \in \mathscr{T}_h$, equality holding when K is equilateral.

The estimator proposed in Sect. 3.2 is inspired by an anisotropic interpolation error estimate derived in this setting [4]. In particular, let I_h^1 be the Clément interpolant of degree 1 for functions $v \in H^1(\Omega)$.

Proposition 1. *Let $v \in H^1(\Omega)$. Then, if $\#\Delta_K \leq \mathscr{D}$ and $\mathrm{diam}(\widehat{\Delta}_K) \leq \delta$, for any $K \in \mathscr{T}_h$, there exists a constant $C = C(\mathscr{D}, \delta)$, such that*

$$\|v - I_h^1(v)\|_{L^2(K)} \leq C\Big(\sum_{i=1}^{2} \lambda_{i,K}^2\,(\mathbf{r}_{i,K}^T G_{\Delta_K}(\nabla v)\,\mathbf{r}_{i,K})\Big)^{1/2}, \tag{3}$$

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

$$[G_{\Delta_K}(\mathbf{w})]_{i,j} = \sum_{T\in\Delta_K} \int_T w_i\,w_j\,d\mathbf{x}, \quad \text{with } i,j = 1,2, \tag{4}$$

for any vector-valued function $\mathbf{w} = (w_1, w_2)^T \in [L^2(\Omega)]^2$, $\#\Delta_K$ the cardinality of the patch, $\mathrm{diam}(\widehat{\Delta}_K)$ the diameter of $\widehat{\Delta}_K = T_K^{-1}(\Delta_K)$, the pullback of Δ_K via the map T_K.

Remark 1. The hypotheses of Proposition 1 constrain the variation of $\{\mathbf{r}_{i,K}\}$ and $\{\lambda_{i,K}\}$ over Δ_K but do not limit the anisotropy of K.

3.2 The Estimator

Driven by Proposition 1 we devise the anisotropic a posteriori error estimator. Let $\mathbf{E}^r_{\Delta_K} = P^r_{\Delta_K}(\nabla u_h) - \nabla u_h|_{\Delta_K}$ be the approximation for the error on the gradient, over Δ_K. We define the anisotropic Zienkiewicz–Zhu local estimator for the H^1-seminorm of the discretization error as

$$\left[\eta^r_{K,\mathrm{A}}\right]^2 = \frac{1}{\lambda_{1,K}\lambda_{2,K}} \sum_{i=1}^{2} \lambda^2_{i,K} \left(\mathbf{r}^T_{i,K}\, G_{\Delta_K}(\mathbf{E}^r_{\Delta_K})\mathbf{r}_{i,K}\right), \tag{5}$$

where the matrix $G_{\Delta_K}(\cdot)$ is defined as in (4). Then the corresponding global error estimator is given by

$$\left[\eta^r_{\mathrm{A}}\right]^2 = \sum_{K\in\mathcal{T}_h} \left[\eta^r_{K,\mathrm{A}}\right]^2. \tag{6}$$

The estimator (5) and (6) is essentially heuristic. The terms summed on the right-hand side of (5) are suggested by (3) with $v = u - u_h$, after substituting the partial derivatives of u with the corresponding components of $P^r_{\Delta_K}(\nabla u_h)$. However some rationale can be provided. The scaling factor $\lambda_{1,K}\lambda_{2,K}$ guarantees a *consistency* with respect to the isotropic case, i.e., when $\lambda_{1,K} = \lambda_{2,K}$, (5) turns into an isotropic Zienkiewicz–Zhu-like estimator based on the patchwise recovered gradient (2), that is

$$\left[\eta^r_{K,\mathrm{I}}\right]^2 = \int_{\Delta_K} |\mathbf{E}^r_{\Delta_K}|^2\, d\mathbf{x} \quad \text{and} \quad \left[\eta^r_{\mathrm{I}}\right]^2 = \sum_{K\in\mathcal{T}_h} \left[\eta^r_{K,\mathrm{I}}\right]^2.$$

Moreover a sort of *equivalence* between $\eta^r_{K,\mathrm{A}}$ and $|u-u_h|_{H^1(\Delta_K)}$ can be proved. In more detail, given a function $v \in H^1(\Omega)$, let $\widehat{v} = v \circ T_K$ be the associated pullback. Virtually, we would like to choose $v = u - u_h$. Then, we have

$$\int_{\widehat{\Delta}_K} |\widehat{\nabla}\,\widehat{v}|^2 d\mathbf{x} = \frac{1}{\lambda_{1,K}\lambda_{2,K}} \sum_{i=1}^{2} \lambda^2_{i,K} \left(\mathbf{r}^T_{i,K} G_{\Delta_K}(\nabla v)\, \mathbf{r}_{i,K}\right),$$

$$s_K^{-1}\, |v|^2_{H^1(\Delta_K)} \le \frac{1}{\lambda_{1,K}\lambda_{2,K}} \sum_{i=1}^{2} \lambda^2_{i,K} \left(\mathbf{r}^T_{i,K} G_{\Delta_K}(\nabla v)\, \mathbf{r}_{i,K}\right) \le s_K\, |v|^2_{H^1(\Delta_K)},$$

where the middle term mimics estimator (5), on replacing ∇v with $\mathbf{E}^r_{\Delta_K}$.

The patch test

We aim to check the consistency of the recovery procedure by computing the local effectivity index $\mathrm{E.I.}^r_{K,\mathrm{A}} = \eta^r_{K,\mathrm{A}}/|u-u_h|_{H^1(K)}$, for $r = 0, 1$. To avoid a bias effect

due to the grid, we consider the case when u is isotropic and the regular patch Δ_K consists of 13 equilateral triangles, each of area $3\sqrt{3}/4$, with pivot element $\widehat{K}$. In particular let $u = a\,x_1^2 + 2b\,x_1x_2 + c\,x_2^2$, with $a, b, c \in \mathbb{R}$, picked such that the Hessian, $H = [a\ b; b\ c]$, has eigenvalues with the same modulus. This happens only when (i) $a = c$ and $b = 0$ or when (ii) $a = -c$ and b is arbitrary. As typical in a patch test, let u_h coincide with the Lagrange affine interpolant of u. It turns out that $|u - u_h|_{H^1(K)} = |a|\,\sqrt{|K|}$, $\|P^0_{\Delta_K}(\nabla u_h) - \nabla u_h\|_{L^2(\Delta_K)} = |a|\,\sqrt{132\,|K|}$, i.e., $\text{E.I.}^0_{K,\text{A}} = \sqrt{132} \simeq 11.5$ in the case (i); the case (ii) leads to $|u - u_h|_{H^1(K)} = \sqrt{2\,|K|\,(a^2+b^2)}$, $\|P^0_{\Delta_K}(\nabla u_h) - \nabla u_h\|_{L^2(\Delta_K)} = \sqrt{1884\,|K|\,(a^2+b^2)/13}$, i.e., $\text{E.I.}^0_{K,\text{A}} = \sqrt{1884/26} \simeq 8.51$. Analogously, for the case $r = 1$, we obtain $\text{E.I.}^1_{K,\text{A}} \simeq 3.44$ in the case (i) and $\text{E.I.}^1_{K,\text{A}} \simeq 3.52$ in the case (ii).

It can be checked that the same values can be obtained after applying either roto-translations or homotheties to Δ_K.

Although this isotropic context may seem favorable, we expect a similar behavior also in the anisotropic case, provided that the mesh is adapted to the solution.

Estimator (5) and (6) is problem-free, i.e., it can be applied to more general problems, such as elasticity or Navier–Stokes equations. In such a case one could replace, e.g., the gradient with the stress (rate) tensor [11]. Alternately, the adaptation can be driven by the gradient of a scalar variable representative of the problem, like the pressure or the speed for the Navier–Stokes equations.

The estimator corresponding to $r = 0$ is extended to the 3D case in [2]. Here an adaptation driven by a scalar quantity (speed for the Navier–Stokes equations and density for a multimaterial application) is also assessed.

4 The Adaptive Procedure

We employ a metric-based adaptive procedure driven by estimator η^r_{A}. In particular, for a fixed accuracy on the numerical solution, we look for the mesh with the least number of elements. The tensor field $\widetilde{M} : \Omega \to \mathbb{R}^2$, is the actual unknown. According to a predictive procedure, at each iteration, j, of the adaptive process, we deal with: (i) the actual mesh $\mathcal{T}_h^{(j)}$, where problem (1) is approximated; (ii) the new metric $\widetilde{M}^{(j+1)}$ piecewise constant on $\mathcal{T}_h^{(j)}$, predicted elementwise through a suitable local optimization procedure; (iii) the new mesh $\mathcal{T}_h^{(j+1)}$ guaranteeing that all the edges are unit length with respect to $\widetilde{M}^{(j+1)}$ [7].

We focus on step (ii), which is at the heart of the whole adaptive procedure. We minimize $\left[\eta^r_{K,\text{A}}\right]^2$ in (5) with respect to stretching and orientation, and then, via an equidistribution criterion, we compute the actual values of $\lambda_{1,K}$ and $\lambda_{2,K}$. For this purpose we first rewrite the estimator as

$$\begin{aligned}\left[\eta^r_{K,\text{A}}\right]^2 &= s_K\,(\mathbf{r}^T_{1,K}\,G_{\Delta_K}(\mathbf{E}^r_{\Delta_K})\,\mathbf{r}_{1,K}) + s_K^{-1}\,(\mathbf{r}^T_{2,K}\,G_{\Delta_K}(\mathbf{E}^r_{\Delta_K})\,\mathbf{r}_{2,K})\\ &= \lambda_{1,K}\,\lambda_{2,K}\,|\widehat{\Delta}_K|\big[s_K\,(\mathbf{r}^T_{1,K}\,\widehat{G}_{\Delta_K}(\mathbf{E}^r_{\Delta_K})\,\mathbf{r}_{1,K}) + s_K^{-1}\,(\mathbf{r}^T_{2,K}\,\widehat{G}_{\Delta_K}(\mathbf{E}^r_{\Delta_K})\,\mathbf{r}_{2,K})\big],\end{aligned} \tag{7}$$

where $\widehat{G}_{\Delta_K}(\cdot)$ is the scaled matrix $G_{\Delta_K}(\cdot)/|\Delta_K|$, and $|\Delta_K| = \lambda_{1,K}\lambda_{2,K}\,|\widehat{\Delta}_K|$. The idea is that we single out the area-dependent information (the multiplicative term) from the quantity in brackets, depending on orientation and stretching. Then we minimize this last term with respect to s_K and $\{\mathbf{r}_{i,K}\}$, as stated by the following result.

Proposition 2. *Let*

$$J(s_K, \{\mathbf{r}_{i,K}\}_{i=1,2}) = s_K\,(\mathbf{r}_{1,K}^T \widehat{G}_{\Delta_K}(\mathbf{E}^r_{\Delta_K})\mathbf{r}_{1,K}) + s_K^{-1}\,(\mathbf{r}_{2,K}^T \widehat{G}_{\Delta_K}(\mathbf{E}^r_{\Delta_K})\mathbf{r}_{2,K}), \tag{8}$$

and let $\{g_i, \mathbf{g}_i\}_{i=1,2}$ be the eigen-pairs associated with $\widehat{G}_{\Delta_K}(\mathbf{E}^r_{\Delta_K})$, where it is understood $g_1 \geq g_2 > 0$ and $\{\mathbf{g}_i\}_{i=1,2}$ are orthonormal. Then $J(\cdot)$ is minimized when

$$s_K = \sqrt{g_1/g_2}, \quad \mathbf{r}_{1,K} = \mathbf{g}_2, \quad \mathbf{r}_{2,K} = \mathbf{g}_1. \tag{9}$$

Proof. The result follows from Proposition 14 in [3].

Notice that the optimal values in (9) equalize the two terms in (8), i.e., $s_K\,g_2 = s_K^{-1}\,g_1 = \sqrt{g_1\,g_2}$. This implies that the minimum of $J(\cdot)$ does not depend on s_K. To construct $\widetilde{M}^{(j+1)}$, we just have to compute $\{\lambda_{i,K}\}_{i=1,2}$. For this purpose we employ the equidistribution criterion, according to which $\left[\eta^r_{K,\mathrm{A}}\right]^2 = \tau^2/\#\mathscr{T}_h^{(j)}$, where τ is the fixed accuracy and $\#\mathscr{T}_h^{(j)}$ is the cardinality of the background mesh. Thanks to Proposition 2, we obtain $\lambda_{1,K}\,\lambda_{2,K} = \tau^2/(\,2\,\#\mathscr{T}_h^{(j)}|\widehat{\Delta}_K|\sqrt{g_1 g_2}\,)$. Since $s_K = \lambda_{1,K}/\lambda_{2,K}$, we have

$$\lambda_{1,K} = g_2^{-1/2}\left(\frac{\tau^2}{2\,\#\mathscr{T}_h^{(j)}|\widehat{\Delta}_K|}\right)^{1/2}, \quad \lambda_{2,K} = g_1^{-1/2}\left(\frac{\tau^2}{2\,\#\mathscr{T}_h^{(j)}|\widehat{\Delta}_K|}\right)^{1/2}. \tag{10}$$

The predicted metric $\widetilde{M}^{(j+1)}$ is formed, elementwise, by $\widetilde{M}_K^{(j+1)} = \widetilde{M}^{(j+1)}|_K = R_K^T \Lambda_K^{-2} R_K$, with $R_K^T = [\mathbf{r}_{1,K}, \mathbf{r}_{2,K}]$ and $\Lambda_K = \mathrm{diag}(\lambda_{1,K}, \lambda_{2,K})$, where $\{\mathbf{r}_{i,K}\}_{i=1,2}$ and $\{\lambda_{i,K}\}_{i=1,2}$ are provided by (9) and (10), respectively.

Now, for task (iii), we employ the function `adaptmesh` in [6]. Since it takes as input a nodewise representation of $\widetilde{M}^{(j+1)}$, we have to average the elementwise information. The nodewise metric is thus $\widetilde{M}_N^{(j+1)} = (3\,|\Delta_N|)^{-1}\,\sum_{K\in\Delta_N}|K|\,\widetilde{M}_K^{(j+1)}$, where Δ_N is the patch of elements sharing node N and $|\Delta_N|$ is the corresponding area. The scaling factor $1/3$ shrinks the reference triangle to a unit edge one.

Remark 2. The hypothesis on the eigenvalues in Proposition 2 can be relaxed by assuming $g_1 \geq g_2 \geq 0$, i.e., that $\widehat{G}_{\Delta_K}(\mathbf{E}^r_{\Delta_K})$ is actually positive semidefinite. This degenerate case is tackled by choosing $g_i = \max(g_i, g_{\min})$, for $i = 1, 2$, where $g_{\min} = \tau^2/\big(h_\Omega^2\,2\,\#\mathscr{T}_h^{(j)}|\widehat{\Delta}_K|\big)$, with h_Ω the diameter of the domain. Thus, if g_i is degenerate, $\lambda_{i,K} = h_\Omega$.

Test Case 1: Pure Diffusion

We solve (1) with $\mu = 1$, $\mathbf{b} = \mathbf{0}$, $\gamma = 0$ on $\Omega = (-1, 1)^2$, with f chosen such that $u(x_1, x_2) = \tanh(10x_2^2 - 20x_1^3)(x_2^2 - 1)$ and Dirichlet compatible boundary conditions. We apply the above adaptive procedure with the choices $\tau = 2, 1, 0.5$ and $r = 0, 1$. Figure 1 gathers the final adapted grids for $\tau = 1$, obtained after eight iterations. The meshes match the anisotropic features of u, as highlighted by the detail on the right. In Tables 1 and 2 a more quantitative analysis is provided. The effectivity index $\text{E.I.}_{\text{A}}^{r} = \eta_{\text{A}}^{r}/|u - u_h|_{H^1(\Omega)}$ is essentially independent of τ. In the case $r = 1$ the meshes are coarser, $\text{E.I.}_{\text{A}}^{1}$ being closer to 1. The error-vs-number of elements behavior is quasi-optimal in both cases, i.e., of the order of about -0.5.

Test Case 2: Advection–Diffusion

We now consider an instance of (1) more complex than test case 1, choosing $\mu = 10^{-3}$, $\mathbf{b} = (x_2, -x_1)^{\text{T}}$, $\gamma = 0$, $f = 1$ on $\Omega = (0, 1)^2$. The exact solution, not explicitly available, exhibits two boundary layers and a circular internal layer. We discretize (1) by the SUPG method. The adaptive procedure is run, picking $\tau = 2, 1$ and $r = 1$. All the layers are sharply detected by the anisotropic estimator (see Fig. 2). The results in Table 3 confirm the reliability of both the estimator and the adaptive procedure. Notice also the large values of the stretching factor, the maximum being reached in correspondence with the two boundary layers.

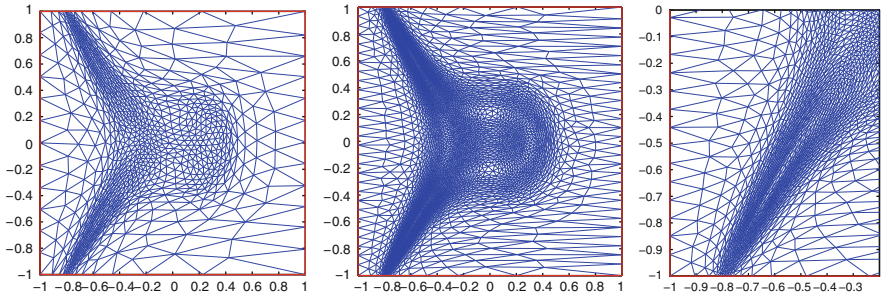


Fig. 1 Final adapted grids for test case 1: $\tau = 1$, $r = 1$ (*left*), $r = 0$ (*middle* and *right*)

Table 1 Test case 1: $r = 0$

τ	$\#\mathcal{T}_h$	max s_K	$\lvert u - u_h\rvert_{H^1(\Omega)}$	η_{A}^{0}	$\text{E.I.}_{\text{A}}^{0}$	η_{I}^{0}	$\text{E.I.}_{\text{I}}^{0}$
2	1,861	25.15	$0.3038 \cdot 10^{+0}$	$0.2108 \cdot 10^{+1}$	6.939	$0.3019 \cdot 10^{+1}$	9.938
1	6,220	31.13	$0.1552 \cdot 10^{+0}$	$0.1081 \cdot 10^{+1}$	6.965	$0.1572 \cdot 10^{+1}$	10.12
0.5	22,388	48.13	$0.8024 \cdot 10^{-1}$	$0.5522 \cdot 10^{+0}$	6.882	$0.8149 \cdot 10^{+0}$	10.16

Table 2 Test case 1: $r = 1$

τ	$\#\mathcal{T}_h$	max s_K	$\lvert u - u_h\rvert_{H^1(\Omega)}$	η_{A}^{1}	$\text{E.I.}_{\text{A}}^{1}$	η_{I}^{1}	$\text{E.I.}_{\text{I}}^{1}$
2	533	19.64	$0.6753 \cdot 10^{+0}$	$0.1747 \cdot 10^{+1}$	2.586	$0.2393 \cdot 10^{+1}$	3.543
1	1,541	17.34	$0.3503 \cdot 10^{+0}$	$0.8802 \cdot 10^{+0}$	2.512	$0.1209 \cdot 10^{+1}$	3.450
0.5	4,699	27.71	$0.1893 \cdot 10^{+0}$	$0.4408 \cdot 10^{+0}$	2.328	$0.6180 \cdot 10^{+0}$	3.264

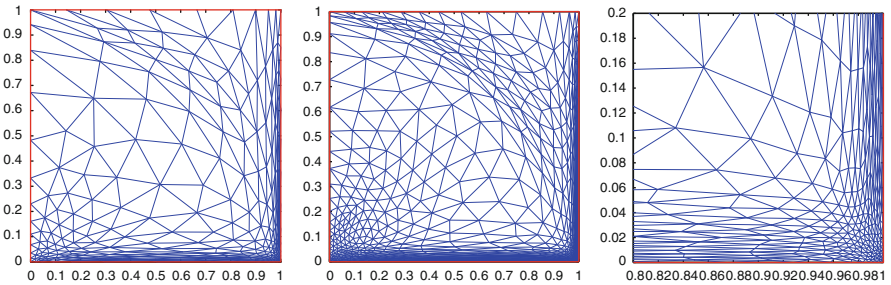


Fig. 2 Final adapted grids for test case 2: $r = 1$, $\tau = 2$ (*left*), $\tau = 1$ (*middle* and *right*)

Table 3 Test case $2 : r = 1$

τ	$\#\mathcal{T}_h$	$\max s_K$	η_{A}^1	η_{I}^1
2	482	57.33	$0.1727 \cdot 10^{+1}$	$0.3884 \cdot 10^{+1}$
1	1,273	109.98	$0.8925 \cdot 10^{+0}$	$0.2062 \cdot 10^{+1}$

5 Conclusions

Despite its heuristic nature, the proposed anisotropic Zienkiewicz–Zhu a posteriori estimator provides satisfactory results. Indeed it detects the anisotropic features of the problem at hand, exhibiting a quasi-optimal error-vs-number of elements behavior as well. This occurs even in the case $r = 0$, which identifies the roughest gradient recovery in the proposed class of estimators.

References

1. Castro-Diaz, M.J., Hecht, F., Mohammadi, B., Pironneau, O.: Anisotropic unstructured mesh adaptation for flow simulations. Int. J. Numer. Meth. Fluids **25**(4), 475–491 (1997)
2. Farrell, P., Micheletti, S., Perotto, S.: An anisotropic Zienkiewicz–Zhu a posteriori error estimator for 3D applications. MOX Report 25/2009, Politecnico di Milano http://mox.polimi.it
3. Formaggia, L., Micheletti, S., 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)
4. Formaggia, L., Perotto, S.: New anisotropic a priori error estimates. Numer. Math. **89**(4), 641–667 (2001)
5. Gruau, C., Coupez, T.: 3D tetrahedral, unstructured and anisotropic mesh generation with adaptation to natural and multidomain metric. Comput. Meth. Appl. Mech. Eng. **194**(48–49), 4951–4976 (2005)
6. Hecht, F.: Freefem++, Version 3.5. http://www.freefem.org/ff++/index.htm
7. Micheletti, S., Perotto, S.: Reliability and efficiency of an anisotropic Zienkiewicz–Zhu error estimator. Comp. Meth. Appl. Mech. Eng. **195**(9–12), 799–835 (2006)
8. Naga, A., Zhang, Z.: A posteriori error estimates based on the polynomial preserving recovery. SIAM J. Numer. Anal. **42**, 1780–1800 (2004)
9. Peraire, J., Vahdati, M., Morgan, K., Zienkiewicz, O.: Adaptive remeshing for compressible flow computations. J. Comput. Phys. **72**(2), 449–466 (1987)

10. Rodríguez, R.: Some remarks on Zienkiewicz–Zhu estimator. Numer. Meth. Part. Differ. Equat. **10**, 625–635 (1994)
11. Zienkiewicz, O., Zhu, J.: The superconvergent patch recovery and a posteriori error estimates. I: The recovery technique. Int. J. Numer. Meth. Eng. **33**, 1331–1364 (1992)
12. Zienkiewicz, O., Zhu, J.: The superconvergent patch recovery and a posteriori error estimates. II: Error estimates and adaptivity. Int. J. Numer. Meth. Eng. **33**, 1365–1382 (1992)

On a Sediment Transport Model in Shallow Water Equations with Gravity Effects

T. Morales de Luna, M.J. Castro Díaz, and C. Parés Madroñal

Abstract Sediment transport by a fluid over a sediment layer can be modeled by a coupled system with a hydrodynamical component, described by a shallow water system, and a morphodynamical component, given by a solid transport flux. Meyer-Peter and Müller developed one of the most known formulae for solid transport discharge, but it has the inconvenient of not including pressure forces. This makes numerical simulations not accurate in zones where gravity effects are relevant, e.g., advancing front of the sand layer. Fowler et al. proposed a generalization that takes into account gravity effects as well as the length of the sediment layer which agrees better to the physics of the problem. We propose to solve this system by using a path-conservative scheme for the hydrodynamical part and a duality method based on Bermudez-Moreno algorithm for the morphodynamical component.

1 Introduction

The study of sediment transport processes includes movement of rocks in a mountain as material diffusion in water, among other processes. Transport is caused by gravity effects and by friction effects with the air or the fluid containing the sediment. Sediment transport is usually divided into three types: bedload, saltation and suspension. Bedload transport is defined as the type of transport where sediment grains roll or slide along the bed. Saltation transport is defined as the type of transport where single grains jump over the bed a length proportional to their diameter, losing for instants the contact with the soil. Sediment is suspended when the flux is intense enough such as the sediment grains reach height over the bed. Here we

T. Morales de Luna (✉)
Dpto. de Matemáticas, Campus de Rabanales, Universidad de Córdoba, 14071 Córdoba (Spain)
e-mail: Tomas.Morales@uco.es

M.J. Castro Díaz and C. Parés Madroñal
Dpto. Análisis Matemático, Campus de Teatinos, Universidad de Málaga, 29071 Málaga
e-mail: castro@anamat.cie.uma.es, pares@anamat.cie.uma.es

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_70,

study the case of bedload sediment transport. Bedload sediment transport process due to the movement of a fluid in contact with the sediment layer may be modeled by a coupled system constituted by a hydrodynamical component and a morphodynamical component. The hydrodynamical component is modeled by the well-known Shallow Water equations. The morphodynamical component is modeled by including a sediment transport equation, which depends on a solid transport flux given by some empirical law. Among the usual formulae for solid transport flux found in literature we shall cite the ones proposed by Grass [3], Meyer-Peter and Müller [4], Van Rijn [8], Nielsen [5] and many others.

2 The Classical Meyer-Peter and Müller Formula

The model proposed by Meyer-Peter and Müller model is one of the most used and well known formula for bedload transport of a sediment layer due to the movement of a fluid. The system can be described as follows.

$$\begin{cases} \partial_t h + \partial_x(hu) = 0, \\ \partial_t(hu) + \partial_x(hu^2 + g\frac{h^2}{2}) + gh\partial_x(z - H) = -f|u|u, \\ \partial_t z + \xi \partial_x q_b = 0, \end{cases} \tag{1}$$

where $z(t, x)$ is the thickness of the sediment layer that can be transported by the fluid, $h(t, x)$ is the thickness of the fluid layer and $u(t, x)$ is the velocity (see Fig. 1). The sediment layer is assumed to be located over a fixed bedrock at depth $H(x)$ from a given reference level.

The solid transport flux q_b is given by the classical Meyer-Peter and Müller formula

$$q_b(h, hu) = \alpha \left(\frac{|\widehat{\tau_e}|}{\beta} - \mu_{crit} \right)_+^{3/2} \operatorname{sgn}(\widehat{\tau_e}), \tag{2}$$

$$\widehat{\tau_e}(h, hu) = \gamma \frac{|u|u}{h^{1/3}}, \tag{3}$$

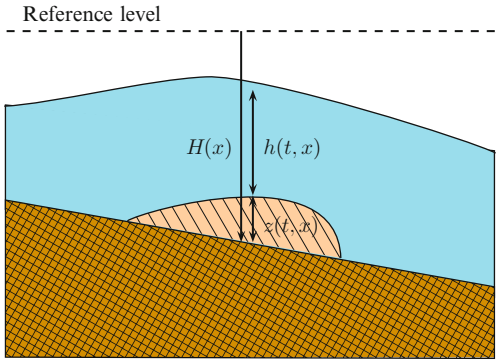


Fig. 1 Sketch of a sediment layer transported by the action of a fluid

where $\alpha, \beta, \gamma, \mu_{crit}$ are constants that depend on the sediment type considered. In particular, they may be described as

$$\alpha = 8\frac{\sqrt{\Delta\rho g D_s^3}}{\rho_w^{1/2}}, \tag{4}$$

$$\beta = \Delta\rho g D_s, \tag{5}$$

$$\gamma = \rho_w g M^2, \tag{6}$$

with D_s the mean grain size of the particles, ρ_w the density of the fluid, $\Delta\rho$ the difference between the density of the sediment and the density of the fluid and M is the Manning coefficient. μ_{crit} is the critical shear stress. Finally $\xi = \frac{1}{1-p}$, where p is the porosity of the sediment layer.

3 A Modified Meyer-Peter and Müller Model

Eventhough the classical Meyer-Peter and Müller formula has been extensively used for sediment bedload transport, it presents two main disadvantages which are also common to other of the formulae cited before.

First, gravity effects are neglected. This makes that particle do not fall due to gravity and motion of particle only depends on the velocity of the fluid. As a consequence, one can observe vertical profiles that are not found in physical situations, for example at the front of a dune.

Second, mass conservation may be lost. For example, imagine a situation like the one described in Fig. 2, where the sediment layer is only present in the interior of the domain. Assume that no sediment comes in or out of the domain through the boundaries during time interval $[0, T]$. By integrating the third equation in (1) in $[0, T] \times [a, b]$, we obtain

$$\int_a^b z|_{t=T}\,dx - \int_a^b z|_{t=0}\,dx = -\xi \int_0^T (q_b|_{x=b} - q_b|_{x=a})\,dt \tag{7}$$

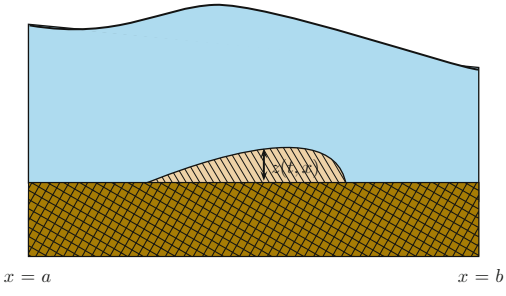


Fig. 2 Sediment layer isolated in the interior of the domain

Given that the solid flux q_b does not take into account the sediment layer thickness and only depends on the variables h and u, the right hand side of (7) can be non-zero and the mass is not preserved.

These two problems lead us consider other alternatives for the solid transport flux.

Fowler et al. proposed in [2] a modified Meyer-Peter and Müller formula where the expression (2)-(3) is replaced by

$$q_b(h, hu, z) = \alpha \frac{z}{z^0} \left(\frac{|\tau_e|}{\beta} - \mu_{crit} \right)_+^{3/2}, \tag{8}$$

$$\tau_e = \gamma \frac{|u|u}{h^{1/3}} - \partial_x(z - H), \tag{9}$$

where z^0 is a constant that represents the mean thickness of the sediment layer.

This new formula grants that the solid flux q_b is zero whenever the sediment layer vanishes so that sediment mass is preserved. Moreover, thanks to the term $\partial_x(z-H)$ in (9), gravity effects are considered.

4 Numerical Scheme

System (1), (8), (9) is no longer an hyperbolic system due to the term $\partial_x(z - H)$. We focus here in the solution of the third equation so that we may assume that h, u are given. Indeed, it is the case in many situations where we can suppose that we are near to a steady state so that they are calculated from the general relations of equilibria for shallow water system. In a more general framework, these variables may be updated by using a two step algorithm as follows.

We shall rewrite the system under the form

$$\partial_t W + \partial_x F(W, \sigma) = B(W)\partial_x W + S(W)\frac{dH}{dx}, \tag{10}$$

with

$$W = (h, hu, z)^t, \quad \widetilde{W} = (W, \sigma, H), \quad \sigma = \partial_x z, \tag{11}$$

$$F(W, \sigma) = (hu, hu^2 + g/2h^2, q_b(W, \sigma))^t \tag{12}$$

$$S(W) = (0, gh, 0)^t \tag{13}$$

$$B(W) = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -gh \\ 0 & 0 & 0 \end{pmatrix}. \tag{14}$$

Assume the approximations W_i^n over some cells I_i at time $t = t^n$. First, we compute the approximations h_i^{n+1}, u_i^{n+1} by solving

$$\begin{cases} \partial_t W + \partial_x F(W, \sigma^*) = B(W)\partial_x W + S(W)\partial_x H, \\ \widetilde{W}(t = t^n, x) = \widetilde{W_i^n} \text{ for } x \in I_i. \end{cases} \tag{15}$$

for some approximation of σ^*.

This can be done by using the general framework of path conservative schemes proposed in [7] and [6].

Second, we compute the variables z_i^n, σ_i^n by solving

$$\partial_t z - \partial_x G(\partial_x z) = 0, \tag{16}$$

with

$$G(v) = \frac{\alpha}{z^0} z \left(\left| \frac{\widehat{\tau_e}}{\beta} - v \right| - \mu_{crit} \right)_+^{3/2} \operatorname{sgn}(\beta v - \widehat{\tau_e}). \tag{17}$$

This second step can be solved by a duality algorithm proposed in [1]. When G is a maximal monotone operator, which is indeed our case, the solution of (16) can be computed iteratively by solving

$$\begin{cases} \frac{z^{l+1} - z^n}{\Delta t} - \omega \partial_{xx} z^{l+1} = \partial_x \theta^l, \\ \theta^{l+1} = G_\omega^\lambda(\partial_x z + \lambda \theta^l), \end{cases} \quad l = 0, 1, \ldots \tag{18}$$

where G_λ^ω is the Yoshida's regularization of G which is defined as

$$G_\lambda^\omega = \frac{1}{\lambda} \left[Id - J_\lambda^\omega \right], \tag{19}$$

$$J_\lambda^\omega = \left[Id + \lambda G^\omega \right]^{-1}, \tag{20}$$

$$G^\omega = G - \omega Id, \tag{21}$$

where λ and ω are constants such that $\lambda\omega = 1/2$. We refer to [1] for further details.

5 Numerical Simulations

We show here two simple numerical simulations where we can see the advantages of this new model. First, consider a sediment layer of height 0.25 and width 2 in the interior of the domain at depth 0.8. The conditions of lake at rest are assumed so that the initial surface is at level 0 and velocity is set to 0. If we use the classical Meyer-Peter and Müller model, this would be a steady state and it would not evolve as the gravity effects are not considered. But using the new model, we see in Fig. 3 that sediment falls due to its own weight until a steady slope is reached.

Now, consider a similar case where the height of the sediment layer is again 0.25 and the width is equal to 2 located at depth 0.95 from the surface and we impose a velocity $u = 1$. We see in Fig. 4 that a dune is formed as expected and now we do not

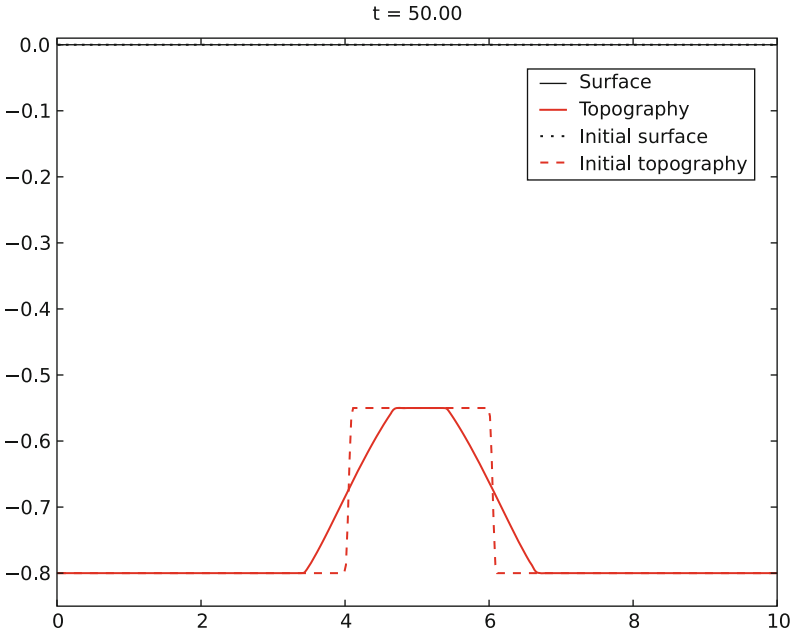


Fig. 3 Sediment layer evolution in lake-at-rest conditions

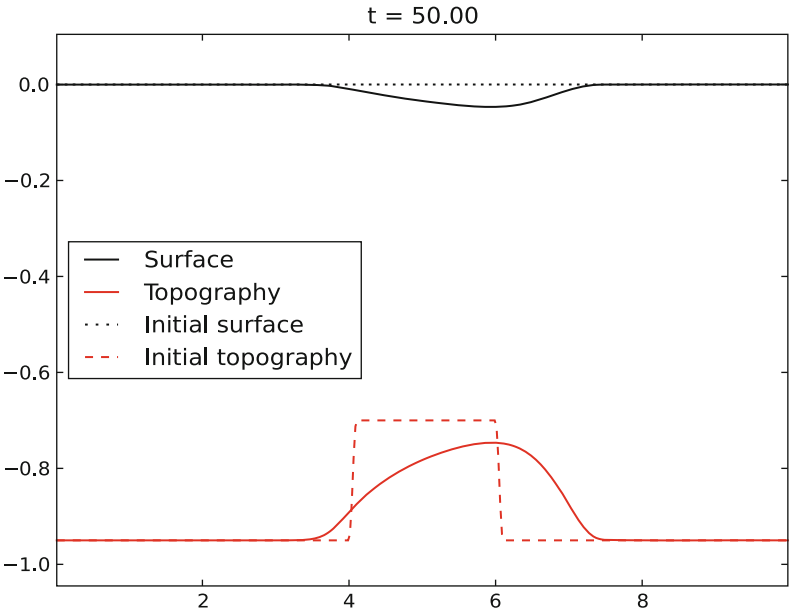


Fig. 4 Formation of a dune

have a vertical profile at the front as it would be the case for classical Meyer-Peter and Müller model.

6 Conclusions

The modified Meyer-Peter and Müller model proposed by Fowler et al. has the advantage of including gravitational effects and takes into account the thickness of the sediment layer. This makes the model to be more physically relevant and can reproduce some phenomena observed in reality that cannot be reproduced with the classical model as it has been shown in Sect. 5. The price to pay is that it is a more complex model which is no longer an hyperbolic system so that computational cost is more expensive.

References

1. A. Bermúdez and C. Moreno. Duality methods for solving variational inequalities. *Comput. Math. Appl.*, 7:43–58, 1981
2. A. C. Fowler, N. Kopteva, and C. Oakley. The formation of river channels. *SIAM Journal on Applied Mathematics*, 67(4):1016–1040, 2007
3. A. Grass. Sediment transport by waves and currents. *SERC London Cent. Mar. Technol*, Report No. FL29, 1981
4. E. Meyer-Peter and R. Müller. Formulas for bed-load transport. In *2nd meeting IAHSR, Stockholm, Sweden*, pages 1–26, 1948
5. P. Nielsen. *Coastal Bottom Boundary Layers and Sediment Transport*. World Scientific Pub Co Inc, Aug. 1992
6. C. Parés. Numerical methods for nonconservative hyperbolic systems: a theoretical framework. *SIAM J. Numer. Anal.*, 44(1):300–321 (electronic), 2006
7. C. Parés and M. Castro. On the well-balance property of Roe's method for nonconservative hyperbolic systems. Applications to shallow-water systems. *M2AN Math. Model. Numer. Anal.*, 38(5):821–852, 2004
8. L. Van Rijn. Sediment transport: bed load transport. *Journal of Hydraulic Engineering - ASCE*, 110(10):1431–1456, 1984

Adaptive SQP Method for Shape Optimization

P. Morin, R.H. Nochetto, M.S. Pauletti, and M. Verani

Abstract We examine shape optimization problems in the context of inexact sequential quadratic programming. Inexactness is a consequence of using adaptive finite element methods (AFEM) to approximate the state equation, update the boundary, and compute the geometric functional. We present a novel algorithm that uses a dynamic tolerance and equidistributes the errors due to shape optimization and discretization, thereby leading to coarse resolution in the early stages and fine resolution upon convergence. We discuss the ability of the algorithm to detect whether or not geometric singularities such as corners are genuine to the problem or simply due to lack of resolution – a new paradigm in adaptivity.

P. Morin
Instituto de Matemática Aplicada del Litoral, Universidad Nacional del Litoral, CONICET, Santa Fe, Argentina
e-mail: pmorin@santafe-conicet.gov.ar

R.H. Nochetto
Department of Mathematics and Institute for Physical Science and Technology, University of Maryland, College Park, USA
e-mail: rhn@math.umd.edu

M.S. Pauletti
Department of Mathematics, University of Maryland, College Park,
and
Department of Mathematics, Texas A&M, USA
e-mail: seba@math.umd.edu

M. Verani (✉)
MOX – Dipartimento di Matematica "F. Brioschi", Politecnico di Milano, Milano, Italy
e-mail: marco.verani@polimi.it

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_71, © Springer-Verlag Berlin Heidelberg 2010

1 Shape Optimization as Adaptive Sequential Quadratic Programming

Shape optimization problems governed by partial differential equations (PDE) can be formulated as constrained minimization problems with respect to the shape of a domain Ω in $\mathbb{R}^d$. If $u = u(\Omega)$ is the solution of a PDE in Ω, the state equation is

$$\mathscr{A}u(\Omega) = f, \tag{1}$$

and $J(\Omega) = J(\Omega, u(\Omega))$ is a cost functional, then we consider the minimization problem

$$\Omega^* \in \mathscr{U}_{ad} : \qquad J(\Omega^*) = \inf_{\Omega \in \mathscr{U}_{ad}} J(\Omega), \tag{2}$$

within the set $\mathscr{U}_{ad}$ of admissible domains in $\mathbb{R}^d$. This is a constrained minimization problem for J.

In this paper we formulate an Adaptive Sequential Quadratic Programming algorithm (or ASQP), that adaptively builds a sequence of domains $\{\Omega_k\}_{k\geq 0}$ converging to a local minimizer of the shape optimization problem (1) and (2). To motivate and briefly describe the ideas underlying ASQP, we need the concept of shape derivative $dJ(\Omega; \mathrm{w})$ of $J(\Omega)$ in the direction of a normal velocity w

$$dJ(\Omega; \mathrm{w}) = \int_\Gamma g(\Omega)\mathrm{w}, \tag{3}$$

see [14] for its precise definition. We observe that $g(\Omega)$, the *Riesz representation* of the shape derivative $dJ(\Omega)$, depends on $u(\Omega)$. We present ASQP in two steps: we first introduce an infinite dimensional Sequential Quadratic Programming (Exact SQP) algorithm, and next we introduce and motivate its adaptive finite dimensional version, responsible for the inexact nature of ASQP.

Exact SQP Algorithm. We let Ω_k be the current iterate and Ω_{k+1} be the new one. We let $\Gamma_k := \partial\Omega_k$ and let $\mathbb{V}(\Gamma_k)$ be a Hilbert space defined on Γ_k, with scalar product $b_{\Gamma_k}(\cdot,\cdot) : \mathbb{V}(\Gamma_k) \times \mathbb{V}(\Gamma_k) \to \mathbb{R}$ and norm $\|\cdot\|_{\mathbb{V}(\Gamma_k)}$. This gives rise to the elliptic selfadjoint operator $\mathscr{B}_k : \mathbb{V}(\Gamma_k) \to \mathbb{V}(\Gamma_k)^*$ defined by $\langle \mathscr{B}_k \mathrm{v}, \mathrm{w}\rangle_{\Gamma_k} = b_{\Gamma_k}(\mathrm{v}, \mathrm{w})$. We then consider the following *quadratic model* $Q_k : \mathbb{V}(\Gamma_k) \to \mathbb{R}$ of J around Ω_k

$$Q_k(\mathrm{w}) := J(\Omega_k) + dJ(\Omega_k; \mathrm{w}) + \frac{1}{2}\langle \mathscr{B}_k \mathrm{w}, \mathrm{w}\rangle. \tag{4}$$

It is easy to check that the unique minimizer v_k of $Q_k(\mathrm{w})$ satisfies

$$\mathrm{v}_k \in \mathbb{V}(\Gamma_k) : \qquad b_{\Gamma_k}(\mathrm{v}_k, \mathrm{w}) = -\langle g_k, \mathrm{w}\rangle_{\Gamma_k} \qquad \forall \mathrm{w} \in \mathbb{V}(\Gamma_k), \tag{5}$$

with $g_k := g(\Omega_k)$; i.e. $\mathrm{v}_k = -\mathscr{B}_k^{-1} g_k$. Moreover, v_k is an admissible descent direction; i.e. $dJ(\Omega_k; \mathrm{v}_k) < 0$ because $b_{\Gamma_k}(\cdot,\cdot)$ is a scalar product. Once v_k has been found, we need to determine a stepsize that is not too small and guarantees

sufficient decrease of the functional J. To accomplish this goal we identify a range of admissible stepsizes by adapting the classical *Armijo–Wolfe conditions* in $\mathbb{R}^n$: given $0 < \alpha < \beta < 1$, we seek a stepsize $\mu \in \mathbb{R}^+$ satisfying

$$J(\Omega_k + \mu \mathbf{v}_k) \leq J(\Omega_k) + \alpha\mu \, dJ(\Omega_k; \mathrm{v}_k), \qquad dJ(\Omega_k + \mu \mathbf{v}_k; \mathrm{v}_k) \geq \beta \, dJ(\Omega_k; \mathrm{v}_k), \tag{6}$$

where $\partial(\Omega_k + \mu \mathbf{v}_k) := \{\mathbf{y} \in \mathbb{R}^d : \mathbf{y} = \mathbf{x} + \mu \mathbf{v}_k(\mathbf{x}), \ \mathbf{x} \in \partial\Omega_k\}$ is the updated domain boundary and $\mathbf{v}_k = \mathrm{v}_k \boldsymbol{\nu}_k$ is a normal vector field. We are now ready to introduce the Exact Sequential Quadratic Programming algorithm for solving the constrained optimization problem (1) and (2): given the initial domain Ω_0, set $k = 0$ and iterate

Compute $u_k = u(\Omega_k)$ by solving (1)
Compute the Riesz representation $g_k = g(\Omega_k)$ of (3)
Compute the search direction v_k by solving (5)
Determine an admissible stepsize μ_k satisfying (6)
Update: $\Omega_{k+1} = \Omega_k + \mu_k \mathbf{v}_k$; $k \leftarrow k + 1$

This algorithm is not feasible as it stands, because it requires the exact computation of the following quantities at each iteration: the solution u_k to the state equation (1); the solution v_k to the linear subproblem (5); the values of the functional J and of its derivative dJ in the line search routine. Replacing all of the above non-computable operations by finite approximations yields a practical algorithm.

Adaptive SQP Algorithm (`ASQP`)**.** This method adjusts the accuracies of the various approximations relative to the energy decrease for each iteration. It is worth noticing that the adaptive procedure driving our algorithm has to deal with two distinct sources of error:

- *PDE Error*: This hinges on the approximation of (1) and the values of the functional J and its derivative (3);
- *Geometric Error*: This relates to the approximation of (5) which yields the new domain.

Since it is wasteful to impose a PDE error finer than the expected geometric error, we have a natural mechanism to balance the computational effort. The `ASQP` algorithm is an iteration of the form:

$$\ldots \to \mathscr{E}_k \to \texttt{APPROXJ} \to \texttt{SOLVE} \to \texttt{RIESZ} \to \texttt{DIRECTION} \to \texttt{LINESEARCH} \to \texttt{UPDATE} \to \mathscr{E}_{k+1} \to \ldots$$

where $\mathscr{E}_k = \mathscr{E}_k(\Omega_k, \mathbb{S}_k, \mathbb{V}_k)$ is the total error incurred in at step k, $\mathbb{S}_k = \mathbb{S}_k(\Omega_k)$ is the finite element space defined on Ω_k and $\mathbb{V}_k = \mathbb{V}_k(\Gamma_k)$ is the finite element space defined on the boundary Γ_k. To describe briefly each module along with the philosophy behind `ASQP`, we let G_k be an approximation to the shape derivative $g_k = g(\Omega_k)$ given by `RIESZ`, and $V_k \in \mathbb{V}_k(\Gamma_k)$ be an approximation to the exact solution $\mathrm{v}_k \in \mathbb{V}(\Gamma_k)$ of (5) given by `DIRECTION`. The discrepancy between v_k and

V_k leads to the geometric error. Upon using a first order Taylor expansion around Ω_k, together with (5) for the exact velocity v_k, we obtain

$$\begin{aligned}\big|J(\Omega_k + \mu_k \mathbf{V}_k) - J(\Omega_k + \mu_k \mathbf{v}_k)\big| &\simeq \mu_k \big|dJ(\Omega_k; V_k - \mathrm{v}_k)| \\ &= \mu_k \big|b_{\Gamma_k}(\mathrm{v}_k, V_k - \mathrm{v}_k)\big| \leq \mu_k \|\mathrm{v}_k\|_{\Gamma_k} \|\mathrm{v}_k - V_k\|_{\Gamma_k}.\end{aligned}$$

Motivated by this expression, we now describe the modules APPROXJ and DIRECTION , in which adaptivity is carried out. These modules are driven by different adaptive strategies and corresponding different tolerances, say a PDE tolerance γ and a geometric tolerance θ. Their relative values allow for different distributions of the computational effort in dealing with the PDE and the geometry. The routine DIRECTION enriches/coarsens the space $\mathbb{V}_k$ to control the quality of the descent direction:

$$\|V_k - \mathrm{v}_k\|_{\Gamma_k} \leq \theta \|V_k\|_{\Gamma_k}, \tag{7}$$

where $\theta \leq 1/2$ guarantees that the angle between V_k and v_k is $\leq \pi/6$; in particular $\|\mathrm{v}_k\|_{\Gamma_k} \leq (1+\theta)\|V_k\|_{\Gamma_k}$. This implies a geometric error proportional to $\mu_k \|V_k\|^2_{\Gamma_k}$, namely

$$\big|J(\Omega_k + \mu_k \mathbf{V}_k) - J(\Omega_k + \mu_k \mathbf{v}_k)\big| \leq \delta \mu_k \|V_k\|^2_{\Gamma_k}, \tag{8}$$

with $\delta := \theta(1+\theta) \leq \frac{3}{2}\theta$. On the other hand, the module APPROXJ enriches/coarsens the space $\mathbb{S}_k$ to control the error in the approximate functional value $J_k(\Omega_k + \mu_k \mathbf{V}_k)$ to the prescribed tolerance $\gamma \mu_k \|V_k\|^2_{\Gamma_k}$,

$$\big|J(\Omega_k + \mu_k \mathbf{V}_k) - J_k(\Omega_k + \mu_k \mathbf{V}_k))\big| \leq \gamma \mu_k \|V_k\|^2_{\Gamma_k}, \tag{9}$$

where $\gamma = \frac{1}{2} - \delta \geq \delta$ prevents excessive numerical resolution relative to the geometric one. This is achieved within the module APPROXJ via the *Dual Weighted Residual* method (DWR) [2], taylored to the approximation of the functional value J. The remaining modules perform the following tasks. The module SOLVE finds approximate solutions $U_k \in \mathbb{S}_k$ of (1) and $Z_k \in \mathbb{S}_k$ of an adjoint equation (necessary for the computation of $g(\Omega_k)$), while RIESZ builds on $\mathbb{S}_k$ an approximation G_k to the shape derivative g_k. Finally, the module LINESEARCH enforces an inexact version of (6).

Energy Decrease. The triangle inequality, in conjunction with conditions (8) and (9), yields

$$\big|J_k(\Omega_k + \mu_k \mathbf{V}_k) - J(\Omega_k + \mu_k \mathbf{v}_k)\big| \leq \frac{1}{2}\mu_k \|V_k\|^2_{\Gamma_k}, \tag{10}$$

which is a bound on the local error incurred in at step k. On the other hand, the exact energy decrease reads

$$\begin{aligned} J(\Omega_k) - J(\Omega_k + \mu_k \mathrm{v}_k) &\approx -\mu_k dJ(\Omega_k; \mathrm{v}_k) = \mu_k b_{\Gamma_k}(\mathrm{v}_k, \mathrm{v}_k) \\ &= \mu_k \|\mathrm{v}_k\|^2_{\Gamma_k} \geq (1-\theta)^2 \mu_k \|\mathbf{V}_k\|^2_{\Gamma_k}, \end{aligned} \tag{11}$$

and leads to the further constraint $(1-\theta)^2 > \frac{1}{2}$ to guarantee the energy decrease $J_k(\Omega_k + \mu_k \mathbf{V}_k) < J(\Omega_k)$.

If ASQP converges to a stationary point, i.e., $\mu_k \|V_k\|^2_{\Gamma_k} \to 0$ as $k \to \infty$, then the routines DIRECTION and APPROXJ approximate the descent direction V_k and functional $J(\Omega_k)$ increasingly better as $k \to \infty$, as dictated by (7) and (9). In other words, this imposes a dynamic error tolerance and progressive improvement in approximating U_k, Z_k and G_k as $k \to \infty$. This argument is a consistency check of ASQP.

We observe that the test (9) is not very demanding for DWR. So we expect coarse meshes at the beginning, and a combination of refinement and coarsening later as DWR detects geometric singularities, such as corners, and sorts out whether they are genuine to the problem or just due to lack of numerical resolution. This aspect of our approach is a novel paradigm in adaptivity and is documented in Sect. 3.

Prior Work. The idea of coupling FEM, a posteriori error estimators and optimal design error estimators to efficiently solve shape optimization problems is not new. The pioneering work [3] presents an iterative scheme, where the Zienkiewicz–Zhu error indicator and the L^2 norm of the shape gradient are both used at each iteration to improve the PDE error and the Geometric error, respectively. However, the algorithm in [3] does not resort to any dynamically changing tolerance, that would allow, as it happens for ASQP, to produce coarse meshes at the beginning of the iteration and a combination of Geometric and PDE refinement/coarsening later on. Moreover, [3] does not distinguish between fake and genuine geometric singularities that may arise on the domain boundary during the iteration process, and does not allow the former to disappear. More recently, the use of adaptive modules for the numerical approximation of PDEs has been employed by several authors [1, 11, 12] to improve the accuracy of the solution of shape optimization problems. However, in these papers the critical issue of linking the adaptive PDE approximation with an adaptive procedure for the numerical treatment of the domain geometry is absent. We address this linkage below.

2 Drag Minimization for Stokes Flow

Let $\Omega \subset \mathbb{R}^d$, $d \geq 2$ be a bounded domain of $\mathbb{R}^d$. Let $\mathbf{u} := \mathbf{u}(\Omega)$ and $p := p(\Omega)$ solve the Stokes problem:

$$-\mathrm{div}\mathbf{T}(\mathbf{u}, p) = 0, \qquad \mathrm{div}\,\mathbf{u} = 0, \qquad \text{in } \Omega, \tag{12}$$

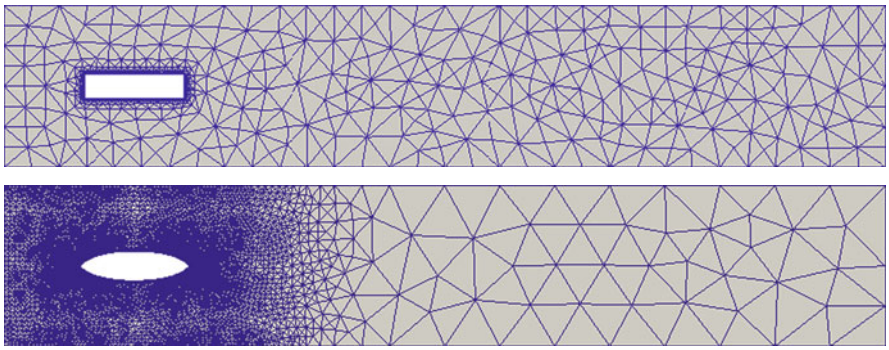


Fig. 1 Initial (*top*) and final (*bottom*) configuration: Γ_s is the deformable part of Ω, Γ_{in} the left-hand part, Γ_{out} the right-hand part and Γ_w the union of the upper and lower part. The algorithm obtains the optimal "rugby ball" shape [10]. The mesh refinement takes place mostly around Γ_s, whereas in the rest of Ω the mesh is rather coarse: this is related to DWR mesh refinement (and coarsening) and the particular expression (13) of the cost functional $J(\Omega)$

with Dirichlet boundary condition $\mathbf{u} = \mathbf{v}_\infty$ on Γ_{in}, $\mathbf{u} = \mathbf{0}$ on $\Gamma_s \cup \Gamma_w$, and traction-free boundary condition $\mathbf{T}(\mathbf{u}, p) \cdot \mathbf{n} = 0$ on Γ_{out} (see Fig. 1). Hereafter, $\mathbf{T}(\mathbf{u}, p) := 2\nu\boldsymbol{\epsilon}(\mathbf{u}) - p\mathbf{I}$ is the stress tensor with $\boldsymbol{\epsilon}(\mathbf{u}) = \frac{\nabla\mathbf{u}+\nabla\mathbf{u}^T}{2}$, and $\mathbf{v}_\infty = V_\infty\hat{\mathbf{v}}_\infty$, with $\hat{\mathbf{v}}_\infty$ being the unit vector directed as the incoming flow and V_∞ a scalar function.

The drag exerted by the fluid on the obstacle surrounded by Γ_s is given by the functional

$$J(\Omega) = J(\Omega, \mathbf{u}, p) := -\int_{\Gamma_s} \hat{\mathbf{v}}_\infty \cdot \mathbf{T}(\mathbf{u}, p) \cdot \mathbf{n}\, d\Gamma. \tag{13}$$

We consider the following shape optimization problem $\min_{\Omega\in\mathscr{U}_{ad}} J(\Omega)$ on the set $\mathscr{U}_{ad}$ of admissible configurations with given volume, obtained by perturbing only the boundary Γ_s of the obstacle [10].

It is possible to prove [9] that, for all sufficiently smooth vector fields $\mathbf{v}$ which are non-zero in a neighborhood of Γ_s, the shape derivative of $J(\Omega)$ in the direction $\mathbf{v}$ is given by

$$dJ(\Omega; \mathbf{v}) = -2\nu \int_{\Gamma_s} \boldsymbol{\epsilon}(\mathbf{u}) : \boldsymbol{\epsilon}(\mathbf{z})\, \mathrm{v}\, d\Gamma, \tag{14}$$

with $\mathrm{v} = \mathbf{v} \cdot \mathbf{n}$ the normal velocity and $\mathbf{z}$ the solution to the adjoint problem

$$-\operatorname{div}\mathbf{T}(\mathbf{z}, q) = 0, \qquad \operatorname{div}\mathbf{z} = 0, \qquad \text{in } \Omega, \tag{15}$$

subject to Dirichlet boundary conditions $\mathbf{z} = -\hat{\mathbf{v}}_\infty$ on Γ_s, $\mathbf{z} = \mathbf{0}$ on $\Gamma_w \cup \Gamma_{in}$, and traction-free condition $\mathbf{T}(\mathbf{z}, q) \cdot \mathbf{n} = 0$ on Γ_{out}.

3 Numerical Experiment: Optimal Shape for Drag Minimization

In this section we briefly describe key aspects of the implementation of ASQP for the successful realization of simulations. A full description of the algorithm can be found in [9]. The implementation of ASQP was done using the toolbox ALBERTA [13], and the graphics were produced with ParaView [7].

Adaptivity. Adaptivity is carried out inside the modules APPROXJ and DIRECTION. In the module APPROXJ, adaptivity is performed using the goal-oriented Dual Weighted Residual estimator (DWR) driven by approximation of the boundary functional $J(\Omega)$ [2]. Briefly, the goal-oriented DWR estimator determines where to refine/coarsen the mesh in Ω in order to improve the functional approximation, without imposing a small error in the global energy norm over the whole domain (see Figs. 1 and 3).

The scalar velocity v_k obeys (5) with $\mathbb{V}(\Gamma_k) := H^1(\Gamma_k)$ and the bilinear form $b_{\Gamma_k}(\mathrm{v}, \mathbf{w}) := \int_{\Gamma_k} \alpha_b \nabla_\Gamma v \cdot \nabla_\Gamma w + \beta_b v w$, where ∇_Γ denotes the surface gradient, and $\alpha_b = 10^{-3}$, $\beta_b = 1$. The module DIRECTION enforces the bound (7) on $\|V_k - \mathrm{v}_k\|_{\Gamma_k}$ using the a posteriori error estimators for the *Laplace–Beltrami* (LB) operator Δ_Γ developed in [8]. They are of residual type and estimate the energy error when solving $\Delta_\Gamma u = f$ on a known surface Γ. They consist of the usual PDE estimator and a new *geometric estimator* that accounts for the approximation

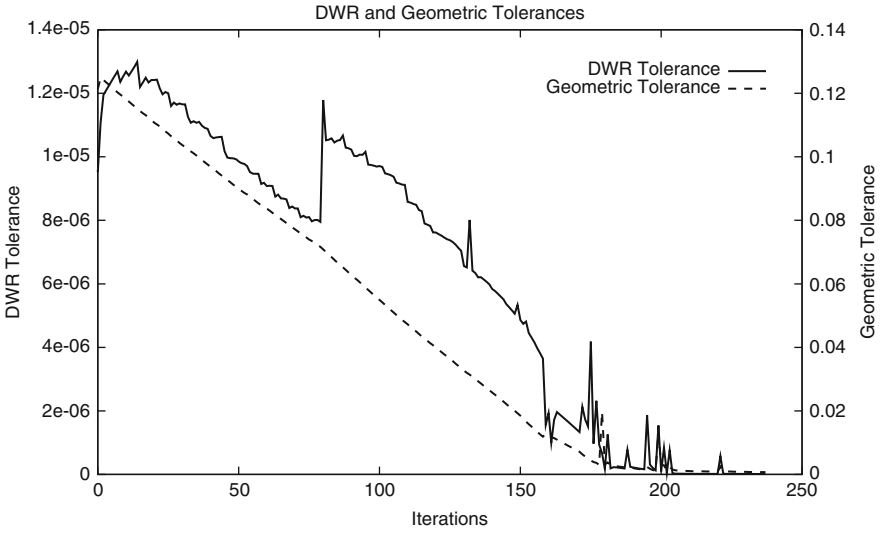


Fig. 2 Dynamic tolerance for both Geometric and PDE approximation: the adaptive SQP method produces coarse meshes at the beginning and a combination of Geometric and PDE refinement/-coarsening later on (see Fig. 3). The zig-zag behavior in the tolerance is due to the combination of refinement/coarsening. Coarsening allows the tolerance to increase (see Table 1)

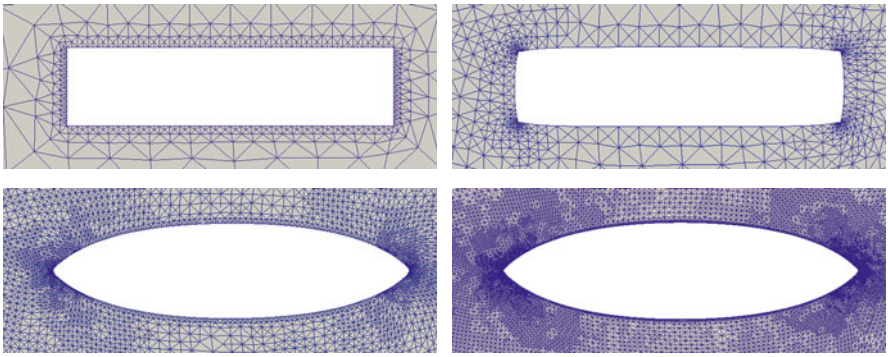


Fig. 3 Combination of DWR and LB refinement/coarsening. Evolution of the initial configuration Γ_s: iterations 0, 44, 184 and 217. The initially refined corners (*top*) are subsequently smoothed out and coarsened (see Fig. 4). The new corners of the rugby ball, instead, are genuine singularities and are preserved and further refined by ASQP (*bottom*)

of Γ by piecewise polynomials. Since Γ is unknown in this context, we mimic the W^1_∞ error between true and discrete surface by properly scaled jumps of the normal vector to the discrete surface. More precisely, the error indicator associated to element T of the k-th surface Γ_k is given by

$$\eta_{\Gamma_k}(T)^2 := h_T^2 \|\mathbb{R}(V_k)\|^2_{L^2(T)} + h_T \|\mathcal{J}(V_k)\|^2_{L^2(\partial T)} + \max_{S\subset\partial T} \mathcal{J}_{n,S}^2 \|\nabla_\Gamma V_k\|^2_{L^2(T)},$$

where $\mathbb{R}(V_k) = -\alpha_b \Delta_\Gamma V_k + \beta_b V_k - g_k$ is the so-called *interior residual*, $\mathcal{J}(V_k)$ is the *jump residual*, namely jump of $\nabla_\Gamma V_k$ normal to the edge, and $\mathcal{J}_{n,S}$ is the *jump of the unit normal vector* (to the surface) across the interelement side S (see Fig. 2 and Table 1).

Geometrically Consistent Mesh Modification (GCMM). The presence of corners (or kinks) on the deformable boundary Γ_k is usually problematic. First, the scalar product $b_{\Gamma_k}(\cdot,\cdot)$ of (5) includes a LB regularization term ($\alpha_b > 0$) which stabilizes the boundary update but cannot remove kinks because V_k is smooth (see (14)). Second, DWR regards kinks as true singularities and tries to refine them accordingly. The combination of these two effects leads to numerical artifacts (ear formation) and halt of computations.

The GCMM method of [4] circumvents this issue; see Figure 4. Whenever the boundary mesh Γ_k is to be modified (refine, coarsen, or smooth out), then the discrete curvature H_k of Γ_k is interpolated and the new position X_k of the free boundary is determined from the fundamental geometric identity $-\Delta_{\Gamma_k} X_k = H_k$. This preserves geometric consistency, which is violated by simply interpolating Γ_k, as well as accuracy [4]. In addition, this computation rounds fake kinks (due to numerics) and preserves genuine kinks (see Fig. 5).

Table 1 Number of marked elements for refinement/coarsening according to Laplace–Beltrami (LB) and dual weighted residual (DWR). The adaptive SQP method, with dynamically changing tolerance, alternates refinement/coarsening for LB and DWR. After the first two iterations, where refinement/coarsening takes place, the algorithm performs 80 iterations of optimization without changing the numerical resolution. Later on, the tolerance is modified by a sequence of DWR and LB refinement/coarsening

Iteration	0	1	81	88	128	150	153	160	161	163	173	175	177	179	181	189	196	200	202	204	213	222
DWR-ref	2,000					218		523	2,428	2,112					7,625			3,786	4,566	32,372	1,051	81,657
DWR-coars	777							697	1,284	1,096	174	176	178	180	2,312			1,994	5,355	3,305		2,234
LB-ref	75	44	4	21	5		25	523			29	65	56	819		191	35	1,379		1,002	0	924
LB-coars	88		22		11		14	49			44		4	54		104	11	57		113		1,000

Fig. 4 Detection of genuine geometric singularities. Evolution of the initial *upper-left* corner of Γ_s (see *top* of Figs. 1 and 3): snapshots of iterations 0, 1, 160 and 190. The adaptive SQP method is able to sort out whether geometric singularities are genuine to the problem or just due to lack of numerical resolution and to coarsen overrefined regions of the computational grid

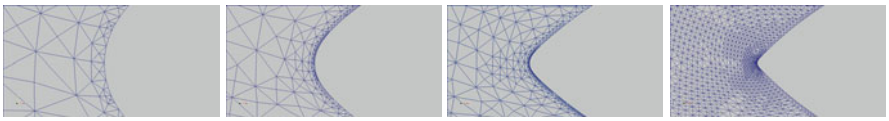


Fig. 5 Detection of genuine geometric singularities. Zoom on the evolution of the left-hand part of the initial configuration Γ_s (see *top* of Fig. 1 and *bottom* of Fig. 3): snapshots of iterations 140,160,180 and 220. The adaptive SQP method is able to recognize the corner of the rugby ball as genuine singularity of the problem and to refine the mesh (combined use of LB and DWR error estimates) to improve both the PDE and the Geometric approximation

Mesh Quality. The mesh is evolved by a prescribed discrete velocity of its boundary. To avoid mesh deterioration a mechanism to maintain good quality must be provided. Remeshing in each iteration is expensive and destroys the binary hierarchical data structure used for refinements and coarsenings [13]. Our approach is to use an optimization routine that works on stars and selectively reallocates the center node so as to improve the star quality and approximately preserve the local mesh size. It does not change the mesh topology so it is compatible with the binary data structure. In each star we minimize the SSU (Simultaneous Smoothing and Untangling) cost functional proposed in [6]. When optimization alone is not sufficient to maintain a good quality we remesh the domain. We refer to [9] for the effect of remeshing.

Time Step. Control of time step is required to satisfy the Armijo conditions (6) as well as to avoid node crossing when evolving the mesh [5]. The latter constraint sometimes dictates the time step, especially when the mesh is fine. We have found that remeshing ameliorates this issue upon drastically improving the mesh quality. In [9] we allow remeshing inside the Armijo condition.

Constraints. The area constraint that defines the class of admissible functions is enforced via a Lagrange multiplier. The algorithm, described in [5], guarantees volume conservation to machine precision in each time iteration and is well suited to be utilized inside the Armijo condition loop.

Acknowledgements This work for Morin has been partially supported by Universidad Nacional del Litoral through Grant CAI+D PI-62-312, and CONICET through Grant PIP 112-200801-02182. This work for Nochetto and Pauletti has been partially supported by NSF grants DMS-0505454 and DMS-0807811. This work for Verani has been partially supported by Italian FIRB RBIP06HF8S.

References

1. P. Alotto, P. Girdinio, P. Molfino, M. Nervi, *Mesh adaption and optimization techniques in magnet design*, IEEE Trans. Magn., 32(4): 2954–2957, 1996
2. W. Bangerth, R. Rannacher, *Adaptive Finite Element Methods for Differential Equations*, Birkhäuser, Boston, 2003
3. N.V. Banichuk, A. Falk, E. Stein, *Mesh refinement for shape optimization*, Struct. Optim., 9:46–51, 1995
4. A. Bonito, R.H. Nochetto, M.S. Pauletti, Geometrically Consistent Mesh Modification, SIAM J. Numer. Anal., to appear.
5. A. Bonito, R.H. Nochetto, M.S. Pauletti, Parametric FEM for Geometric Biomembranes, J. Comput. Phys., 229: 3171–3188, 2010
6. J.M. Escobar, E. Rodriguez, R. Montenegro, G. Montero, J.M. Gonzalez-Yuste, *Simultaneous untangling and smoothing of tetrahedral meshes*, Comput. Meth. Appl. Mech. Eng., 192:2775–2787, 2003
7. A. Henderson, *ParaView Guide, A Parallel Visualization Application* Kitware Inc., 2007
8. K.Mekchay, P.Morin, R.H. Nochetto, AFEM for Laplace Beltrami operator on graphs: Design and conditional contraction property, Math. Comp., to appear
9. P. Morin, R.H. Nochetto, M.S. Pauletti, M. Verani, AFEM for shape optimization (in preparation)
10. O. Pironneau, *On optimum profiles in Stokes flow*, J. Fluid Mech. 59:117–128, 1973
11. J.R. Roche, *Adaptive method for shape optimization*, 6th World Congresses of Structural and Multidisciplinary Optimization, Rio de Janeiro, 2005
12. A. Schleupen, K. Maute, E. Ramm, *Adaptive FE-procedures in shape optimization*, Struct. Multidisc. Optim., 19:282–302, 2000
13. A. Schmidt, K.G. Siebert, *Design of Adaptive Finite Element Software. The Finite Element Toolbox ALBERTA*, Lecture Notes in Computational Science and Engineering 42, Springer, Berlin, 2005
14. J. Sokołowski, J.-P. Zolésio, *Introduction to Shape Optimization*, Springer, Berlin, 1992

Convergence of Path-Conservative Numerical Schemes for Hyperbolic Systems of Balance Laws

M.L. Muñoz-Ruiz, C. Parés, and M.J. Castro Díaz

Abstract We are concerned with the numerical approximation of Cauchy problems for hyperbolic systems of balance laws, which can be studied as a particular case of nonconservative hyperbolic systems. We consider the theory developed by Dal Maso, LeFloch, and Murat to define the weak solutions of nonconservative systems, and path-conservative numerical schemes (introduced by Parés) to numerically approximate these solutions. In a previous work with Le Floch we have studied the appearance of a convergence error measure in the general case of noconservative hyperbolic systems, and we have noticed that this lack of convergence cannot always be observed in numerical experiments. In this work we study the convergence of path-conservative schemes for the special case of systems of balance laws, specifically, the experiments performed up to now show that the numerical solutions converge to the right weak solutions for the correct choice of path-conservative scheme.

1 Introduction

We are interested in the numerical approximation of the initial value problem for hyperbolic systems that have the form

$$w_t + F(w)_x = S(w)\sigma_x \ , \quad x \in \mathbf{R},\ t > 0 \ , \tag{1}$$

where $w(x,t) \in \Omega$, an open convex set of $\mathbf{R}^N$, $\sigma(x)$ is a known function from $\mathbf{R}$ to $\mathbf{R}$, F is a regular function from Ω to $\mathbf{R}^N$, and S is also a function from Ω to $\mathbf{R}^N$. These systems are called hyperbolic systems of conservation laws with source term or balance laws.

M.L. Muñoz-Ruiz (✉), C. Parés, and M.J. Castro Díaz
Dpt. Análisis Matemático, Universidad de Málaga, 29071-Málaga, Spain
e-mail: munoz@anamat.cie.uma.es, pares@anamat.cie.uma.es, castro@anamat.cie.uma.es

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_72,

Following an idea by LeFloch [3], the equation

$$\sigma_t = 0 \tag{2}$$

is added to the system in order to rewrite it in the quasilinear form

$$W_t + A(W)\, W_x = 0\,, \tag{3}$$

with

$$W = \begin{bmatrix} W \\ \sigma \end{bmatrix} \tag{4}$$

and

$$A(W) = \begin{bmatrix} J(W) & -S(W) \\ 0 & 0 \end{bmatrix}, \tag{5}$$

being

$$J(W) = \frac{\partial F}{\partial W}(W)\,. \tag{6}$$

To begin with, we point out below the non trivial difficulties that arise in the study of systems (3) with general W and A (not necessarily with the structure (4)–(6)), which are related to the presence of the nonconservative products $A(W)W_x$. In fact, when the solutions to (3) are discontinuous, which is the common feature, these products have no sense in the distributional framework and the usual concept of weak solution can not be used.

We suppose that system (3) is strictly hyperbolic and that the characteristic fields are either genuinely nonlinear or linearly degenerate, and we assume the definition of nonconservative products introduced by Dal Maso, LeFloch and Murat in [2], which is associated to the choice of a family of paths in the phase space Ω. A family of paths in Ω is a locally Lipschitz map $\Phi\colon [0,1]\times\Omega\times\Omega\to\Omega$ that satisfies

$$\Phi(0; W_L, W_R) = W_L \text{ and } \Phi(1; W_L, W_R) = W_R\,, \text{ for any } W_L, W_R \in \Omega\,,$$

together with certain regularity conditions. Once a family of paths Φ is chosen, the nonconservative product $A(W)\,W_x$ can be defined as a bounded measure for $W \in (L^\infty(\mathbf{R}\times\mathbf{R}^+)\cap BV(\mathbf{R}\times\mathbf{R}^+))^N$. A notion of weak solution (consequently depending on the chosen paths) can now be given. Across a discontinuity, a weak solution satisfies the generalized Rankine–Hugoniot condition

$$\int_0^1 \big(\xi I - A(\Phi(s; W^-, W^+))\big)\Phi_s(s; W^-, W^+)\, ds = 0\,, \tag{7}$$

where ξ is the speed of propagation of the discontinuity, I is the identity matrix, and W^- and W^+ are the left and right limits of the solution at the discontinuity. Notice that, when $A(W)$ is the Jacobian matrix of some flux function $F(W)$ we are concerned with a conservation law, and the proposed definition of nonconservative

product does not depend on the choice of paths and coincides with the distributional derivative of $F(W)$. In addition, the definition of the weak solution reduces to the distributional one and the generalized Rankine–Hugoniot condition to the usual one. A notion of entropy weak solution can also be given and the classic theory for hyperbolic systems of conservation laws concerning simple waves and the solutions of Riemann problems can be extended to nonconservative systems (3) (see [2,4] for details).

The choice of the family of paths is highly important, as it determines the speed of propagation of discontinuities. It should be based on the physical background of the problem, taking into account the effects related to dispersion, diffusion, ... but this seems to be a difficult task in practical applications. On the other hand, some mathematical conditions can be required for the paths in order to obtain good properties for the related weak solutions [5].

Concerning the discretization of system (3) together with an initial condition $W(x,0) = W_0(x),\ x \in \mathbf{R}$, computing cells $I_i = [x_{i-1/2}, x_{i+1/2}]$ with constant size Δx are considered, being $x_{i+\frac{1}{2}} = i\Delta x$ and $x_i = (i-1/2)\Delta x$, the center of the cell I_i. Let Δt be the constant time step and define $t^n = n\Delta t$. The approximation of the cell averages of the exact solution obtained by means of the numerical scheme is denoted by W_i^n. In this work we consider the class of numerical schemes called path-conservative that was proposed in [6]. A numerical scheme is said to be Ψ-conservative, for any given family of paths Ψ, if it can be written in the form

$$W_i^{n+1} = W_i^n - \frac{\Delta t}{\Delta x}\left(D_{i-1/2}^{n,+} + D_{i+1/2}^{n,-}\right), \tag{8}$$

where $D_{i+1/2}^{n,\pm} = D^{\pm}\left(W_{i-q}^n, \ldots, W_{i+p}^n\right)$ and D^- and D^+ are two Lipschitz-continuous functions from Ω^{p+q+1} to Ω satisfying

$$D^{\pm}(W, \ldots, W) = 0 \quad \forall W \in \Omega\ , \tag{9}$$

and

$$D^-(W_{-q}, \ldots, W_p) + D^+(W_{-q}, \ldots, W_p) = \int_0^1 A\big(\Psi(s; W_0, W_1)\big)\Psi_s(s; W_0, W_1)\, ds\ , \tag{10}$$

for every $W_i \in \Omega,\ i = -q, \ldots, p$.

This definition generalizes that of conservative scheme for conservative problems: in the particular case of a system of conservation laws, a numerical scheme is conservative if and only if it is Ψ-conservative for any family of paths Ψ.

Although it should be desirable to use the same family of paths to define weak solutions and to construct the numerical scheme, it is not always possible or reasonable from the computational point of view the construction of Φ-conservative schemes, as the election of Φ is restricted by the matematical conditions mentioned above.

In the following section we recall the conditions imposed for the paths in the case of systems of balance laws and some results concerning path-conservative schemes for this particular case of nonconservative systems. In Sect. 3 some aspects concerning the convergence of path-conservative schemes when applied to balance laws are discussed and they are illustrated with some numerical experiments.

2 Systems of Balance Laws

In this section we emphasize in some aspects concerning the definition of weak solutions of systems of balance laws and their approximation by means of a path-conservative scheme. For that, we consider systems (1) written in the form (3)–(6).

If $J(W)$ has N distinct real and non-vanishing eigenvalues $\lambda_1(W), \ldots, \lambda_N(W)$, then the system (3)–(6) is strictly hyperbolic with eigenvalues $\lambda_1(W), \ldots, \lambda_N(W), 0$. We will assume that the characteristic field associated to the zero eigenvalue is the only one that is linearly degenerate.

In order to define weak solutions to this nonconservative system a family of paths

$$\Phi(s; W_L, W_R) = \begin{bmatrix} \Phi_w(s; W_L, W_R) \\ \Phi_\sigma(s; W_L, W_R) \end{bmatrix}$$

in $\Omega \times \mathbf{R}$ has to be chosen.

It is convenient to impose the family of paths to satisfy that

$$\Phi_\sigma(s; W_L, W_R) = \bar{\sigma} \ , \quad s \in [0, 1] \ , \tag{11}$$

for all W_L and W_R such that $\sigma_L = \sigma_R = \bar{\sigma}$, as it assures that, if W is a weak solution of the nonconservative system (3)–(6) with constant σ, $\sigma(x) = \bar{\sigma}$, then w is a weak solution of the conservative problem $w_t + F(w)_x = 0$.

We also assume that the path connecting two states that belong to the same integral curve of the linearly degenerate field is a parametrization of the arc of the integral curve linking them. And finally, we assume that the path connecting two states for wich the associated Riemann problem has a unique self-similar weak solution composed by simple waves linking constant intermediate states, is equal to the union of the paths linking the intermediate states [5].

In fact, the previous requirements on the family of paths completely determine the notion of weak solution for these systems, which contain two types of discontinuities: shock waves across which σ is continuous that satisfy the usual Rankine–Hugoniot conditions and stationary contact discontinuities placed at the jumps of σ that connect two states belonging to the same integral curve of the linearly degenerate field.

In order to numerically approximate the initial value problem we consider path-conservative schemes. For that, we choose a family of paths

$$\Psi(s; W_L, W_R) = \begin{bmatrix} \Psi_w(s; W_L, W_R) \\ \Psi_\sigma(s; W_L, W_R) \end{bmatrix}.$$

The path-conservative schemes constructed using the family of paths Φ used to define weak solutions can be characterized by the equality

$$D^-\big(W_{-q},\dots,W_p\big)+D^+\big(W_{-q},\dots,W_p\big) = F(w_1)-F(w^+_{1/2})+F(w^-_{1/2})-F(w_0)\ , \tag{12}$$

where

$$W^-_{1/2} = \begin{bmatrix} w^-_{1/2} \\ \sigma_0 \end{bmatrix} \qquad \text{and} \qquad W^+_{1/2} = \begin{bmatrix} w^+_{1/2} \\ \sigma_1 \end{bmatrix}$$

are the left and right limits at $x = 0$ of the solution of the Riemann problem that has W_0 and W_1 as initial condition. This equality also allows us to prove that a Φ-conservative numerical scheme reduces to a conservative one where σ is constant.

In fact, at least for $q = 0$ and $p = 1$, the Ψ-conservative schemes which reduce to a conservative scheme in regions where σ is constant can be characterized and we can prove that they can be written in the form

$$w_i^{n+1} = w_i^n - \frac{\Delta t}{\Delta x}\big(G^n_{i+1/2} - G^n_{i-1/2}\big) + \frac{\Delta t}{\Delta x}\big(H^{n,+}_{i-1/2} + H^{n,-}_{i+1/2}\big)\ , \tag{13}$$

where $G^n_{i+1/2} = G\big(w_i^n, w_{i+1}^n\big)$ and $H^{n,\pm}_{i+1/2} = H^{\pm}\big(W_i^n, W_{i+1}^n\big)$, being G, H^- and H^+ Lipschitz continuous functions such that

$$G(w,w) = F(w)\ , \tag{14}$$

$$H^{\pm}\big(W,W\big) = 0\ , \tag{15}$$

$$H^{\pm}\big(\overline{W}_0,\overline{W}_1\big) = 0\ , \tag{16}$$

and

$$H^-\big(W_0,W_1\big) + H^+\big(W_0,W_1\big) = \int_0^1 S(\Psi_w(s;W_0,W_1))\,(\Psi_\sigma)_s\,(s;W_0,W_1))\,ds\ . \tag{17}$$

The notation $\overline{W}$ has been used for values W such that $\sigma = \bar\sigma$, being $\bar\sigma$ a fixed value.

In addition, a necessary condition for a Ψ-conservative scheme to be well-balanced (see [7] for more details on the well-balance property) is that (12) is satisfied for every W_0 and W_1 in an integral curve of the linearly degenerate field.

3 Convergence Properties

The convergence of path-conservative schemes for general nonconservative systems (3) has been recently studied in [1]. In fact, it has been proven that, if the approximations generated by a path-conservative scheme converge in the uniform sense of graphs (see [2] for more details on this notion of convergence), then the

limit is a weak solution of the nonconservative system. Unfortunately, this convergence is rather strong and usually fails in applications (although it holds for random choice methods). It can also be proved that, if the approximations just converge almost everywhere (which is a more realistic assumption from the practical point of view), a convergence error term appears in the limiting system.

In the particular case of systems of balance laws this lack of convergence is not appreciated in the cases presented below. The numerical experiments have been performed for the particular example of the shallow water system governing the flow of a shallow layer of inviscid homogeneous fluid through a straight channel with a constant rectangular cross-section:

$$\begin{aligned} &\frac{\partial h}{\partial t} + \frac{\partial q}{\partial x} = 0 \,, \\ &\frac{\partial q}{\partial t} + \frac{\partial}{\partial x}\left(\frac{q^2}{h} + \frac{g}{2}h^2\right) = gh\frac{\mathrm{d}H}{\mathrm{d}x} \,. \end{aligned}$$

In these equations $q(x,t)$ represents the mass-flow and and $h(x,t)$, the thickness, g is the gravity, and $H(x)$ is the depth measured from a fixed level of reference.

The first experiment concerns a dam-break problem over a non-flat bottom topography given by $H(x) = 1 - 0.5e^{-(x-5)^2}$. The initial condition considered is

$$q(x,0) = 0 \,, \qquad h(x,0) = \begin{cases} H(x) \,, & x \geq 4 \,, \\ H(x) + 0.5 \,, & x < 4 \,. \end{cases}$$

The approximations have been obtained by means of a Roe and a modified Lax–Friedrichs scheme, path-conservative with respect to a very simple family of paths: the family of segments. In Fig. 1 we observe that both schemes converge to the same solution as the mesh is refined.

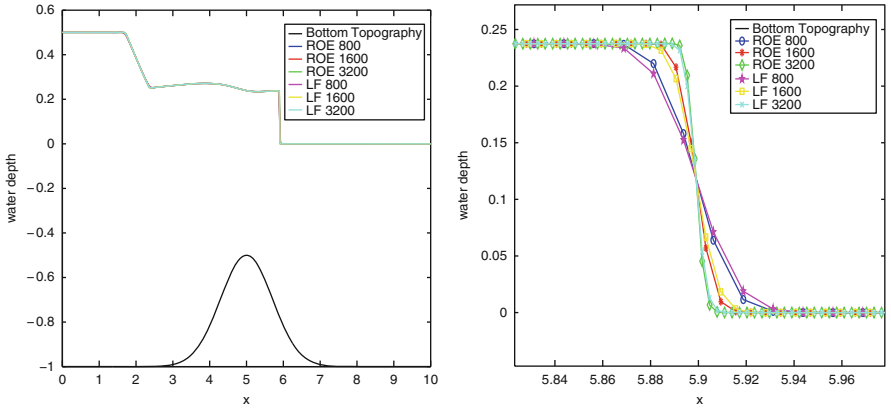


Fig. 1 Dam-break problem over a non-flat bottom topography. *Left*: bottom topography and free surface. *Right*: free surface (zoom)

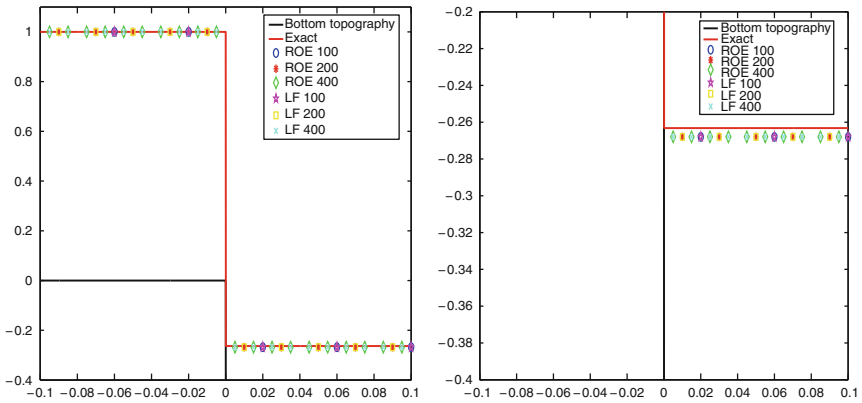


Fig. 2 Stationary contact discontinuities. *Left*: bottom topography and free surface (stationary solution). *Right*: zoom

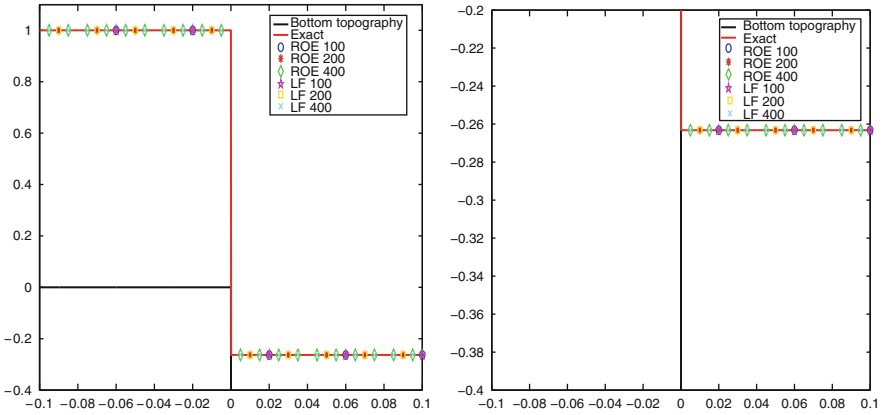


Fig. 3 Stationary contact discontinuities. *Left*: bottom topography and free surface (stationary solution) obtained with exactly well-balanced schemes. *Right*: zoom

The convergence of the approximations generated by a Φ-conservative scheme, provided $\sigma \in W^{1,1}$, the approximations are bounded in $L^{\infty}_{\text{loc}}(\mathbf{R} \times [0,\infty))^N$ and converge to some function in $L^{1}_{\text{loc}}(\mathbf{R} \times [0,\infty))^N$ can be proved (a previous result of that type for a class of well-balanced numerical schemes for solving scalar conservation laws can be found in [8]). In fact, when σ is continuous, the approximations provided by any Ψ-conservative scheme that can be written in the form (13)–(17) converge to the weak solution of the nonconservative problem.

The second experiment concerns the approximation of stationary contact discontinuities. The bottom topography is now given by

$$H(x) = \begin{cases} 0\,, & x < 0, \\ 1\,, & x \geq 0, \end{cases}$$

and the initial condition is a pair of states in the same integral curve of the linearly degenerate field. We first use a Roe and a modified Lax–Friedrichs scheme, path-conservative with respect to the family of segments. We observe in Fig. 2 that both schemes converge to the same discontinuous function, but not to the exact solution.

Nevertheless, if we use a more appropriate family of paths to construct the numerical schemes, a family such that the path linking two states in the same integral curve of the linearly degenerate field is an arc of the integral curve (that is also the necessary condition for well-balancing), both Roe and the modified Lax–Friedrichs scheme capture the stationary contact discontinuity exactly. This is observed in Fig. 3.

This leads us to the conclusion that, in certain special situations, as systems of balance laws, if the family of paths satisfies the condition related to the linearly degenerate characteristic fields stated above, then all of the discontinuities are correctly approximated and the scheme converge to exact solutions.

Acknowledgements This research has been partially supported by the Spanish Government Research project MTM2006-08075.

References

1. Castro, M. J., LeFloch, P. G., Muñoz-Ruiz, M. L., Parés, C.: Why many theories of shock waves are necessary. Convergence error in formally path-consistent schemes. J. Comput. Phys. **227**, 8107–8129 (2008)
2. Dal Maso, G., LeFloch, P. G., Murat, F.: Definition and weak stability of nonconservative products. J. Math. Pures Appl. **74**, 483–548 (1995)
3. LeFloch, P. G.: Shock waves for nonlinear hyperbolic systems in nonconservative form. Inst. Math. Appl. Minneap. **593** (1989)
4. LeFloch, P. G., Liu, T. P.: Existence theory for nonlinear hyperbolic systems in nonconservative form. Forum Math. **5**, 261–280 (1993)
5. Muñoz-Ruiz, M. L., Parés, C.: Godunov method for nonconservative hyperbolic systems. M2AN Math. Model. Numer. Anal. **41**, 169–185 (2007)
6. Parés, C.: Numerical methods for nonconservative hyperbolic systems: a theoretical framework. SIAM J. Numer. Anal. **44**, 300–321 (2006)
7. Parés, C., Castro, M. J.: On the well-balance property of Roe's method for nonconservative hyperbolic systems. Applications to shallow water systems. M2AN Math. Model. Numer. Anal. **38**, 821–852 (2004)
8. Perthame, B., Simeoni, C.: Convergence of the upwind interface source method for hyperbolic conservation laws. In: Proc. of Hyp 2002, pp. 61–78. Springer, Berlin (2003)

A Two-Level Newton–Krylov–Schwarz Method for the Bidomain Model of Electrocardiology

M. Munteanu, L.F. Pavarino, and S. Scacchi

Abstract A two-level Newton–Krylov–Schwarz (NKS) solver is constructed and analyzed for implicit time discretizations of the Bidomain reaction-diffusion system. This multiscale system describes the bioelectrical activity of the heart by coupling two degenerate parabolic equations with several ordinary differential equations at each point in space. The proposed NKS Bidomain solver employs an outer inexact Newton iteration to solve the nonlinear finite element system originating at each time step of the implicit discretization. The Jacobian update during the Newton iteration is solved by a Krylov method employing a two-level overlapping Schwarz preconditioner. A convergence rate estimate is proved for the resulting preconditioned operator, showing that its condition number is independent of the number of subdomains (scalability) and bounded by the ratio of the subdomains characteristic size and the overlap size. This theoretical result is confirmed by several parallel simulations employing up to more than 2,000 processors for scaled and standard speedup tests in three dimensions.

1 Introduction

The aim of this work is the construction and analysis of a two-level overlapping Schwarz preconditioner that leads to a scalable Newton–Krylov–Schwarz (NKS) solver for the Bidomain system of electrocardiology.

The Bidomain system consists of two degenerate parabolic reaction-diffusion equations modeling the evolution of the intra- and extracellular potentials of the anisotropic cardiac tissue (macroscale), coupled through the nonlinear reaction term with a stiff system of ordinary differential equations describing the ionic currents evolution through the cellular membrane (microscale). The numerical solution of this coupled multiscale reaction-diffusion model is very expensive and in order

M. Munteanu, L.F. Pavarino, and S. Scacchi (✉)
Department of Mathematics, University of Milan, via Saldini 50, 20133 Milan, Italy
e-mail: marilena_munteanu@hotmail.com, luca.pavarino@unimi.it, simone.scacchi@unimi.it

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_73,

to reduce the computational costs involved many previous works have considered semi-implicit (Imex) time discretizations and/or operator splitting schemes, where the reaction and diffusion terms are treated separately, see e.g., [3, 4, 7, 11, 13–15, 17, 20–22]. In an attempt to relax the stability constraints of these methods, some authors have proposed fully implicit [10] or decoupled implicit [8] Bidomain discretizations, requiring the solution of a nonlinear system at each time step.

In this paper, we construct a scalable NKS Bidomain solver based on a two-level overlapping Schwarz preconditioner. The condition number of the resulting preconditioned operator is independent of the number of subdomains (scalability) and bounded by the ratio of the subdomains characteristic size and the overlap size. This result is confirmed by parallel simulations employing more than 2,000 processors for scaled and standard speedup tests in three dimensions. A theoretical analysis of the proposed NKS Bidomain solver and additional parallel tests can be found in [9].

2 The Bidomain Model

The macroscopic Bidomain model represents the cardiac tissue as the superposition of two anisotropic continuous media, the intra (i) and extra (e) cellular media, coexisting at every point of the tissue and separated by a distributed continuous cellular membrane. The intra- and extracellular electric potentials u_i, u_e in the cardiac domain Ω are described in the Bidomain model by the following parabolic reaction-diffusion system coupled with a system of ODEs for the ionic variables w:

$$\begin{cases} c_m \dfrac{\partial v}{\partial t} - \operatorname{div}(\mathbf{D}_i \nabla u_i) + I_{ion}(v,w) = I^i_{app} & \text{in } \Omega \times (0,T) \\ -c_m \dfrac{\partial v}{\partial t} - \operatorname{div}(\mathbf{D}_e \nabla u_e) - I_{ion}(v,w) = I^e_{app} & \text{in } \Omega \times (0,T) \\ \dfrac{\partial w}{\partial t} - R(v,w) = 0, & \text{in } \Omega \times (0,T), \end{cases} \tag{1}$$

with boundary conditions $\mathbf{n}^T \mathbf{D}_{i,e} \nabla u_{i,e} = 0$ on $\partial\Omega \times (0,T)$ and initial conditions $v(\mathbf{x},0) = v_0(\mathbf{x}), w(\mathbf{x},0) = w_0(\mathbf{x})$ in Ω. Here c_m is the capacitance per unit area times the surface to volume ratio; $v = u_i - u_e$ is the transmembrane potential; $\mathbf{D}_{i,e}$ are the anisotropic intra- and extracellular conductivity tensors; $I^{i,e}_{app}$ are applied currents. I_{ion} and R model the ionic currents and depend on the choice of membrane model, here assumed to be the Luo-Rudy phase I (LR1) model [6]. We refer to e.g., [4, 12, 19] for a mathematical analysis of the Bidomain system under the compatibility conditions $\int_\Omega (I^i_{app} + I^e_{app}) dx = 0$.

3 Discretization and Numerical Methods

System (1) is discretized by the finite element method in space and a decoupled implicit method in time. The space discretization is obtained by meshing the cardiac domain Ω with a structured grid $\mathcal{T}_h$ of hexahedral Q_1 elements and introducing the associated finite element space V_h. A semidiscrete problem is obtained by applying a standard Galerkin procedure. Let M be the symmetric mass matrix, $\mathrm{A}_{i,e}$ the symmetric stiffness matrices associated with the intra and extra-cellular anisotropic conductivity tensors, respectively, and I^h_{ion}, $\mathrm{I}^{i,e,h}_{app}$ the finite element interpolants of I_{ion} and $I^{i,e}_{app}$, respectively. The time discretization is performed by the following implicit method with time step τ. Given $\mathbf{v}^n = \mathbf{u}^n_i - \mathbf{u}^n_e$ and $\mathbf{w}^n$ at time t_n,

a. Solve for $\mathbf{w}^{n+1}$ the ODE ionic system: $\mathbf{w}^{n+1} - \tau\, R(\mathbf{v}^n, \mathbf{w}^{n+1}) = \mathbf{w}^n$;
b. Solve for $\mathbf{u}^{n+1} = (\mathbf{u}^{n+1}_i, \mathbf{u}^{n+1}_e)$ the nonlinear system:

$$\mathbf{F}(\mathbf{u}^{n+1}) = \left(\frac{c_m}{\tau}\begin{bmatrix} \mathrm{M} & -\mathrm{M} \\ -\mathrm{M} & \mathrm{M} \end{bmatrix} + \begin{bmatrix} \mathrm{A}_i & 0 \\ 0 & \mathrm{A}_e \end{bmatrix}\right)\begin{pmatrix} \mathbf{u}^{n+1}_i \\ \mathbf{u}^{n+1}_e \end{pmatrix} +$$

$$\begin{pmatrix} \mathrm{M}[\ \mathbf{I}^h_{ion}(\mathbf{v}^{n+1}, \mathbf{w}^{n+1}) - \mathbf{I}^{i,h}_{app}] \\ \mathrm{M}[-\mathbf{I}^h_{ion}(\mathbf{v}^{n+1}, \mathbf{w}^{n+1}) - \mathbf{I}^{e,h}_{app}] \end{pmatrix} - \frac{c_m}{\tau}\begin{pmatrix} \mathrm{M}[\ \mathbf{u}^n_i - \mathbf{u}^n_e] \\ \mathrm{M}[-\mathbf{u}^n_i + \mathbf{u}^n_e] \end{pmatrix} = 0. \qquad (2)$$

4 A Newton–Krylov–Schwarz (NKS) Bidomain Solver

The nonlinear system (2) arising at each time iteration of the decoupled implicit method, described in the previous section, is solved by a nested Newton–Krylov–Schwarz (NKS) method, see e.g., [2,5]. In this class of methods, a Newton scheme is used as outer iteration and the Jacobian linear system arising at each Newton iteration is solved by a Krylov method with a Schwarz-type preconditioner. In this paper, we will consider the Preconditioned Conjugate Gradient (PCG) method accelerated by a two-level overlapping Additive Schwarz preconditioner.

The outer Newton iteration The outer Newton iteration reads as follows:

a. Choose a starting value $\mathbf{u}^0 = (\mathbf{u}^0_i, \mathbf{u}^0_e)$;
b. For $k \geq 0$ and until a Newton stopping criterion is met, find the solution $\mathbf{s}^{k+1} = (\mathbf{s}^{k+1}_i, \mathbf{s}^{k+1}_i)$ of the Jacobian linear system:

$$\mathcal{J}^k \mathbf{s}^{k+1} = -\mathbf{F}(\mathbf{u}^k), \qquad (3)$$

where $\mathbf{u}^k = (\mathbf{u}^k_i, \mathbf{u}^k_e)$ is the finite element approximation of the intra- and extracellular potentials at the k^{th}–Newton iteration at the current time step

and $\mathscr{J}^k$ is the Jacobian of $\mathbf{F}(\cdot)$ computed in $\mathbf{u}^k$. Then, choose $\mathbf{s}_e^{k+1}$ such that $\int_\Omega s_e^{k+1} dx = 0$;

c. Update the Newton solution: $\mathbf{u}^{k+1} = \mathbf{u}^k + c_k \mathbf{s}^{k+1}$, with the scaling factor c_k determined by a line search technique (PETSc default Newton update, see [1]).

By defining the subspace of zero average finite element functions

$$\widetilde{V}_h = \{\varphi \in V_h : \int_\Omega \varphi = 0\} \quad \text{and the product space } \mathbf{V}_h = V_h \times \widetilde{V}_h,$$

the Jacobian system (3) can be written in an abstract elliptic variational form, by introducing the bilinear form $a_{bid}(\cdot,\cdot) : \mathbf{V}_h \times \mathbf{V}_h \longrightarrow R$, defined by

$$a_{bid}(s^{k+1}, \phi) = (s_i^{k+1} - s_e^{k+1}, \varphi_i - \varphi_e) + \tau a_i(s_i^{k+1}, \varphi_i) + \tau a_e(s_e^{k+1}, \varphi_e) +$$
$$+ \tau \left(\sum_{l=1}^{N} \frac{\partial I_{ion}}{\partial v_l}(v_l^k)(s_{il}^{k+1} - s_{el}^{k+1})\varphi_l, \varphi_i - \varphi_e \right),$$

where $a_{i,e}(\cdot,\cdot)$ are the H^1-bilinear forms induced by the diffusion tensors $D_{i,e}$, $(\cdot,\cdot)$ is the L^2-inner product and N is the dimension of the finite element space V_h.

A two-level Additive Schwarz preconditioner for the NKS Jacobian system Following the classical overlapping Schwarz theory (see [16, 18]), applied to the variational problem associated to the bilinear form a_{bid}, we construct the two-level Additive Schwarz (AS) operator $\mathbf{T}_{AS}$ and we estimate its condition number. We refer to [9] for further details and a proof of results below. Let $\mathscr{T}_S$ be a coarse shape-regular triangulation of Ω constituted by N_S nonoverlapping hexahedral subdomains Ω_m, $m = 1, \ldots, N_S$, of diameter H_m and set $H_S = \max_m H_m$. We assume that $\mathscr{T}_S$ is such that the fine triangulation $\mathscr{T}_1 = \mathscr{T}_h$, introduced in Sect. 3, is nested in $\mathscr{T}_S$. Let then $\mathscr{T}_0$ be an additional coarse triangulation of Ω, nested in $\mathscr{T}_S$, finer than or equal to $\mathscr{T}_S$ and coarser than $\mathscr{T}_1$, i.e., $\mathscr{T}_S \subseteq \mathscr{T}_0 \subset \mathscr{T}_1$. Let H denote the characteristic mesh size of $\mathscr{T}_0$. The standard technique of adding to each subdomain Ω_m all the fine elements $\tau_j \in \mathscr{T}_1$ within a distance δ from its boundary $\partial\Omega_m$ is used in order to construct an overlapping partition of Ω. Ω'_m denotes the overlapping subdomains obtained by such extension of each Ω_m. Associated with each subdomain Ω'_m, we define the following local finite element spaces

$$V_m := \{u_i \in V_h \,:\, u_i(x) = 0 \; x \in \Omega \backslash \Omega'_m\}, \quad \text{and} \quad \mathbf{V}_m := V_m \times V_m.$$

Let V_0 be the coarse space of trilinear finite elements associated to the coarse triangulation $\mathscr{T}_0$, $\widetilde{V}_0 = V_0 \cap \widetilde{V}_h$ and $\mathbf{V}_0 := V_0 \times \widetilde{V}_0$. Let us introduce the standard local interpolation operators $R_m^T : V_m \longrightarrow V_h$, $\mathbf{R}_m^T = (R_m^T, R_m^T) : \mathbf{V}_m \longrightarrow V_h \times V_h$ for $m = 1, ..., N_S$, the coarse to fine interpolation operator $\mathbf{R}_0^T : \mathbf{V}_0 \longrightarrow \mathbf{V}_h$.

Following the abstract Schwarz framework, see e.g., [18, Chap. 2], we define the projection-like operators $\widetilde{\mathbf{T}}_m : \mathbf{V}_h \longrightarrow \mathbf{V}_m$, $\quad m = 0, 1, \ldots, N_S$, by

$$a_{bid}(\mathbf{I}_s\mathbf{R}_m^T\widetilde{\mathbf{T}}_m u, \mathbf{I}_s\mathbf{R}_m^T v) = a_{bid}(u, \mathbf{I}_s\mathbf{R}_m^T v) \quad \forall v \in \mathbf{V}_m, \quad m = 0, 1, \ldots, N_S,$$

and $\mathbf{T}_m : \mathbf{V}_h \longrightarrow \mathbf{I}_s\mathbf{R}_m^T\mathbf{V}_m \subset \mathbf{V}_h$ by $\mathbf{T}_m = \mathbf{I}_s\mathbf{R}_m^T\widetilde{\mathbf{T}}_m$.
The two-level Additive Schwarz (AS) operator is defined as

$$\mathbf{T}_{AS} := \sum_{m=0}^{N_S} \mathbf{I}_s \left(\mathbf{R}_m^T \mathbf{A}^{-1} \mathbf{R}_m\right) \mathscr{A}_{bid}, \tag{4}$$

where $\mathscr{A}_{bid}$ is the matrix operator associated to $a_{bid}(\cdot,\cdot)$, $\mathbf{A}_m = \mathbf{R}_m\mathscr{A}_{bid}\mathbf{R}_m^T$, and $\mathbf{I}_s$ is a shift operator that restore the zero average property in the extracellular component. See [9] for further details and a proof of the following result.

Theorem 1. *In the hypotheses of Lemma 5.4 of [8], the condition number of the two-level Additive Schwarz operator for the NKS Bidomain system is bounded by*

$$\kappa_2(\mathbf{T}_{AS}) \leq C\left(1 + \frac{H}{\delta}\right),$$

with a constant C independent of the mesh size h, subdomain size H, overlap size δ.

5 Numerical Results

We now present the results of parallel numerical experiments performed on the Linux Cluster IBM BCX/5120 of the Cineca Consortium (www.cineca.it). Our FORTRAN code is based on the parallel library PETSc, from the Argonne National Laboratory [1].

The Bidomain system coupled to the LR1 membrane model is integrated by the decoupled implicit method described in the previous sections. At each time step, the nonlinear system is solved by an inexact Newton scheme. The Newton initial guess is the solution at the previous time step and the stopping criterion is a 10^{-4} reduction of the residual l^2-norm. The symmetric Jacobian linear system at each Newton iteration is solved by the preconditioned conjugate gradient method, with zero initial guess, stopping criterion a 10^{-4} reduction of the relative residual l^2-norm, and preconditioned by the two-level AS preconditioner. Inexact ILU(0) local solvers are used for the local problems on the subdomains, while the coarse problem is solved in parallel by PCG with Block-Jacobi preconditioner with ILU(0) on each block, run to machine precision reduction of the relative residual.

Computational domain The domain Ω is either a cartesian slab or the image of a cartesian slab using ellipsoidal coordinates, yielding a portion of a truncated ellipsoid (see Fig. 1). These two choices allow us also to test the performance of the two-level NKS solver in absence or presence of severe domain deformations.

Fine and coarse meshes For both types of domains (cartesian slabs and truncated ellipsoids), we denote the cartesian mesh used by $\mathscr{T} = \mathscr{T}_i \times \mathscr{T}_j \times \mathscr{T}_k$,

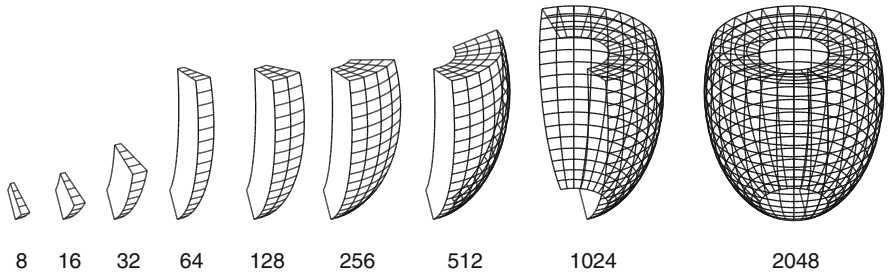


Fig. 1 Ellipsoidal domains for scaled speedup Test 1 (Table 1), decomposed in 8, 16, ..., 2,048 subdomains, each one composed of $32 \times 32 \times 32$ finite elements (not shown)

indicating the number of elements in each coordinate direction. This notation applies to both fine and coarse meshes. When we scale up the mesh by a factor c, for brevity we define $c\mathcal{T} = c\mathcal{T}_i \times c\mathcal{T}_j \times c\mathcal{T}_k$, i.e., the number of elements in $c\mathcal{T}$ is c^3 times the number of elements in $\mathcal{T}$.

Stimulation site, initial and boundary conditions The depolarization process is started by applying a stimulus of $I_{app} = 200\,\mu\text{A}\,\text{cm}^{-3}$ lasting 1 ms on the face of the domain modeling the endocardial surface. The initial conditions are at resting values for all the potentials and LR1 gating variables, while the boundary conditions are for insulated tissue.

Test 1: Scaled speedup on ellipsoidal domains We consider a scaled speedup test on deformed ellipsoidal domains, shown in Fig. 1. The local size of each subdomain on the finest mesh is kept fixed at the value $32 \times 32 \times 32$ (before adding the overlap) and the number of processors (procs.) is increased from 8 to 2,048. The corresponding processor meshes vary from $\mathcal{T}_S = 1 \times 2 \times 4$ to $32 \times 16 \times 4$, forming increasing portions of ellipsoidal domains Ω as shown in Fig. 1. The fine mesh is chosen proportionally to the processor mesh as $\mathcal{T}_1 = 32\mathcal{T}_S$ so as to keep the local mesh size on each processor fixed at $32 \times 32 \times 32$. The coarse mesh is chosen as $\mathcal{T}_0 = 2\mathcal{T}_S$ and the overlap size is $\delta = h$. The simulation is run for 3 time steps of 0.05 ms during the depolarization phase.

Table 1 reports the average number of Newton iterations per time step (*nit*), the average number of PCG iteration per Newton iteration (*lit*) and average CPU times (*ltime*) per Newton iteration in seconds.

These results confirm the scalability of the two-level NKS Bidomain solver (see Theorem 1), because *nit = 2* and *lit* is bounded by 24. The average CPU times *ltime* show a slight growth with the processor count but only by a factor of about 3.4 in comparison with the factor 256 in the processor and d.o.f. growth.

Test 2: Standard speedup on cartesian domains, complete cardiac cycle We now consider the simulation of a complete cardiac cycle over 400 ms with a fixed time step size $\tau = 0.05$ ms, for a total amount of 8,000 time steps. This simulation encompasses all the main phases of a heartbeat (depolarization, plateau and repolarization). The cardiac domain Ω considered in this test is a slab of dimensions $1.28 \times 1.28 \times 0.16\,\text{cm}^{-3}$, discretized by a fine mesh $\mathcal{T}_1 = 128 \times 128 \times 16$

Table 1 Test 1: ellipsoid. Scaled speedup test with coarse mesh $\mathscr{T}_0 = 2\mathscr{T}_S$, three time steps with $\tau = 0.05$ ms. nit = average nonlinear Newton iterations per time step; lit = average linear PCG iterations per nonlinear iteration; ltime = average CPU times per nonlinear iteration in seconds

Procs.	Procs. mesh $\mathscr{T}_S$	Fine mesh $\mathscr{T}_1$	d.o.f.	*nit*	*lit*	*ltime*
8	$1 \times 2 \times 4$	$32 \times 64 \times 128$	0.5M	2	21.8	6.9
16	$1 \times 4 \times 4$	$32 \times 128 \times 128$	1.1M	2	21.5	6.1
32	$1 \times 8 \times 4$	$32 \times 256 \times 128$	2.2M	2	21.5	7.5
64	$1 \times 16 \times 4$	$32 \times 512 \times 128$	4.4M	2	21.3	7.9
128	$2 \times 16 \times 4$	$64 \times 512 \times 128$	8.6M	2	22.0	9.2
256	$4 \times 16 \times 4$	$128 \times 512 \times 128$	17.1M	2	22.3	11.5
512	$8 \times 16 \times 4$	$256 \times 512 \times 128$	34.1M	2	22.8	12.9
1,024	$16 \times 16 \times 4$	$512 \times 512 \times 128$	67.9M	2	23.5	15.8
2,048	$32 \times 16 \times 4$	$1{,}024 \times 512 \times 128$	135.7M	2	23.7	23.7

Table 2 Test 2: complete cardiac cycle. Tnit = total nonlinear Newton iterations; nit = average nonlinear Newton iterations per time step; Tlit = total linear PCG iterations; lit = average linear PCG iterations per nonlinear iteration; Ttime = total CPU time in seconds; dtime = average CPU times per time step in seconds

Procs	$\mathscr{T}_0 = \mathscr{T}_S$	*Tnit*	*nit*	*Tlit*	*lit*	*Ttime*	*dtime*	Speedup	
4	$2 \times 2 \times 1$	13,189	1.6	483,067	36.6	334,890	41.9	1.0	(1)
8	$4 \times 2 \times 1$	13,153	1.6	520,377	39.5	209,030	26.1	1.6	(2)
16	$4 \times 4 \times 1$	13,158	1.6	426,450	32.4	85,614	10.7	3.9	(4)
32	$8 \times 4 \times 1$	9,081	1.1	328,589	36.2	40,664	5.1	8.2	(8)
64	$8 \times 8 \times 1$	13,190	1.6	249,551	18.9	18,886	2.4	17.5	(16)

($h = 0.01$ cm). For this problem, we perform a standard speedup test (strong scaling) by keeping the fine mesh $\mathscr{T}_1$ fixed while increasing the number of subdomains (= number of processors) from 4 to 64, hence increasing the coarse mesh $\mathscr{T}_0 = \mathscr{T}_S$ from $2 \times 2 \times 1$ to $8 \times 8 \times 1$. The small number of processors in this test is due to the long simulation times needed by a complete cardiac cycle (reported in the column *Ttime* of Table 2), since we cannot afford to run for such long times on a larger number of processors. The overlap size is fixed to $\delta = h$.

Table 2 reports in each row the total number of nonlinear Newton iterations (*Tnit*) over the 8,000 time steps performed, the average number of nonlinear Newton iterations per time step (*nit*), the total number of PCG iteration (*Tlit*), the average number of PCG iteration per Newton iteration (*lit*), the total CPU time (*Ttime*), the average CPU time per time step (*dtime*) and the speedup of the total *Ttime* (or the average *dtime* since they only differ by the constant factor 8,000) with respect to the 4 processor run defined as

$$\text{speedup (procs)} := \frac{\text{Ttime (procs)}}{\text{Ttime(4)}}$$

(and equivalently for the average *dtime*). The results in Table 2 show that, as in the previous test, the Newton iterations (both total and average) are independent of the number of subdomains, except an unexpected reduction in the run with 32

processors. The total linear iterations *Tlit* show a consistent reduction for increasing processors (except for the first increase from 4 to 8 processors), in agreement with the main bound of Theorem 1, since H (and also the ratio H/h because h is fixed) is reduced when the number of subdomains (processors) is increased.

References

1. Balay, S., Buschelman, K., Gropp, W.D., Kaushik, D., Knepley, M., Curfman McInnes, L., Smith, B.F., Zhang, H.: PETSc Users Manual. Tech. Rep. ANL-95/11 - Revision 2.1.5, Argonne National Laboratory (2002)
2. Cai, X.-C., Keyes, D.: Nonlinearly preconditioned inexact Newton algorithms. SIAM J. Sci. Comput. 24 (1), 183–200 (2002)
3. Colli Franzone, P., Pavarino, L.F.: A parallel solver for reaction-diffusion systems in computational electrocardiology. Math. Mod. Meth. Appl. Sci. 14 (6), 883–911 (2004)
4. Colli Franzone, P., Pavarino, L.F., Savaré G.: Computational electrocardiology: mathematical and numerical modeling. In: Quarteroni, A., et al. (eds.) Complex Systems in Biomedicine, pp. 187–241. Springer, Berlin (2006)
5. Hwang, F.-N., Cai, X.-C.: A class of parallel two-level nonlinear Schwarz preconditioned inexact Newton algorithms. Comput. Meth. Appl. Mech. Eng. 196 (8), 1603–1611 (2007)
6. Luo, C., Rudy, Y.: A model of the ventricular cardiac action potential: depolarization, repolarization, and their interaction. Circ. Res. 68 (6), 1501–1526 (1991)
7. Mardal, K.-A., Nielsen, B.F., Cai, X., Tveito, A.: An order optimal solver for the discretized bidomain equations. Numer. Lin. Algebra Appl. 14 (2), 83–98 (2007)
8. Munteanu, M., Pavarino, L.F.: Decoupled Schwarz algorithms for implicit discretization of nonlinear Monodomain and Bidomain systems. Math. Mod. Meth. Appl. Sci. 19 (7), 1065–1097 (2009)
9. Munteanu, M., Pavarino, L.F., Scacchi, S.: A scalable Newton–Krylov–Schwarz method for the Bidomain reaction-diffusion system. SIAM J. Sci. Comput. 31 (5), 3861–3883 (2009)
10. Murillo, M., Cai, X.-C.: A fully implicit parallel algorithm for simulating the non-linear electrical activity of the heart. Numer. Lin. Algebra Appl. 11 (2–3), 261–277 (2004)
11. Pavarino, L.F., Scacchi S.: Multilevel additive Schwarz preconditioners for the Bidomain reaction–diffusion system. SIAM J. Sci. Comput. 31 (1), 420–443 (2008)
12. Pennacchio, M., Savaré, G., Colli Franzone, P.: Multiscale modeling for the bioelectric activity of the heart. SIAM J. Math. Anal. 37 (4), 1333–1370 (2006)
13. Plank, G., Liebmann, M., Weber dos Santos, R., Vigmond, E.J., Haase, G.: Algebraic multigrid preconditioner for the cardiac bidomain model. IEEE Trans. Biomed. Eng. 54 (4), 585–596 (2007)
14. Potse, M., Dubè, B., Richer, J., Vinet, A., Gulrajani, R.: A comparison of Monodomain and Bidomain reaction–diffusion models for action potential propagation in the human heart. IEEE Trans. Biomed. Eng. 53 (12), 2425–2434 (2006)
15. Scacchi, S.: A hybrid multilevel Schwarz method for the bidomain model. Comput. Meth. Appl. Mech. Eng. 197 (45–48), 4051–4061 (2008)
16. Smith, B.F., Bjørstad, P., Gropp, W.D.: Domain Decomposition: Parallel Multilevel Methods for Elliptic Partial Differential Equations. Cambridge University Press, Cambridge (1996)
17. Sundnes, J., Lines, G.T., Tveito, A.: An operator splitting method for solving the bidomain equations coupled to a volume conductor model for the torso. Math. Biosci. 194 (2), 233–248 (2005)
18. Toselli, A., Widlund, O.B.: Domain Decomposition Methods: Algorithms and Theory. Computational Mathematics, Vol. 34. Springer, Berlin (2004)
19. Veneroni, M.: Reaction–Diffusion systems for the macroscopic Bidomain model of the cardiac electric field. Nonlinear Anal. R. World Appl. 10 (2), 849–868 (2009)

20. Vigmond, E.J., Aguel, F., Trayanova, N.A.: Computational techniques for solving the bidomain equations in three dimensions. IEEE Trans. Biomed. Eng. 49, 1260–1269 (2002)
21. Vigmond, E.J., Weber dos Santos, R., Prassl, A.J., Deo, M., Plank, G.: Solvers for the cardiac bidomain equations. Progr. Biophys. Molec. Biol. 96, 3–18 (2008)
22. Whiteley, J.: An efficient numerical technique for the solution of the monodomain and bidomain equations. IEEE Trans. Biomed. Eng. 53 (11), 2139–2147 (2006)

On a Shallow Water Model for Non-Newtonian Fluids

G. Narbona-Reina and D. Bresch

Abstract The aim of this work is to modelize the evolution of a viscoelastic fluid through a Shallow Water system.

The fluid hydrodynamic in this situation comes from the Navier-Stokes equations but the difficulty lies in the definition of the stress tensor for this non-Newtonian fluid. In order to get an expression for it we focus on the microscopic properties of the fluid by considering a diluted solution of polymer liquids. A kinetic theory for this type of solutions gives us "constitutive equations" that relate the stress tensor to the velocity. They are known as the Fokker–Planck equations.

Once the stress tensor is defined we shall derive the model by developing the asymptotic analysis of the joined system of equations to obtain a Shallow Water type model following [6]. Finally we show a numerical test to check the influence of the polymers in the behavior of the flow.

1 The Fokker–Planck Equation

We consider a diluted solution of polymers to modelize a non-Newtonian fluid. There are many structures for the polymer to be (cf. [1]); in this work we consider the simplest model able to account for noninteracting polymer chains, the so called *elastic dumbbell model*. So a polymer is represented as two beads connected by an entropic spring, see Fig. 1.

This configuration is characterized at time t by the position of the center of the mass $r(t)$ and its elongation $q(t)$ that follow a system of stochastic differential

G. Narbona-Reina (✉)
Dpto. de Matemática Aplicada I, E.T.S. Arquitectura. Universidad de Sevilla. Avda. Reina Mercedes 2. 41012 Sevilla, Spain
e-mail: gnarbona@us.es

D. Bresch
LAMA, UMR5127 CNRS, Université de Savoie, 73376 Le Bourget du Lac, France
e-mail: Didier.Bresch@univ-savoie.fr

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_74, © Springer-Verlag Berlin Heidelberg 2010

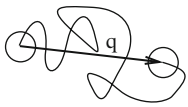


Fig. 1 Elastic dumbbell polymer

equations (see [1] for details). From this system the kinetic theory of polymer liquids gives us the diffusion equation (called Fokker–Planck eq.) for the function $f(t,x,q)$ that is defined as the probability density of a polymer to be at the time t at the point x in the configuration q (that represents the vector connecting the two beads). If we assume that the fluid moves at velocity u, this equation reads as:

$$\partial_t f + u\cdot\nabla f = -\mathbf{div}_q\left(HF + IF + BF\right), \text{ for } (t,x,q)\in\mathbb{R}^+\times\Omega\times B, \qquad (1)$$

with Ω the fluid domain and B the range for the elongation q.

This equation collects all the force effects experienced by the polymer that we enumerate next:

1. The hydrodynamic drag force: $HF = (\nabla u\cdot q)f$. This is the force of resistance of the polymer as it moves through the fluid.
2. The intramolecular force: $IF = -\frac{2}{\zeta}F(q)f$. This is the force resulting from the spring in the dumbbell, being $F(q)$ the spring force.
3. The Brownian force: $BF = -\frac{2\kappa\theta}{\zeta}\nabla_q f$. This force includes the effects of the thermal fluctuation of the fluid on the polymer.

The constants involved in the equation are ζ the friction coefficient between the beads, θ the temperature and κ the Boltzman constant. The definition of the spring force $F(q)$ depends on the structure of the polymers. In particular for the kind of polymer that we are considering we choose the following expression:

$$F(q) = \frac{Hq}{1-\frac{q^2}{q_0^2}}, \quad q\le q_0. \qquad (2)$$

In this case q_0 is the maximum elongation of the spring in the polymer. It is called the *Finitely Extensible Nonlinear Elastic (FENE)* connector force [1,3].

1.1 The Solution of the Fokker–Planck Equation

In order to get the stress tensor we must solve the Fokker–Planck equation. For this aim we firstly must write this equation under a dimensionless form.

Adimensional Fokker-Planck equation

First note that since q_0 is the maximum dumbbell extension, then we take $B = B(0,q_0)$. We define characteristic variables for the length (L_*), the velocity (U_*) and for q (Q_*), so the equation (1) reads:

$$\partial_t f + u \cdot \nabla f = -\mathbf{div}_q \left(\nabla u \cdot q\, f - \frac{1}{2\mathscr{D}e} F(q) f - \frac{1}{2\mathscr{D}e} \frac{1}{b} \nabla_q f \right), \tag{3}$$

where now $(t, x, q) \in \mathbb{R}^+ \times \Omega \times B(0, \sqrt{\delta})$ and being δ, $\mathscr{D}e$ and b adimensional parameters given by $\mathscr{D}e = \frac{\xi U_*}{4L_* H}$, $\delta = \frac{q_0^2}{Q_*^2}$ and $b = \frac{H q_0^2}{k\theta}$.

The number $\mathscr{D}e$ is called the *Deborah* number and it is an indicator of how fluid a material is. Thus, for $\mathscr{D}e << 1$ we find a fluid behavior while for $\mathscr{D}e >> 1$ the material acts as a solid. So we shall consider $\mathscr{D}e << 1$ to have a purely viscous fluid behavior.

According to [7], the parameter δ is roughly the number of monomer units represented by a bead; thus it is generally larger than 10. To simplify the Fokker–Planck equation we usually assume that $b = 1$ when $Q_* \sim 1$ cm, (see [3] for example). With these variables the spring force in equation (3) is now done by: $F(q) = \frac{q}{1 - \frac{q^2}{\delta}}$.

The Chapman-Enskog procedure

One way to find a solution for the Fokker–Planck equation is to use the Chapman-Enskog procedure (see [1,4] for details) that allows to find successive approximate solutions to general kinetic equations of the form:

$$Kf = \frac{1}{\epsilon} Q\, f \tag{4}$$

in terms of the small parameter ϵ being K and Q two operators.

To find a solution of the Fokker–Planck equation we assume $\mathscr{D}e \sim \epsilon$ and we define the next operators:

- T, the transport operator: $Tf = \partial_t f + u \cdot \nabla f$,
- B, the operator corresponding to the drag force on the beads: $Bf = \nabla_q \cdot (\nabla u \cdot q\, f)$,
- A, the operator due to the motion by the springs: $Af = \frac{1}{2} \nabla_q \cdot (\nabla_q f + F(q)\, f)$.

Then the Fokker–Planck equation (3) can be written as follows:

$$Tf + Bf = \frac{1}{\epsilon} Af. \tag{5}$$

So taking $K = T + B$ and $Q = A$ we can write the Fokker–Planck equation as in (4).

Due to the good properties of the operator A, we can write it as $Af = \frac{1}{2} \nabla_q \cdot \left(M \nabla_q \left(\frac{f}{M} \right) \right)$, with M being the normalized Maxwellian defined by

$$M(q) = \frac{1}{J} \left(1 - \frac{q^2}{\delta} \right)^{\frac{\delta}{2}} \quad \text{with } J = \int_{B(0,\sqrt{\delta})} \left(1 - \frac{q^2}{\delta} \right)^{\frac{\delta}{2}} dq. \tag{6}$$

That helps us to find an unique solution f of (5) given by (cf. [4]):

$$f(q) = n_0 M(q) \quad \text{being } n_0 \text{ a constant.}$$

2 Deduction of the Model

We consider a viscous fluid in a periodic domain $\Omega = \Omega(t) = \{(x,z) \in \mathbb{R}^2 / 0 \leq x \leq L,\ 0 \leq z \leq h\}$ modeled by the Navier-Stokes equations:

$$\eta \partial_t u + \eta \mathbf{div}(u \otimes u) + \nabla p = \mathbf{div}(\sigma) - \eta g e_z \quad \text{and } \mathbf{div}(u) = 0 \quad \text{in } \mathbb{R}^+ \times \Omega \quad (7)$$

being η the density, $u = (v, w)$ the velocity, p the pressure, g the gravity constant and σ the stress tensor.

We consider the Fokker–Planck equation to modelize the movement of polymers immersed into the newtonian fluid. So, essentially, we have a new definition for the stress tensor divided in two parts: $\sigma = \sigma_S + \sigma_P$, where σ_S comes from the newtonian fluid, that is usually defined as $\sigma_S = \mu(\nabla u + \nabla^t u)$, being μ the viscosity and σ_P directly related to the forces acting on the polymer. It is defined from the distribution density function f –solution of (1)– through the following expression (cf. [1]):

$$\sigma_P(t,x) = \langle F(q) \otimes q \rangle_d - \kappa\theta \langle Id \rangle_d , \quad (8)$$

where the q-average $< \cdot >_d$, is defined as $\langle \phi \rangle_d = \int_B \phi(q) f(q) dq$.

Therefore, to write the whole model we must consider both Navier-Stokes and Fokker–Planck equations together with the boundary conditions. In particular we consider the free surface condition and the effect of the atmospheric pressure. On the flat bottom we consider the no penetration condition and the friction effect. We write the problem as follows:

$$\begin{cases} \eta \partial_t u + \eta \mathbf{div}(u \otimes u) + \nabla p = \mathbf{div}(\sigma) - \eta g e_z; \\ \mathbf{div}(u) = 0; \\ \partial_t f + u \cdot \nabla f = -\mathbf{div}_q \left(\nabla u \cdot q\, f - \frac{2}{\xi} F(q) f - \frac{2\kappa\theta}{\xi} \nabla_q f \right); \\ \left.\begin{array}{l} (\sigma - p) \cdot n_S = \alpha_S\, k \cdot n_S \\ \partial_t h + v \cdot \partial_x h = w \end{array}\right\} \quad \text{on } z = h \\ \left.\begin{array}{l} ((\sigma - p) \cdot n_B)_\tau = \alpha_B u_\tau \\ u \cdot n_B = 0 \end{array}\right\} \quad \text{on } z = 0 \end{cases} \quad (9)$$

Where we consider the following notation: h the free surface, α_S the tension coefficient at the surface, α_B the friction coefficient at the bottom, n_S the normal vector to the surface, n_B the normal vector to the bottom and $k = \mathbf{div}(n_S)$ the mean curvature on the surface.

For the derivation of the model we follow the work developed by Gerbeau and Perthame in [6]. It consists of several steps: adimensionalization, hydrostatic approximations, vertical integration and asymptotic analysis. The originality of our work lies in the addition of the Fokker–Planck equation to the classical problem, so for the sake of brevity we are only going to show the contribution of this equation into the process. Finally we'll show the model obtained. The complete deduction can be found in [2].

With regard to the first step, the dimensionless Navier-Stokes equations are got as usual under the Shallow-Water hypothesis, i.e., the characteristic length of the domain (L_*) is larger than the characteristic height (H_*). Thus we can assume $\epsilon = \frac{H_*}{L_*}$ to be small and we obtain the equations in function of this parameter. Besides, for the asymptotic regime we chose: $\mu = \epsilon\mu_0$, $\alpha_S = \epsilon\alpha_{0S}$, $\alpha_B = \epsilon\alpha_{0B}$, $\mathscr{D}e = \epsilon\mathscr{D}e_0$.

Note that in this case we also assume that the characteristic vertical velocity is ϵU_*, that must be taken into account on the dimensionless Fokker–Planck equation where ∇u appears. In this sense we have to correct the equation (3). First we denote:

$$\nabla u = \frac{1}{\epsilon}\underbrace{\begin{pmatrix} 0 & \partial_z v \\ 0 & 0 \end{pmatrix}}_{C} + \underbrace{\begin{pmatrix} \partial_x v & 0 \\ 0 & \partial_z w \end{pmatrix}}_{G} + \epsilon \underbrace{\begin{pmatrix} 0 & 0 \\ \partial_x w & 0 \end{pmatrix}}_{E}, \tag{10}$$

so we write the operator B as follows:

$$Bf = \nabla_q \cdot \left(\frac{1}{\epsilon} Cqf + Gqf + \epsilon Eqf \right) = \frac{1}{\epsilon}\nabla_q \cdot (Cqf) + \nabla_q \cdot (Gqf) + \epsilon \nabla_q \cdot (Eqf).$$

The Fokker–Planck equation reads now as:

$$Tf + \nabla_q \cdot (Gqf) + \epsilon \nabla_q \cdot (Eqf) = \frac{1}{\epsilon \mathscr{D}e_0}(Af - \mathscr{D}e_0 \nabla_q \cdot (Cqf))$$

that can be written as a revised equation (5):

$$Tf + \widetilde{B} f = \frac{1}{\epsilon}\widetilde{A} f, \tag{11}$$

being $\widetilde{B} f = \nabla_q \cdot ((G + \epsilon E)qf)$ and $\widetilde{A} f = \frac{1}{2\mathscr{D}e_0}\nabla_q \cdot (\nabla_q f + F(q) f - 2\mathscr{D}e_0 Cqf)$.

We must also write the dimensionless expression for the stress tensor σ_P:

$$\sigma_P(t, x) = \lambda \left(b \left\langle F(q) \otimes q \right\rangle - n_0 Id \right), \tag{12}$$

being now the new $q-$average $\langle \phi \rangle = \displaystyle\int_{B(0,\sqrt{\delta})} \phi(q) f(q) dq$ and $n_0 = \displaystyle\int_{B(0,\sqrt{\delta})} f(q) dq$ the density of the polymer chains. The parameter $\lambda = \kappa\theta Q_*$ can be roughly interpreted as the mean kinetic energy of the Brownian motion.

Once we get the dimensionless coupled system we neglect the second order terms to obtain the hydrostatic system and we work out the vertical integration.

To obtain the Shallow Water system we first take the development of the variables in function of ϵ up to second order:

$$v = v_0 + \epsilon v_1 + \mathcal{O}(\epsilon^2),\ w = w_0 + \epsilon w_1 + \mathcal{O}(\epsilon^2),\ p = p_0 + \epsilon p_1 + \mathcal{O}(\epsilon^2),$$
$$h = h_0 + \epsilon h_1 + \mathcal{O}(\epsilon^2),\ \sigma_P = \sigma_{P0} + \epsilon\sigma_{P1} + \mathcal{O}(\epsilon^2)$$

and $k = \partial_x^2 h + \mathcal{O}(\epsilon^2)$. If we write the equations up to principal order, $\frac{1}{\epsilon}$, we have:

$$\mu_0\partial_z^2 v_0 = -\partial_z\sigma_{P0}^{12},\quad \mu_0\partial_z v_{0|z=h} = -\sigma_{P0|z=h}^{12},\quad \mu_0\partial_z v_{0|z=0} = -\sigma_{P0|z=0}^{12}.$$

The classical way to deduce the Shallow Water system is based on the fact that the velocity v does not depend on z up to first order, so $v_0 = v_0(t,x)$ and $\partial_z v_0 = 0$.

Accordingly to this and to continue with the deduction of the model we are obliged to prove that $\sigma_{P0}^{12} = 0$. As we know σ_P comes from the solution of the Fokker–Planck equation according to (12), so we must solve it.

Theorem 1. *We consider the matrix* $C = \begin{pmatrix} 0 & \partial_z v \\ 0 & 0 \end{pmatrix}$ *with* $\partial_z^2 v = \frac{1}{\mu_0}\partial_z\sigma_P^{12}$ *and*

$$\sigma_P^{12} = \lambda\int_{\mathscr{D}} \frac{q_1 q_2}{1-\frac{q^2}{\delta}} f(q)dq$$

with $q = (q_1, q_2)$. *Then the equation* $\widetilde{A}f = 0$ *admits an unique radial solution* $f = f(|q|)$ *and it is of the form* $f = n_0 M(q)$, *being* n_0 *the density of the polymer chains, solution of* $\partial_t n_0 + u_0\cdot\nabla n_0 = 0$ *and* M *given by (6).*

Now we follow the Chapman-Enskog procedure to find the solution $f = f_0 + \epsilon f_1$ of the Fokker–Planck equation and we calculate $\sigma_P = \sigma_{P0} + \epsilon\sigma_{P1}$ to get:

$$\sigma_{P0} = \gamma_0 n_0 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},\ \text{for } \gamma_0 = \frac{\lambda(\pi\beta(3)-1)}{J} \tag{13}$$

and

$$\sigma_{P1} = \gamma_1 n_0 \begin{pmatrix} \partial_x v_0 & \partial_z v_1 \\ \partial_z v_1 & -\partial_x v_0 \end{pmatrix},\ \text{for } \gamma_1 = \frac{\lambda\beta(5)}{4J}, \tag{14}$$

being the function $\beta(p) = \int_0^{\sqrt{\delta}} r^p \left(1-\frac{r^2}{\delta}\right)^{\delta/2-1} dr$.

Note that γ_0 and γ_1 correspond to the contribution of the polymers present in the fluid into the friction effect.

Now that we know σ_P, we can continue with the deduction of the model in the classical way to get the Shallow Water system.

2.1 Final Systems

In this section we show the two models obtained by taking the first and the second order approximation respectively. We write them in the dimensional form so we denote $\tilde{\alpha}_B = \frac{\alpha_B}{\eta}$, $\tilde{\alpha}_S = \frac{\alpha_S}{\eta}$ and $\nu = \frac{\mu}{\eta}$.

First order approximation

$$\text{(S1)} \begin{cases} \partial_t h + \partial_x(hv) = 0; \\ \partial_t(hv) + \partial_x(hv^2) + \frac{1}{2}g\partial_x(h^2) = -\tilde{\alpha}_B v; \\ \partial_t(hn_0) + \partial_x(v\,hn_0) = 0. \end{cases} \tag{15}$$

As we can see in these equations neither the tension nor viscosity effects appear.

Second order approximation

$$\text{(S2)} \begin{cases} \partial_t h + \partial_x(hv) = 0; \\ \partial_t(hv) + \partial_x(hv^2) + \frac{1}{2}g\partial_x(h^2) - 4\nu\partial_x(h\partial_x v) = -\xi\tilde{\alpha}_B v + \tilde{\alpha}_S h\partial_x^3 h; \\ \partial_t(hn_0) + \partial_x(v\,hn_0) = 0. \end{cases} \tag{16}$$

Where now, $\xi = \left(1 + \frac{1}{3}\frac{\tilde{\alpha}_B h}{\nu\tau(n_0)}\right)^{-1}$ and $\tau(n_0) = 1 + \epsilon\frac{\gamma_1 n_0}{\nu}$.

As usually we obtain a corrected friction term for the second order approximation. If we look at the system (15), the friction term, $\tilde{\alpha}_B v$, depends only on the friction coefficient $\tilde{\alpha}_B$ while in the system above the friction terms reads as $\xi\tilde{\alpha}_B v$. This new coefficient ξ contains the polymer effects into the fluid, represented by γ_1.

3 Numerical Results

In this section we solve a dam break problem for the two models obtained previously in order to check the polymer effect into the fluid.

To solve numerically these systems we have used the WAF method of second order accuracy (see [5]). For the sake of simplicity we don't consider the surface tension effect by taking $\tilde{\alpha}_S = 0$.

We consider a domain of length $L = 50$ and take $\Delta x = 0.5$, the CFL condition is fixed as 0.9 and the final time is 3 seconds. We take the following initial conditions:

$$h(t=0) = \begin{cases} 3 & x < 10 \\ 0.1 & x \geq 10 \end{cases}; \; hv(t=0) = \begin{cases} 3.5 & x < 10 \\ 0 & x \geq 10 \end{cases}; \; n_0(t=0) = 1.$$

We have fixed the following constants: $\eta_0 = 0.95$, $\mu_0 = 0.5$, $\delta = 10$, with $\nu = \frac{\mu_0}{\eta_0}10^{-3}$. Considering that $\lambda \in [0.1, 5]$ (cf. [4]), we have taken the following values: $\tilde{\alpha}_B = 10^{-1}, 10^{-2}, 10^{-3}$; $\quad \lambda = 0.1, 1, 3, 5$.

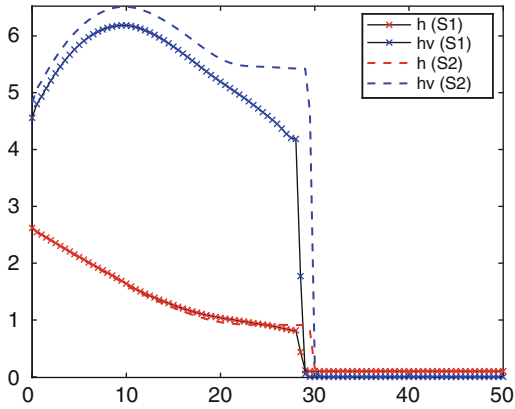


Fig. 2 Solutions for (S1) and (S2)

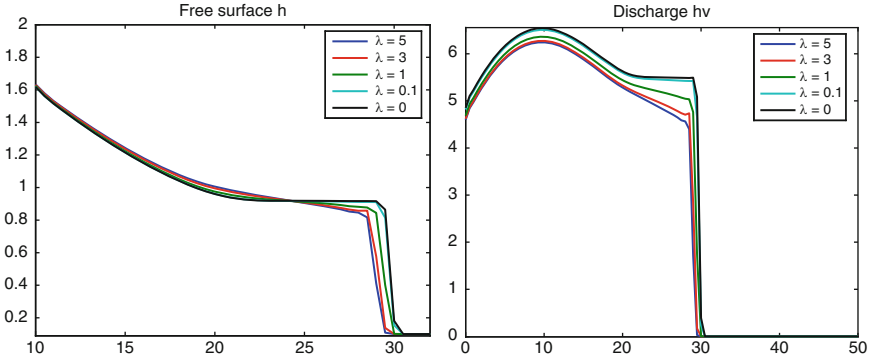


Fig. 3 Solutions for (S2) for all values of λ (zoom for h)

We show the numerical results in Fig. 2, where we can see that the main difference is found in the discharge. For these results we have taken $\tilde{\alpha}_B = 10^{-1}$ and $\lambda = 0.1$ that gives us the minimum value for ξ, for which we obtain the largest difference between the two models.

The innovation of this work is the presence of polymers into the fluid that comes from the coefficient γ_1. So another important issue is to check the influence of γ_1 in the model. For this aim we shall also take $\gamma_1 = 0$ to compare this solution of second order model with those when $\gamma_1 \neq 0$. These solutions are shown in Fig. 3 where we can see that this influence is not insignificant at all.

4 Conclusions

In summary, we have deduced a new Shallow-Water model for a non-newtonian fluid focused on its microscopic properties. Regarding the numerical results, as usual we find important differences between first and second order models due to the viscosity effect. But we evidence that the influence of the polymers in the evolution of the fluid is also important.

Acknowledgements The research of G. Narbona-Reina to develop this work was partially supported by the Spanish Government Research project MTM2006-01275.

References

1. R. B. Bird, C. F. Curtiss, R. C. Armstrong , O. Hassager, *Dynamics of Polymeric Liquids*, (Wiley, NY, 1987)
2. D. Bresch, G. Narbona-Reina, *A shallow water model for viscoelastic fluid from the kinetic theory of polymer solutions*, (in preparation)
3. L. Chupin, *The FENE viscoelastic model and thin film flows*, C. R. Acad. Sci. Paris, Ser. I, **347**(17–18), pp. 1041–1046 (2009)
4. P. Degond, M. Lemou, M. Picasso, *Viscoelastic fluid models derived from kinetic equations for polymers*, SIAM J. Appl. Math. **62**(5), pp. 1501–1019 (2002)
5. E.D. Fernàndez-Nieto, G. Narbona-Reina, *Extension of waf type methods to nonhomogeneous shallow water equations with pollutant*, J. Sci. Comput. **36**(2), pp. 193–217 (2008)
6. J.F. Gerbeau, B. Perthame, *Derivation of viscous Saint-Venant system for laminar shallow water; numerical validation*, Discrete Contin. Dyn. Syst. Ser. B **1**(1), pp. 89–102 (2001)
7. H.C. Ottinger, *Stochastic process in polymeric fluids*, (Springer, Berlin, 1996)

On Stationary Viscous Incompressible Flow Through a Cascade of Profiles with the Modified Boundary Condition on the Outflow and Large Inflow

Tomáš Neustupa

Abstract The paper is concerned with the analysis of the model of incompressible, viscous, stationary flow through a plane cascade of profiles. The problem is formulated in a bounded domain of the form of one space period with suitable boundary conditions on the boundary. Let us recall that there is usually imposed the condition on smallness of the inflow velocity or the condition on smallness of fluxes between various components of the boundary (Specially that the balance of fluid entering and leaving domain is zero for each component of boundary) in known theorems on existence of a weak solution of the boundary-value problem for the Navier–Stokes equation with the nonzero Dirichlet boundary condition, (see e.g., Mathematical Methods in Fluid Dynamics (1993), An Introduction to the Mathematical Theory of the Navier–Stokes Equations (1994), Finite Element Approximation of the Navier–Stokes Equations (1979), Navier–Stokes Equations (1977)). In this paper the case of a large inflow is considered, however the possibility of the large inflow is compensated by certain modification of the boundary condition on the outflow and by a specification on the shape of the domain.

1 Introduction and the Geometry of the Problem

We study the steady flow through a simplified plane cascade of profiles. The model of cascade of profiles describes e.g., the flow through a turbine. If we consider the intersection of the real 3D region filled by the moving fluid with a circular cylindrical surface, whose axis coincides with the axis of rotation of the turbine, and expand the surface in the x_1, x_2-plane. We can naturally arrive at a 2D domain. The obtained domain is unbounded, however periodic in the x_2–direction. Its complement in R^2 consists of the infinite number of profiles, numbered from $-\infty$ to $+\infty$.

T. Neustupa
Faculty of Mechanical Engineering, Czech Technical University Prague, Karlovo nám. 13, 121 35 Prague, Czech Republic
e-mail: tneu@centrum.cz

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_75,

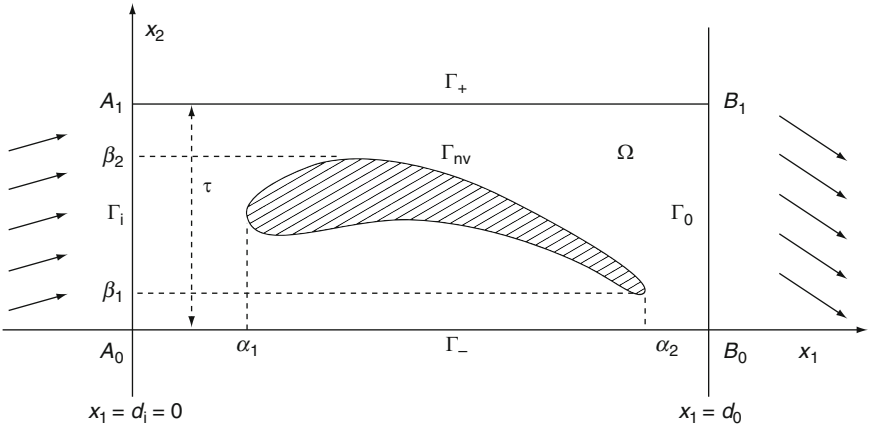


Fig. 1 Domain Ω

We suppose that the boundary of the profile No. 0 is a simple closed curve C_0 in R^2, piecewise of the class C^2, whose interior and exterior are domains with a Lipschitz-continuous boundary. We put $C_k = \{(x_1, x_2 + k\tau);\ (x_1, x_2) \in C_0\}$ (for $k \in Z$), where τ is a positive constant. We assume that τ is so large that the curves C_k are mutually disjoint. The set $M := \bigcup_{k=-\infty}^{+\infty} \overline{\mathrm{Int}\, C_k}$ is called a *cascade of profiles*. ($\mathrm{Int}\, C_k$ denotes the interior of curve C_k.) Number τ is called the *period* of the cascade.

It is reasonable to assume that the flow through the cascade is periodic in the x_2-direction with the period τ. Consequently, we can study the flow just in one spatial period of the whole domain. The chosen period is denoted Ω. Its boundary consists of the curves Γ_i, Γ_o, Γ_+, Γ_- and Γ_w. See Fig. 1.

We suppose that the profiles in the cascade have a shape which enables us to choose the artificial periodic boundaries as strait lines. Furthermore, we choose for simplicity the origin of the system of coordinates so that $A_0 = [0, 0]$. Then $d_i = 0$ and $A_1 = [0, \tau]$, $B_0 = [d_o, 0]$ and $B_1 = [d_o, \tau]$. The corresponding shape of domain Ω is now obvious from Fig. 1. For technical reasons we assume that the curve Γ_w is of the class C^2.

2 Auxiliary Functions and Results

2.1 The Auxiliary Cut-Off Function θ_ϵ

Suppose that $\epsilon > 0$ is a small positive number. We assume that

$$\mathrm{e}^{-1/\epsilon} < \max\{\tfrac{1}{6}\alpha_1;\ \tfrac{1}{3}[\tau - \beta_2];\ \tfrac{1}{3}\beta_1\}. \tag{1}$$

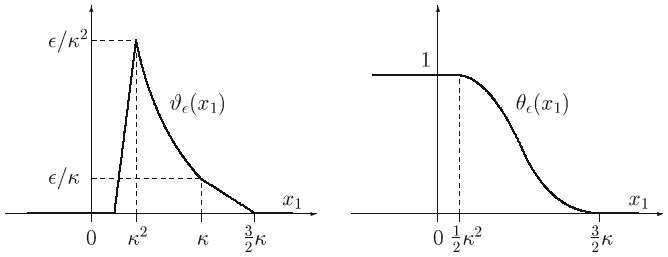
We set $\kappa = \mathrm{e}^{-1/\epsilon}$ (hence $\mathrm{e}^{-2/\epsilon} = \kappa^2$) and

$$\vartheta_\epsilon(x_1) := \begin{cases} 0 & \text{for} \quad x_1 \le \frac{1}{2}\kappa^2, \\ \text{linear} & \text{for} \quad \frac{1}{2}\kappa^2 \le x_1 \le \kappa^2, \\ \epsilon/x_1 & \text{for} \quad \kappa^2 \le x_1 \le \kappa, \\ \text{linear} & \text{for} \quad \kappa \le x_1 \le \frac{3}{2}\kappa, \\ 0 & \text{for} \quad \frac{3}{2}\kappa \le x_1. \end{cases}$$

Further, we define

$$\theta_\epsilon(x_1) = \frac{1}{K}\int_{x_1}^{\frac{3}{2}\kappa} \vartheta_\epsilon(t)\,\mathrm{d}t \qquad \text{where} \qquad K := \int_0^{\frac{3}{2}\kappa} \vartheta_\epsilon(t)\,\mathrm{d}t.$$

An elementary calculation shows that $K = \frac{1}{2}\epsilon + 1 \ge 1$.



We can observe that

$$\sup_{x_1>0} |\theta'_\epsilon(x_1)| = \frac{\epsilon}{K\kappa^2} \le \frac{\epsilon}{\kappa^2},$$

$$|\theta'_\epsilon(x_1)| \le \frac{3\epsilon}{2Kx_1} \le \frac{3\epsilon}{2x_1} \qquad \text{for } x_1 > 0, \quad x_1 \ne \tfrac{1}{2}\kappa^2,\ \kappa^2,\ \kappa,\ \tfrac{3}{2}\kappa.$$

and

$$\sup_{x_1>0} |\theta''_\epsilon(x_1)| = \frac{2\epsilon}{K\kappa^4} \le \frac{2\epsilon}{\kappa^4},$$

$$|\theta''_\epsilon(x_1)| \le \frac{9\epsilon}{2Kx_1^2} \le \frac{9\epsilon}{2x_1^2} \qquad \text{for } x_1 > 0, \quad x_1 \ne \tfrac{1}{2}\kappa^2,\ \kappa^2,\ \kappa,\ \tfrac{3}{2}\kappa.$$

2.2 The Auxiliary Cut-Off Function χ_ϵ

(By analogy with Temam [6]) we define $\rho(\mathbf{x}) := \mathrm{dist}(\mathbf{x}, \Gamma_w)$ and put

$$\chi_\epsilon(\mathbf{x}) := 1 - \theta_\epsilon\big(\rho(\mathbf{x})\big) \qquad \text{for } \mathbf{x} \in \Omega.$$

The function χ_ϵ equals zero in a neighborhood of the profile Γ_w (for $\rho(\mathbf{x}) < \frac{1}{2}\kappa^2$ and equals one far from Γ_w (for $\rho(\mathbf{x}) > \frac{3}{2}\kappa$).

As the function $\rho(\mathbf{x})$ is twice continuously differentiable in $\overline{\Omega}$, we can derive from estimates for $\theta'_\epsilon(x_1), \theta''_\epsilon(x_1)$ that there exist positive constants c_1, c_2, c_3 and c_4 (independent of ϵ) such that

$$\sup_{x\in\Omega} |\nabla\chi_\epsilon| \,=\, c_1\,\frac{\epsilon}{\kappa^2}, \qquad |\nabla\chi_\epsilon(\mathbf{x})| \,\le\, c_2\,\frac{\epsilon}{\rho(\mathbf{x})} \qquad \text{for a.a. } \mathbf{x}\in\Omega,$$

$$\sup_{x\in\Omega} |\nabla^2\chi_\epsilon(\mathbf{x})| \,\le\, c_3\,\frac{\epsilon}{\kappa^4}, \qquad |\nabla^2\chi_\epsilon(\mathbf{x})| \,\le\, c_4\,\frac{\epsilon}{\rho^2(\mathbf{x})} \qquad \text{for a.a. } \mathbf{x}\in\Omega.$$

(Here and in the following a.a. is an abbreviation for almost all and a.e. is an abbreviation for almost everywhere).

2.3 A Special Extension of the Inflow Profile g to the Domain Ω

We suppose that function $\mathbf{g}$ represents the given velocity profile on the inflow. Function $\mathbf{g}$ satisfies the condition $\mathbf{g}12(A_0) = \mathbf{g}(A_1)$. We assume that $\mathbf{g}^*$ is the extension of the function $\mathbf{g}$ from Γ_i onto Ω, constructed in [2], such that $\mathbf{g}^* = \mathbf{0}$ on Γ_w, $\mathbf{g}^*$ satisfies the condition of periodicity $\mathbf{g}^*(x_1, x_2+\tau) = \mathbf{g}^*(x_1, x_2)$ for $(x_1, x_2) \in \Gamma_-$ and the estimate $\|\mathbf{g}^*\|_1 \le c\,\|\mathbf{g}\|_{s;\,\Gamma_i}$ holds.

There exists a stream function $\psi^* \in H^2(\Omega)$ such that $\mathbf{g}^* = \left(\frac{\partial\psi^*}{\partial x_2}, -\frac{\partial\psi^*}{\partial x_1}\right)$ in Ω. (This can be deduced from Theorem 3.1 in [4, p. 37]. Since domain Ω is not simply connected, it is here important that the trace of $\mathbf{g}^*$ on Γ_w is zero.) Moreover, there exists a constant $c_5 > 0$ (independent of $\mathbf{g}^*$) such that

$$\|\psi^*\|_{H^2(\Omega)} \,\le\, c_5\,\|\mathbf{g}^*\|_{H^1(\Omega)^2}. \tag{2}$$

Now we modify the stream function ψ^* by means of the cut-off functions θ_ϵ and χ_ϵ and we define

$$\psi^{**}(x_1, x_2) := \psi^*(x_1, x_2)\,\theta_\epsilon(x_1) + \frac{\Phi}{\tau}\,x_2\left[1-\theta_\epsilon(x_1)\right]\chi_\epsilon(x_1, x_2), \tag{3}$$

$$\mathbf{g}^{**}(x_1, x_2) := \Big(\frac{\partial\psi^{**}}{\partial x_2}(x_1, x_2), -\frac{\partial\psi^{**}}{\partial x_1}(x_1, x_2)\Big) \tag{4}$$

where $\Phi := \int_0^\tau g_1(s)\,\mathrm{d}s$ (the flux into Ω through Γ_i). Thus, $\mathbf{g}^{**}$ is the divergence-free vector function whose stream function is ψ^{**} and

$$\|\mathbf{g}^{**}\|_{H^1(\Omega)^2} \,\le\, c(\epsilon)\,\|\mathbf{g}\|_{H^s(\Gamma_i)^2}. \tag{5}$$

(For complete proof see [5].)

The idea of the definition of the stream function ψ^{**} is as follows: We first use the cut-off function θ_ϵ in order to interpolate between the stream function ψ^* (generating the flow $\mathbf{g}^*$) and the stream function $(\Phi/\tau)\, x_2$ (generating the constant one-dimensional flow $(\Phi/\tau,\, 0)$). The interpolation in fact takes place in the area $\frac{1}{2}\kappa^2 < x_1 < \frac{3}{2}\kappa$. Then we multiply the stream function of the constant flow $(\Phi/\tau,\, 0)$ in the area $x_1 > \frac{3}{2}\kappa$ by the cut-off function $\chi_\epsilon(\mathbf{x})$ in order to modify the flow in the neighborhood of the profile Γ_w.

3 The Problem with the Large Inflow

We assume that the moving fluid is viscous and incompressible. From the definition of our mathematical model follows that the velocity can be considered to create a 2D vector field $\mathbf{u} = (u_1, u_2)$. We further denote by p the kinematic pressure, by $\mathbf{n}$ the outer normal to the boundary, $\mathbf{f}\,(= (f_1,\, f_2))$ the specific volume force and by ν is the kinematic coefficient of viscosity. We study the flow described by 2D steady Navier–Stokes equation in the form

$$(\mathbf{u}\cdot\nabla)\mathbf{u} \;=\; \mathbf{f} \,-\, \nabla p \,+\, \nu\,\Delta\mathbf{u}. \tag{6}$$

Equation (6) must be necessarily completed by the condition of incompressibility

$$\operatorname{div}\mathbf{u} \;=\; 0. \tag{7}$$

It is natural to prescribe the inhomogeneous Dirichlet boundary condition on the inlet:

$$\mathbf{u}\,|_{\Gamma_i} \;=\; \mathbf{g}. \tag{8}$$

We assume that the fluid satisfies the no slip Dirichlet boundary condition on the profile:

$$\mathbf{u}\,|_{\Gamma_w} \;=\; \mathbf{0}. \tag{9}$$

We further suppose that the following conditions of periodicity are fulfilled on the artificial boundaries Γ_+ and Γ_-:

$$\mathbf{u}(x_1, x_2 + \tau) = \mathbf{u}(x_1, x_2) \qquad \text{for } (x_1, x_2) \in \Gamma_-, \tag{10}$$

$$\frac{\partial\mathbf{u}}{\partial\mathbf{n}}(x_1, x_2 + \tau) = -\frac{\partial\mathbf{u}}{\partial\mathbf{n}}(x_1, x_2) \qquad \text{for } (x_1, x_2) \in \Gamma_-, \tag{11}$$

$$p(x_1, x_2 + \tau) \;=\; p(x_1, x_2) \qquad \text{for } (x_1, x_2) \in \Gamma_-. \tag{12}$$

The boundary condition on the outflow Γ_o, arises from the weak formulation of the problem and has a form

$$-\nu \frac{\partial \mathbf{u}}{\partial \mathbf{n}} + p\,\mathbf{n} - \frac{1}{2}(\mathbf{u}\cdot\mathbf{n})^{-}(\mathbf{u}-\mathbf{g}^{**}) = \mathbf{h} \tag{13}$$

which can, due to the special form of $\mathbf{g}^{**}$ on the outflow (following from (3)), be written in the form

$$-\nu \frac{\partial \mathbf{u}}{\partial \mathbf{n}} + p\,\mathbf{n} - \frac{1}{2}(\mathbf{u}\cdot\mathbf{n})^{-}\mathbf{u} + \frac{1}{2}(\mathbf{u}\cdot\mathbf{n})^{-}\big(\Phi/\tau,\,0\big) = \mathbf{h}.$$

Here $\mathbf{h} = (h_1, h_2)$ is a given function on Γ_o.

Equations (6) and (7) and the boundary conditions (8)–(13) represent the classical formulation of the considered boundary-value problem.

4 Weak Formulation of the Problem in Domain Ω and Existence of a Weak Solution

We denote by $H^1(\Omega)$ the usual Sobolev space of functions defined a.e. in Ω. The space of vector-functions (with values in R^2) whose each component belongs to $H^1(\Omega)$ is denoted by $H^1(\Omega)^2$. Furthermore, V denotes the space of vector-functions $\mathbf{v} = (v_1, v_2) \in H^1(\Omega)^2$ such that $\operatorname{div}\mathbf{v} = 0$ a.e. in Ω, $\mathbf{v} = \mathbf{0}$ a.e. in $\Gamma_i \cup \Gamma_w$ and $\mathbf{v}(x_1, x_2 + \tau) = \mathbf{v}(x_1, x_2)$ for a.a. $(x_1, x_2) \in \Gamma_-$. (The conditions on the curves Γ_i, Γ_w and Γ_- are interpreted in the sense of traces.) We equip the linear space V by the norm

$$|||\mathbf{v}||| := \left(\int_\Omega \sum_{i,j=1}^{2} \left(\frac{\partial v_i}{\partial x_j}\right)^2 \mathrm{d}\mathbf{x}\right)^{1/2}$$

which is equivalent with the norm of the space $H^1(\Omega)^2$.

Let us multiply (6) by an arbitrary test function $\mathbf{v} = (v_1, v_2) \in V$, integrate over Ω and use Green's theorem. We get

$$\int_\Omega \mathbf{f}\cdot\mathbf{v}\,\mathrm{d}x = \int_\Omega \left(\frac{\partial \mathbf{u}}{\partial t} - \nu\Delta\mathbf{u} + (\mathbf{u}\cdot\nabla)\,\mathbf{u} + \nabla p\right)\cdot\mathbf{v}\,\mathrm{d}x = \tag{14}$$

$$= \int_\Omega \frac{\partial \mathbf{u}}{\partial t}\cdot\mathbf{v}\,\mathrm{d}x + \nu\int_\Omega \sum_{i,j=1}^{2} \frac{\partial u_i}{\partial x_j}\frac{\partial v_i}{\partial x_j}\,\mathrm{d}x - \nu\int_{\partial\Omega} \frac{\partial \mathbf{u}}{\partial \mathbf{n}}\cdot\mathbf{v}\,\mathrm{d}S$$

$$+ \int_\Omega \sum_{i,j=1}^{2} u_j \frac{\partial u_i}{\partial x_j} v_i\,\mathrm{d}x - \int_\Omega p\,\operatorname{div}\mathbf{v}\,\mathrm{d}x + \int_{\partial\Omega} p\,\mathbf{v}\cdot\mathbf{n}\,\mathrm{d}S.$$

Using the properties of the function $\mathbf{v}$, the boundary conditions (10)–(13) and the relation $\mathbf{n}(x_1, x_2) = -\mathbf{n}(x_1, x_2 + \tau)$for $(x_1, x_2) \in \Gamma_-$, we obtain the identity

$$\int_\Omega \mathbf{f}\cdot\mathbf{v}\,\mathrm{d}x = \nu\int_\Omega \sum_{i,j=1}^{2}\frac{\partial u_i}{\partial x_j}\frac{\partial v_i}{\partial x_j}\,\mathrm{d}x + \tag{15}$$

$$+\int_\Omega \sum_{i,j=1}^{2} u_j\frac{\partial u_i}{\partial x_j}v_i\,\mathrm{d}x + \int_{\Gamma_o}\frac{1}{2}(\mathbf{u}\cdot\mathbf{n})^-(\mathbf{u}-\mathbf{g}^{**})\cdot\mathbf{v}\,\mathrm{d}S + \int_{\Gamma_o}\mathbf{h}\cdot\mathbf{v}\,\mathrm{d}S.$$

For $\mathbf{u} = (u_1, u_2)$, $\mathbf{v} = (v_1, v_2)$, $\mathbf{w} = (w_1, w_2) \in H^1(\Omega)^2$ we introduce the following forms:

$$(\mathbf{v},\mathbf{w}) = \int_\Omega \mathbf{v}\cdot\mathbf{w}\,\mathrm{d}x, \qquad a_1(\mathbf{u},\mathbf{v}) = \nu\int_\Omega \sum_{i,j=1}^{2}\frac{\partial u_i}{\partial x_j}\frac{\partial v_i}{\partial x_j}\,\mathrm{d}x,$$

$$a_2(\mathbf{u},\mathbf{v},\mathbf{w}) = \int_\Omega \sum_{i,j=1}^{2} u_j\frac{\partial v_i}{\partial x_j}w_i\,\mathrm{d}x, \qquad b(\mathbf{h},\mathbf{v}) = -\int_{\Gamma_o}\mathbf{h}\cdot\mathbf{v}\,\mathrm{d}S,$$

$$a_3(\mathbf{u},\mathbf{v},\mathbf{w}) = \int_{\Gamma_o}\frac{1}{2}(\mathbf{u}\cdot\mathbf{n})^-(\mathbf{v}-\mathbf{g}^{**})\cdot\mathbf{w}\,\mathrm{d}S,$$

$$a(\mathbf{u},\mathbf{v}) = a_1(\mathbf{u},\mathbf{v}) + a_2(\mathbf{u},\mathbf{u},\mathbf{v}) + a_3(\mathbf{u},\mathbf{u},\mathbf{v}). \tag{16}$$

Using this notation we arrive at the integral equation:

$$a(\mathbf{u},\mathbf{v}) = (\mathbf{f},\mathbf{v}) + b(\mathbf{h},\mathbf{v}). \tag{17}$$

Here you can see that the boundary condition on the outflow (Γ_o) is constructed according to the deriving of the weak solution. The special form of $\mathbf{g}^{**}$ enables us to prove of the coercivity of the form a without any restrictions on the incoming flow (restrictions on the smallness of the function $\mathbf{g}$). However we must use this more complicated form of the boundary condition on the outflow.

Definition 1. Let function $\mathbf{g} \in H^s(\Gamma_i)^2$ (for some $s \in (\frac{1}{2}, 1]$) satisfy the condition $\mathbf{g}(A_1) = \mathbf{g}(A_0)$ (where A_0 and A_1 are the end points of Γ_i). Let $\mathbf{f} \in L^2(\Omega)^2$ and $\mathbf{h} \in L^2(\Gamma_o)^2$. We seek a vector function $\mathbf{u} \in H^1(\Omega)^2$ which satisfies the equation of continuity (7) a.e. in Ω, the boundary conditions (8) (respectively (9)) in the sense of traces on Γ_i (respectively on Γ_w), the condition of periodicity (10) a.e. on Γ_- and such that identity (17) holds for all test functions $\mathbf{v} \in V$. The solution of this problem is called a *weak solution in the domain* Ω.

Now we shall seek for the weak solution $\mathbf{u}$ in the form $\mathbf{u} = \mathbf{g}^{**} + \mathbf{z}$ where $\mathbf{z} \in V$ is a new unknown function. This guarantees that $\mathbf{u}$ satisfies all the boundary and periodicity conditions (8)–(13). Substituting this form of $\mathbf{u}$ into (17), we derive the following problem: Find a function $\mathbf{z} \in V$ such that it satisfies the equation

$$a(\mathbf{g}^{**} + \mathbf{z}, \mathbf{v}) = (\mathbf{f},\mathbf{v}) + b(\mathbf{h},\mathbf{v}) \tag{18}$$

for all $\mathbf{v} \in V$.

The following theorem can be proved.

Theorem 1 (On the existence of a weak solution). *The weak problem (18) has a solution* $\mathbf{z}$ *that satisfies the estimate*

$$|||\mathbf{z}||| \ \le \ R. \tag{19}$$

where R *is a constant. Consequently, the weak problem from Definition 1 has a solution* $\mathbf{u}$ $(= \mathbf{z} + \mathbf{g}^{**})$ *that satisfies*

$$\|\nabla \mathbf{u}\|_{L^2(\Omega)^2} \ \le \ R + \|\nabla \mathbf{g}^{**}\|_{L^2(\Omega)^2} \ \le \ R + c\,\|\mathbf{g}\|_{H^s(\Gamma_i)^2} \ := \ R'. \tag{20}$$

The proof of that theorem follows the usual way, i.e., we need to prove to coercivity of the form a*. Using the definition of* $a(\mathbf{g}^{**} + \mathbf{z}, \mathbf{z})$*, we obtain:*

$$\begin{aligned} a(\mathbf{g}^{**} + \mathbf{z}, \mathbf{z}) \ &= \ a_1(\mathbf{g}^{**}, \mathbf{z}) + a_1(\mathbf{z}, \mathbf{z}) + a_2(\mathbf{g}^{**}, \mathbf{g}^{**}, \mathbf{z}) \\ &+ a_2(\mathbf{z}, \mathbf{g}^{**}, \mathbf{z}) + a_2(\mathbf{g}^{**} + \mathbf{z}, \mathbf{z}, \mathbf{z}) + a_3(\mathbf{g}^{**} + \mathbf{z}, \mathbf{g}^{**} + \mathbf{z}, \mathbf{z}). \end{aligned} \tag{21}$$

In order to prove coercivity we have to estimate all of the terms on the right hand side of (21). The term $a_1(\mathbf{z}, \mathbf{z})$ *is the "good" term as it is equal to* $\nu|||\mathbf{z}|||^2$*. Terms where* $\mathbf{z}$ *is just linear do not cause any problems but we have to estimate the remaining terms very carefully. Here the formulation of* $\mathbf{g}^{**}$ *is a key part. Putting all the estimates together we obtain the following inequality*

$$\begin{aligned} a(\mathbf{g}^{**} + \mathbf{z}, \mathbf{z}) \ &\ge \ \frac{\nu}{2}\,|||\mathbf{z}|||^2 - \nu\, c\,|||\mathbf{z}|||\,\|\mathbf{g}\|_{H^s(\Gamma_i)^2} - c\,\|\mathbf{g}\|^2_{H^s(\Gamma_i)^2}\,|||\mathbf{z}||| \\ &= |||\mathbf{z}|||\left(\frac{\nu}{2}\,|||\mathbf{z}||| - \nu\, c(\epsilon)\,\|\mathbf{g}\|_{H^s(\Gamma_i)^2} - c\,\|\mathbf{g}\|^2_{H^s(\Gamma_i)^2}\right). \end{aligned} \tag{22}$$

Constant c *is a generic constant coming from the estimates of the individual terms. From (22) is obvious that for* $|||\mathbf{z}||| \to +\infty$ *is* $a(\mathbf{g}^{**}+\mathbf{z}, \mathbf{z}) \to +\infty$ *and the coercivity of the form* a *is ensured. (The complete proof can be found in [5].)*

Conclusion The main goal of this paper is to show that the usual restriction on the incoming flow, which has the form of restriction on the prescribed velocity profile $\mathbf{g}$ or some additional restrictions imbued on the integral form a, can be avoided. In fact my idea is that problem arise from the strong flow between two parts of boundary (here Γ_i and Γ_o). In the construction of the $\mathbf{g}^{**}$ we respected this idea and we put a stronger control on the incoming flow. As you can see it is then possible to prove the existence without any restrictions. The payment for this result is the more complicated form of the outflow boundary condition and, for technical reasons, the rectangular shape of the domain Ω.

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. M. Feistauer: Mathematical Methods in Fluid Dynamics. Pitman Monographs and Surveys in Pure and Applied Mathematics 67, Longman Scientific & Technical, Harlow, 1993
2. M. Feistauer, T. Neustupa: On Some Aspects of Analysis of Incompressible Flow Through Cascades of Profiles. Operator Theory, Advances and Applications, Vol. 147, Birkhauser, Basel, 257–276, 2004
3. G. P. Galdi: An Introduction to the Mathematical Theory of the Navier–Stokes Equations, Vol. I: Linearized Steady Problems. Springer Tracts in Natural Philosophy, Vol. 38, Springer, Berlin, 1994
4. V. Girault, P.-A. Raviart: Finite Element Approximation of the Navier–Stokes Equations. Lecture Notes in Mathematics 749, Springer, Berlin, 1979
5. T. Neustupa: Mathematical Modelling of Viscous Incompressible Flow through a Cascade of Profiles. Dissertation Thesis. Faculty of Mathematics and Physics, Charles University Prague, 2007 (Informations on tneu@centrum.cz)
6. R. Temam: Navier–Stokes Equations. North-Holland, Amsterdam, 1977

Variational and Heterogeneous Multiscale Methods

Jan Martin Nordbotten

Abstract Both the variational and heterogeneous multiscale methods are presented for non-linear variational problems. We show that the variational multiscale method can be seen as a subset of the methods which can be defined within the heterogeneous multiscale method framework. Our results extend to the approximate forms of the multiscale methods which are of interest to applications.

1 Introduction

Classical numerical discretization methods for partial differential equations are formulated at a single scale. Text book examples include the both finite elements and finite difference methods (see e.g., [2, 8, 18]). Single scale methods are well suited for relatively smooth solutions, as is apparent from the derived error estimates (see e.g., [2]). However, as applications have become ever more complex, the need has arisen for numerical methods requiring less smoothness of the solution. Methods addressing these aspects can be called "multiscale."

Among the influential developments of multiscale methods are generalized finite elements [1], (numerical) upscaling methods (see e.g., [7,9] and references therein). Additionally, linear solver for single scale methods have increasingly been adapted to tackle solutions with multiscale nature [3,4,19].

In this contribution, we will illustrate how two modern relatives to the generalized finite element and numerical upscaling approaches are related. In particular, we consider the finite element formulations of the Variational MultiScale (VMS) method [11, 12] and the Heterogeneous Multiscale Method (HMM) [10]. For relationships between existing generalized finite element methods, see e.g., [5, 15]. For error analysis and discussion of both MsFEM (a special form of VMS) and HMM for elliptic problems, see e.g., [6]. While the VMS and HMM are developed from

J.M. Nordbotten
Department of Mathematics, University of Bergen, Bergen, Norway
e-mail: jan.nordbotten@math.uib.no

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_76,

different perspectives, we show that VMS can be interpreted as an instance of the Finite Element HMM (HMFEM).

The relationship between VMS and HMFEM may seem trivial when the methods are presented in similar notation. Yet, since the fundamental ideas and conventional expositions of the methods are different, the similarities have not been widely appreciated. It is thus our goal to give a clear and concise presentation of the relationship between VMS and HMFEM.

We state minimization problems as: Denote by $u \in V$ the element

$$u = \arg\min_{v \in V} A(v) - B(v) \tag{1}$$

for convex and linear functionals A and B, respectively, and a suitable function space V. We describe the variational problem as: Find $u \in V$ such that

$$a(u, v) = b(v) \quad \forall\, v \in V. \tag{2}$$

When a and b are the (Gâteaux) derivatives of A and B, (2) will be satisfied for all solutions of (1). A special case arises when A is quadratic, which implies that a will be bi-linear.

2 VMS

In this section we will recall the VMS methodology for solving problems of the type given in (2). We restrict ourselves in this section to problems where there exists a unique solution. Furthermore, we consider only cases where a and b are linear in v, which is the case when (2) is derived as the weak form of some equation

$$\mathscr{L} u = b.$$

To facilitate the later discussion, the original notation of Hughes for linear problems, see e.g., [12], has been adapted slightly.

The standard VMS approach to solving (2) is to split V into a direct sum of the resolved (usually finite dimensional) space V_H and an unresolved fine scale V'. Thus $V = V_H \oplus V'$. The decomposition will in general not be orthogonal with respect to the energy norm induced by a. We impose the restriction that this decomposition is chosen such that all the fine and coarse scale problems refered to in the continuation have unique solutions (at least one such decompositions exists since the original problem has a unique solution). Thus we can consider the variational problem: Find $u_H \in V_H$ s.t.:

$$a(u_H + u', v_H) = b(v_H) \quad \forall\, v_H \in V_H, \tag{3}$$

where $u' \in V'$ satisfies:

$$a(u' + u_H, v') = b(v') \quad \forall\, v' \in V'. \tag{4}$$

Formally, when a is bi-linear, we can write the solution of (4) using the Green's operator: $u' = G'(\mathscr{L}u_H) - G'(b)$. We apply the notion of Green's functions to non-linear problems, and adapt the notation that $u' = G'(u_H, b)$.

We may then substitute for u' in (3) to obtain a problem posed only on the coarse scale: Find $u_H \in V_H$ such that

$$a(u_H + G'(u_H, b), v_H) = b(v_H) \quad \forall\, v_H \in V_H. \tag{5}$$

This is the form of the VMS method referred to by Hughes et al. as a paradigm for multiscale modelling [12].

The variational form given in (5) can be considered as a Petrov–Galerkin type method for the original problem (the solution is in a subspace of $V_H \oplus G'(V_H)$, while the trial space is V_H). We can achieve an equivalent Galerkin method by observing from (4) that $a(u_H + G'(u_H, b)), G'(v_H, b)) = b(G'(v_H, b))$. Since a is linear in the second argument, by addition we then have: Find $u_H \in V_H$ such that

$$a(u_H + G'(u_H, b)), v_H + G'(v_H, b)) = b(v_H + G'(v_H, b)) \quad \forall\, v_H \in V_H. \tag{6}$$

We refer to this as the symmetric form of VMS.

When a and b are functional derivatives, we then have that the solution of (5) is a stationary point for $u_H \in V_H$ of the expression

$$A(u_H + G'(u_H, b)) - B(u_H + G'(u_H, b)) \tag{7}$$

with respect to variations of the form $\epsilon(v_H + G'(v_H, b))$. Similarly, (5) is a stationary point of (7) with respect to variations of the form ϵv_H.

Remark 1. Equations (5)–(7) are all exact equations for the component of the solution of the original problem in V_H, in terms of the projection introduced by the direct sum decomposition into V_H and V'.

Remark 2. The advantage of the symmetric vs. the non-symmetric form of VMS is not clear. Early experience in one implementation indicates that the non-symmetric form is computationally more efficient due to fewer elements in the coarse scale system matrix, however it tends to be less accurate [16].

In practice, the computation of G' is too expensive, and we use the approximation $\widetilde{G}' \approx G'$. The choice of $\widetilde{G}'$ is essential for the successful application of the VMS framework.

3 HMFEM

We will in this section present the derivation of the Heterogeneous Multiscale Finite Element Method, as originally described in [10].

Consider this time the minimization problem given in (1), and assume it has a unique solution. By introducing a compression operator $\mathscr{Q} : V \to V_D$, where V_D is some coarse scale solution domain, we have the minimization problem equivalent to (1):

$$\min_{v \in V} A(v) - B(v) = \min_{v_D \in V_D} \min_{v : \mathscr{Q}v = v_D} A(v) - B(v). \tag{8}$$

We restrict the choice of compression operators under consideration such that also (8) has a unique solution v_D. We denote a reconstruction operator as any operator $\mathscr{R} : V_D \to V$ such that $u_D - \mathscr{Q}\mathscr{R}u_D = 0$ for all $u_D \in V_D$. We note that an "exact" reconstruction operator with respect to the minimization problem can be defined from (8):

$$\mathscr{R}_e u_D = \arg \min_{v : \mathscr{Q}v = u_D} A(v) - B(v). \tag{9}$$

We now have the coarse scale HMFEM minimization problem

$$\min_{v_D \in V_D} A(\mathscr{R}v_D) - B(\mathscr{R}v_D). \tag{10}$$

This problem is termed "exact" if the reconstruction is exact, $\mathscr{R} = \mathscr{R}_e$.

We now give three variational forms of HMFEM based on (10). If we apply a standard variational approach to (10), we obtain: Find $u_D \in U_D$ such that

$$a(\mathscr{R}u_D, r(u_D, v_D)) - b(r(u_D, v_D)) = 0 \qquad \forall\, v_D \in U_D. \tag{11}$$

Here, a, b and r are the derivatives of A, B and $\mathscr{R}$, respectively.

There are several other ways to derive a variational form of the minimization problem. We will consider two natural choices, which both avoid the (possibly complex) calculation of r. Our first approach takes coarse scale variations around the reconstructed solution, e.g., $\mathscr{R}u_D + \epsilon v_D$, leading to: Find $u_D \in U_D$ such that

$$a(\mathscr{R}u_D, v_D) - b(v_D) = 0 \qquad \forall\, v_D \in U_D. \tag{12}$$

Alternatively, we may consider reconstructed variations, e.g., $\mathscr{R}u_D + \epsilon\mathscr{R}v_D$, which leads to: Find $u_D \in U_D$ such that

$$a(\mathscr{R}u_D, \mathscr{R}v_D) - b(\mathscr{R}v_D) = 0 \qquad \forall\, v_D \in U_D. \tag{13}$$

This latter formulation preserves symmetry of the operator a with respect to the multiscale approach.

Remark 3. For exact reconstructions $\mathscr{R}_e$, all three variational formulations presented above are consistent with the fine scale problem (1), and when unique solutions exist, they will be equivalent.

Remark 4. For problems where $\mathscr{R}$ is linear, we have that $r(u_D, v_D) = \mathscr{R}v_D$, and thus formulations (11) and (13) are identical.

For practical purposes, calulating the exact $\mathscr{R}_e$ is excessively expensive, and an approximation is introduced; $\widetilde{\mathscr{R}} \approx \mathscr{R}_e$. When employing approximate reconstructions, an appreciable difference may be seen between the three variational formulations (11)–(13), see e.g., [16].

It is usually advocated (see e.g., [10]) that since u_D is a macro-scale function, it should vary smoothly, thus it is sufficient to evaluate the integrals appearing in the variational formulation at quadrature points. This allows for great flexibility in localization strategies for approximating $\widetilde{\mathscr{R}}$.

4 The Relationship Between VMS and HMFEM

The motivation and development of VMS and HMFEM is clearly different, which is apparent both from the descriptions given in the preceeding sections as well as from the analysis conducted by previous authors (in addition to the cited works, see also [14, 17]). Nevertheless, we see an immediate similarity when we consider (5)–(7) and (10)–(13). In this section we will formalize these relationships.

4.1 The Relationship Between Exact VMS and HMFEM

This section will make clear the relationships between the exact VMS and HMFEM formulations. In particular, we show that any VMS method is equivalent to a HMFEM, but that the converse is not true.

We denote by "exact," when the Green's operator (for VMS) or the reconstruction operator (for HMFEM) is solved exactly. It follows that

Lemma 1. *Given the solution u_H of exact VMS method, the sum $u = u_H + G'(u_H, b)$ solves the original problem.*

This lemma is a consequence of the observation that no approximations were introduced in the development of (5). This has been commented on since the original derivation as one of the advantages of the VMS framework [12]. Lemma 1 extends to the solution of (6)–(7) when valid.

Similarly, it holds for the HMFEM that

Lemma 2. *Given the solution u_D of exact HMFEM method, the reconstruction $u = \mathscr{R}_e u_D$ solves the original problem.*

Again, this holds for the solution obtained from any of (10)–(11).

We now turn our attention to the relationship between the coarse scale solutions obtained by VMS and HMFEM. Before we state the main result, we summarize properties of the coarse scale VMS and HMFEM solutions.

The coarse scale VMS solution is uniquely defined by the direct sum decomposition of V into V_H and V'. Indeed, let the (linear) projection operators of the decomposition be denoted $\mathscr{P}_H$ and $\mathscr{P}'$. Then the solution u_H is simply $u_H = \mathscr{P}_H u$. This is trivially seen, since we have from Lemma 1 that $u = u_H + G'(u_H, b)$, and thus $\mathscr{P}_H u = \mathscr{P}_H u_H + \mathscr{P}_H G' = u_H$.

For the HMFEM method, a direct sum decomposition is not defined, and the coarse solution is defined by the choice of coarse space U_D and the compression operator $\mathscr{Q}$.

Theorem 3. *Let a and b be the derivatives of A and B. Further, take the coarse spaces V_H and U_D to be identical. Define the coarse VMS solution uniquely by a choice of V' such that $V = V_H \oplus V'$, and associate with the decomposition the projection operator $\mathscr{P}_H$. Then, if $\mathscr{Q} = \mathscr{P}_H$, the coarse solutions u_H of the VMS method and u_D of the HMFEM are identical.*

Proof. We have from Lemmas 1 and 2 that

$$u = u_H + G'(u_H, b) = \mathscr{R}_e u_D.$$

Now, since $\mathscr{Q} = \mathscr{P}_H$, then $\mathscr{P}_H \mathscr{R}_e u_D = u_D$, as follow from the definition of a reconstruction operator in Sect. 3. Thus, by projecting the above relationship into V_H, we have

$$u_H = u_D.$$

Remark 5. Theorem 3 tell us that any VMS method can be desribed by as a HMFEM by setting $\mathscr{Q} = \mathscr{P}_H$. However, the converse need not necessarily be true. This can be observed by noting that there is no point in the derivation of the HMFEM where $\mathscr{Q}$ is required to be linear, however this property is needed for the direct sum property utilized by the VMS method.

Remark 6. An important consequense of the VMS derivation (as contrasted with HMFEM), is the direct application to the variational form. This implies that the VMS method can be applied directly without explicit knowledge of the functional, which need not have an extremum at the solution. Our results above show that these problems may indeed also be treated with HMFEM.

4.2 The Relationship Between Approximate VMS and HMFEM

For applications, it is not meaningful to calculate the exact subscale Green's functions or the reconstruction operator. This is a consequense of the fact that no approximations have been made, and the computational complexity is therefore

(at least) the same as for the original fine scale problem. Therefore, approximate Green's functions and reconstruction operators will invariably be used, and the purpose of both VMS and HMFEM is primarily to provide a framework for these approximations. The results will mainly be straight applications of the results from the previous section, and we will therefore in this section take $\mathscr{Q} = \mathscr{P}_H$.

Throughout this section, we will assume that the coarse spaces $V_H = U_D$ are identical and finite dimensional. Approximation strategies all involve finite dimensional approximations to the fine scale problems give by (4) and (9). We will not go into details of these strategies herein, however we will assume that the same localization strategy is applied to both VMS and HMFEM. The interested reader can see e.g., [10, 12, 13, 16], for detailed discussion on localization strategies.

We will make precise the meaning of equivalent localization strategies. The variational form of the reconstruction problem for HMFEM can be stated as: Find $u = \mathscr{R}u_D \in V$ such that $\mathscr{Q}u = u_D$ and:

$$a(u, v) = b(v) \qquad \forall\ v \text{ such that } q(u, v) = 0. \tag{14}$$

Here q is the derivative of $\mathscr{Q}$. Recall that $\mathscr{Q} = \mathscr{P}_H$ is a linear projection, so that $q(u, v) = \mathscr{Q}v$, and introduce $u' = \mathscr{Q}u - u \equiv \mathscr{P}'u$. We can now write (14) as: Find u' in V' such that:

$$a(u' + u_D, v') = b(v') \qquad \forall\ v' \in\ V'. \tag{15}$$

We recognize that (15) is identical to the fine scale problem derived for the VMS method, and we introduce the definition:

Definition 4. The approximation applied to the restriction operator of HMFEM is termed *equivalent* to the approximation of the fine scale Green's function in VMS if the following relationship holds: $\widetilde{\mathscr{R}}u_D = u_D + \widetilde{G}'(u_D, b)$ for all $u_D \in U_D$.

We now give three corollaries which follow directly from Theorem 3 and Definition 4.

The following result is applicable to the Petrov–Galerkin and Galerkin type formulations for variational problems:

Corollary 5. *Under the same assumptions as in Theorem 3, and let the approximations of the fine scale problems be equivalent. Then the coarse solutions u_H of the non-linear VMS method given by (5) and u_D of the HMFEM on variational form given by (12) are identical. Equivalently, the solution u_H of (6) and u_D of (13) are identical.*

Having discussed the variational formulations, it remains to consider the minimization form of the HMFEM. Based on Remark 4, and Corollary 5, we have the final corollary:

Corollary 6. *Under the same assumptions as in Theorem 3, and let the approximations of the fine scale problems be equivalent. If in addition the reconstruction*

operator $\widetilde{\mathcal{R}}$ is linear, then the coarse solutions u_H of the non-linear VMS method given by (6) and u_D of the HMFEM on minimization form given by (10) are identical.

Acknowledgements The author wishes to thank Weinan E and Talal Rahman for interesting discussions during the writing of this manuscript.

References

1. I. Babuska and J. E. Osborn, *Generalized Finite Element Methods: Their Performance and Their Relation to Mixed Methods*, SIAM J. Numer. Anal., 20(3), 510–536, 1983
2. D. Braess, *Finite elements*, Cambridge University Press, Cambridge, 1997
3. A. Brandt, *Barriers to Achieving Textbook Multigrid Efficiency (TME) in CFD*, URL: http://hdl.handle.net/2002/14809, 1998
4. A. Brandt, *Multiscale Solvers and Systematic Upscaling in Computational Physics*, Comp. Phys. Com., 169, 438–441, 2005
5. F. Brezzi, L. P. Franca, T. J. R. Hughes and A. Russo, $b = \int g$, Comput. Meth. Appl. Mech. Eng., 145, 329–339, 1997
6. Z. Chen, *Multiscale Methods for Elliptic Homogenization Problems*, Numer. Meth. Part. Diff. Equat., 22, 317–360, 2006
7. Z. Chen and T. Y. Hou, *A Mixed Multiscale Finite Element Method for Elliptic Problems with Oscillating Coefficients*, Math. Comp. 72(242), 541–576, 2002
8. Z. Chen, G. Huan and Y. Ma, *Computational Methods for Multiphase Flows in Porous Media*, Society of Industrial and Applied Mathematics, Philadelphia, 2006.
9. L. Durlofsky, *Coarse Scale Models of Two Phase Flow in Heterogeneous Reservoirs: Volume Averaged Equations and their Relationship to Existing Upscaling Techniques*, Comp. Geosci., 2, 73–92, 1998
10. W. E and B. Engquist, *The heterogeneous multi-scale methods*, Comm. Math. Sci., 1, 87–133, 2003
11. T. J. R. Hughes and G. Sangalli, *Variational Multiscale Analysis: The Fine-scale Green's Function, Projection, Optimization, Localization, and Stabilized Methods* SINUM, 45(2), 539–557, 2007
12. T. J. R. Hughes, G. R. Feijoo, L. Mazzei and J. B. Quincy, *The Variational Multiscale Method – A Paradigm for Computational Mechanics*, Comp. Meth. Appl. Mech. and Eng., 166(1–2), 3–24, 1998
13. M. G. Larson, A. Malqvist, *Adaptive Variational Multiscale Methods Based on a Posteriori Error Estimation: Energy Norm Estimates for Elliptic Problems*, Comp. Meth. Appl. Mech. Eng., 196(21–24), 2313–2324, 2007
14. P. B. Ming and P.-W. Zhang, *Analysis of the Heterogeneous Multiscale Method for Parabolic Homogenization Problems*, Math. Comput., 76(257), 153–157, 2007
15. J. Nolen, G. Papanicolaou and O. Pironneau, *A Framework for Adaptive Multiscale Methods for Elliptic Problems*, SIAM Multiscale Model. Simul., 7, 171–196, 2008
16. J. M. Nordbotten, *Adaptive Variational Multiscale Methods for Multi-Phase Flow in Porous Media*, SIAM Multiscale Model. Simul., 7(3), 1455–1473, 2009
17. M. Ohlberger, *A Posteriori Error Estimates for the Heterogeneous Multiscale Finite Element Method for Elliptic Homogenization Problems*, SIAM Multiscale Model. Simul., 4(1), 88–144, 2005
18. M. Shashkov, Conservative finite-difference methods on general grids, CRC, FL, 1996
19. B. F. Smith, P. E. Bjørstad and W. D. Gropp, Domain Decomposition, Cambridge University Press, Cambridge, 1996

Discrete Dislocation Dynamics and Mean Curvature Flow

Petr Pauš, Michal Beneš, and Jan Kratochvíl

Abstract This contribution deals with the numerical simulation of dislocation dynamics by means of parametric mean curvature flow. Dislocations are described as an evolving family of closed and open smooth curves driven by the normal velocity. The equation is solved using direct approach by semi-discrete scheme based on finite difference method. Numerical stability is improved by tangential redistribution of curve points which allows long time computations and better accuracy. Our method contain an algorithm which allows topological changes. The results of dislocation dynamics simulation are presented.

1 Introduction

The dislocations are defined as irregularities or errors in crystal structure of the material. The presence of dislocations strongly influences many of material properties. Plastic deformation in crystalline solids is carried by dislocations. Theoretical description of dislocations is widely provided in literature such as [5, 8, 9, 16]. Dislocation is a line defect of the crystalline lattice. Along the dislocation curve the regularity of the crystallographic arrangement of atoms is disturbed. The dislocation can be represented by a curve closed inside the crystal or by a curve ending on the surface of the crystal. At low homologous temperatures the dislocations can move only along crystallographic planes (gliding planes) with the highest density of atoms. The motion results in mutual slipping of neighboring parts of the crystal along the gliding planes.

P. Pauš (✉) and M. Beneš
Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, Czech Technical University, Prague
e-mail: petr.paus@fjfi.cvut.cz, michal.benes@fjfi.cvut.cz

J. Kratochvíl
Department of Physics, Faculty of Civil Engineering, Czech Technical University Prague
e-mail: kratochvil@fsv.cvut.cz

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_77,

This justifies the importance of developing suitable mathematical models [2,6,7, 12–15, 18–20]. From the mathematical point of view, the dislocations can be represented by smooth closed or open plane curves which evolve in time. Their motion is two-dimensional as they move in glide planes. The evolving curves can be mathematically described in several ways. One possibility is to use the *level-set method* [4, 17, 22], where the curve is defined by the zero level of some surface function. One can also use the *phase-field method* [1].

2 Dislocations and Mean Curvature Flow

The interaction of dislocations and bulk elastic field can be approximately described using the curvature flow as follows (see [21]). We consider perfect dislocation curves with the Burgers vector $\mathbf{b} = (b, 0, 0)$ oriented in the x-direction of the x,y,z coordinate system. The discloation curve motion Γ is located in a glide plane, in our case in the xz-plane. The glide of dislocation is governed by the relaxation law in the form of the mean curvature flow equation in the direction of the normal vector

$$Bv = L\kappa + b\tau_{app}, \tag{1}$$

where B is a drag coefficient, and $v(x,t)$ is the normal velocity of a dislocation at $x \in \Gamma$ and time t. The term $L\kappa$ represents self-force expressed in the line tension approximation as the product of the line tension L and local curvature $\kappa(x,t)$. The term τ_{app} represents the local shear stress acting on the dislocation segment produced by the bulk elastic field. In our simulations, we consider "stress controlled regime" where the applied stress in the channel is kept uniform. This is an upper bound limit case. The other limiting case is "strain controlled regime" as described in [6, 7]. The applied stress τ_{app} is the same in every point of the line and for numerical computations we use $\tau_{app} = const$.

3 Parametric Description

The motion law (1) in the case of dislocation dynamics is treated by parametrization where the planar curve $\Gamma(t)$ is described by a smooth time-dependent vector function $X : S \times I \to \mathbb{R}^2$, where $S = [0, 1]$ is a fixed interval for the curve parameter and $I = [0, T]$ is the time interval. The curve $\Gamma(t)$ is then given as the set

$$\Gamma(t) = \{X(u,t) = (X^1(u,t), X^2(u,t)), u \in S\}.$$

The evolution law (1) is transformed into the parametric form as follows. The unit tangential vector $\mathbf{T}$ is defined as $\mathbf{T} = \partial_u X/|\partial_u X|$. The unit normal vector $\mathbf{N}$ is perpendicular to the tangential vector and $\mathbf{N} \cdot \mathbf{T} = 0$ holds. The curvature κ is

defined as

$$\kappa = \frac{\partial_u X^\perp}{|\partial_u X|} \cdot \frac{\partial_{uu} X}{|\partial_u X|^2} = \mathbf{N} \cdot \frac{\partial_{uu} X}{|\partial_u X|^2},$$

where $X^\perp$ is a vector perpendicular to X. The normal velocity v is defined as the time derivative of X projected into the normal direction, $v = \partial_t X \cdot \partial_u X^\perp / |\partial_u X|$. Equation (1) can now be written as

$$B \partial_t X \cdot \frac{\partial_u X^\perp}{|\partial_u X|} = L \frac{\partial_{uu} X}{|\partial_u X|^2} \cdot \frac{\partial_u X^\perp}{|\partial_u X|} + b \tau_{app},$$

which holds provided the vectorial evolution law is satisfied

$$B \partial_t X = L \frac{\partial_{uu} X}{|\partial_u X|^2} + b \tau_{app} \frac{\partial_u X^\perp}{|\partial_u X|}. \tag{2}$$

This equation is accompanied by the periodic boundary conditions for closed curves, or by fixed-end boundary condition for open curves, and by the initial condition. These conditions are considered similarly as in [3].

The solution of (2) exhibits a natural redistribution property which is useful for short-time curve evolution [10,18]. For long time computations with time and space variable force, the algorithm for curvature adjusted tangential velocity is used. This algorithm moves points along the curve according to the curvature, i.e., areas with higher curvature contain more points than areas with lower curvature. To incorporate a tangential redistribution, a tangential term α has to be added to (2).

$$B \partial_t X = L \frac{\partial_{uu} X}{|\partial_u X|^2} + L\alpha \frac{\partial_u X}{|\partial_u X|} + b \tau_{app}, \frac{\partial_u X^\perp}{|\partial_u X|}. \tag{3}$$

This improves numerical stability and also accuracy of computation. Details are described in [13,23].

4 Numerical Scheme

For numerical approximation we consider a regularized form of (3) which reads as

$$B \partial_t X = L \frac{\partial_{uu} X}{Q(\partial_u X)^2} + L\alpha \frac{\partial_u X}{Q(\partial_u X)} + b \tau_{app}, \frac{\partial_u X^\perp}{Q(\partial_u X)}, \tag{4}$$

where $Q(x_1, x_2) = \sqrt{x_1^2 + x_2^2 + \varepsilon^2}$ is a regularization term and ε a small parameter. We use the backward Euler semi-implicit scheme for numerical solution of the differential equation (3). The first derivative is discretized by backward difference as follows

$$\partial_u X|_{u=jh} \approx \left[\frac{X_j^1 - X_{j-1}^1}{h}, \frac{X_j^2 - X_{j-1}^2}{h}\right],$$

and the second derivative as

$$\partial_{uu} X|_{u=jh} \approx \left[\frac{X_{j+1}^1 - 2X_j^1 + X_{j-1}^1}{h^2}, \frac{X_{j+1}^2 - 2X_j^2 + X_{j-1}^2}{h^2}\right].$$

The approximation of the first derivative is denoted as $X_{\bar{u},j}$ and the second derivative as $X_{\bar{u}u,j}$.

The semi-implicit scheme for (4) has the following form

$$BX_j^{k+1} - Lt\frac{X_{\bar{u}u,j}^{k+1}}{Q^2(X_{\bar{u},j}^k)} - Lt\alpha_j\frac{X_{\bar{u},ej}^{k+1}}{Q(X_{\bar{u},j}^k)} = BX_j^k + tb\tau_{app}\frac{X_{\bar{u},j}^{\perp k}}{Q(X_{\bar{u},j}^k)}, \qquad (5)$$
$$j = 1, \cdots, m-1, k = 0, \cdots, N_T - 1,$$

where $Q(x_1, x_2)$ is a regularization term, $X_{\bar{u},j}^{\perp}$ is a vector perpendicular to $X_{\bar{u},j}$, and α_j is redistribution coefficient. The term ε serves as a regularization to avoid singularities when the curvature tends to infinity. $X_j^k \approx X(jh, kt)$, t is a time step and N_T is the number of time steps. The matrix of the system (5) for one component of X^{k+1} has the following tridiagonal structure:

$$\begin{pmatrix} B + \frac{2tL}{h^2Q^2} - \frac{tL\alpha}{hQ} & \frac{-tL}{h^2Q^2} & 0 & \cdots \\ \frac{-tL}{h^2Q^2} + \frac{tL\alpha}{hQ} & \ddots & \ddots & \ddots \\ 0 & \ddots & & \\ \vdots & \ddots & & \end{pmatrix}.$$

The scheme (5) is solved for each k by means of matrix factorization. Since there are two components of X, two linear systems are solved in each timestep.

5 Application in Dislocation Dynamics

Dislocation curves as defects in material evolve in time. The dislocation evolution history contains shape changes of open curves, closing of open dislocation curves up to collision of dipolar loops (see [9, 16]). Interaction of dislocation curves and dipolar loops has been studied, e.g., in [6,7,12–15]. Our numerical simulations were performed under the following set of parameters:

Burgers vector magnitude	$b = 0.25\,\text{nm}$
Line tension	$L = 2\,\text{nN}$
Drag coefficient	$B = 1.0 \cdot 10^{-5}\,\text{Pa} \cdot \text{s}$
Applied stress	$\tau_{app} = 40\,\text{MPa}$

Dislocations can interact with other defects through the stress field. In this case, dislocation curve can be blocked by a potential barrier. Figure 1 illustrates the evolution of an open dislocation curve through an obstacle in material (a precipitate). In the example, the curve is fixed at $[-300\,\text{nm}, 0\,\text{nm}]$ and $[300\,\text{nm}, 0\,\text{nm}]$ which may be caused by some impurities in the material or it can continue in another slip plane. The obstacle has a form of a circle located at $[0\,\text{nm}, 400\,\text{nm}]$ with a radius of 40 nm. Due to external stress, the dislocation curve expands but the obstacle blocks the evolution. The curve surrounds it. At a certain time, it touches itself and splits into two curves, an open curve and a closed curve. The closed curve cannot evolve anymore because of the obstacle. The open curve continues expansion. The simulation was performed with the following parameters. The number of discretization points is $M = 200$, the external stress applied to the dislocation $\tau_{app} = 40\,\text{MPa}$, the time of simulation $t \in (0, 0.088)$.

The example in Fig. 2 shows the simulation of the Frank–Read mechanism (see [5, 16]) which describes how new dislocation loops are created. The open dislocation curve is fixed at $[-150\,\text{nm}, 0\,\text{nm}]$ and $[150\,\text{nm}, 0\,\text{nm}]$, and is forced to evolve under the applied stress $\tau_{app} = 40\,\text{MPa}$. The evolution continues until it touches

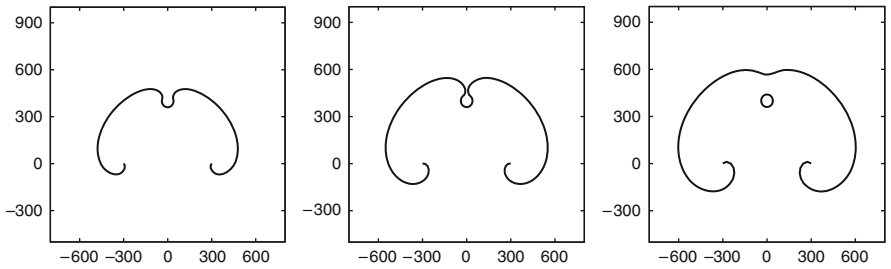


Fig. 1 Evolution through a strong obstacle, $F_O = 0.01\,\text{nN}$, $\tau_{app} = 40\ MPa$, $t \in (0, 0.088)$, curve discretized by $M = 200$ nodes, for $t = 0.054\,\text{s}$, $t = 0.074\,\text{s}$ and $t = 0.088\,\text{s}$

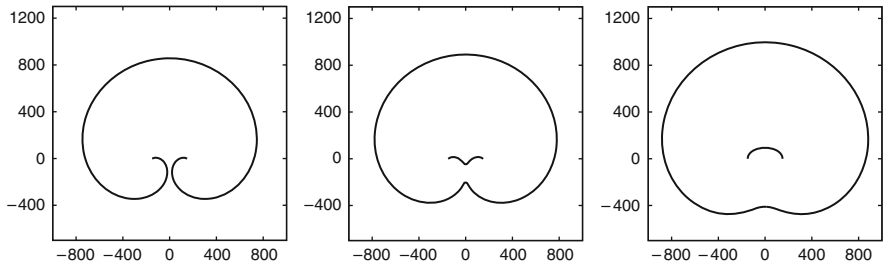


Fig. 2 Frank-Read source, $\tau_{app} = 40\,\text{MPa}$, $t \in (0, 0.29)$, curve discretized by $M = 400$ nodes, for $t = 0.25\,\text{s}$, $t = 0.26\,\text{s}$ and $t = 0.29\,\text{s}$

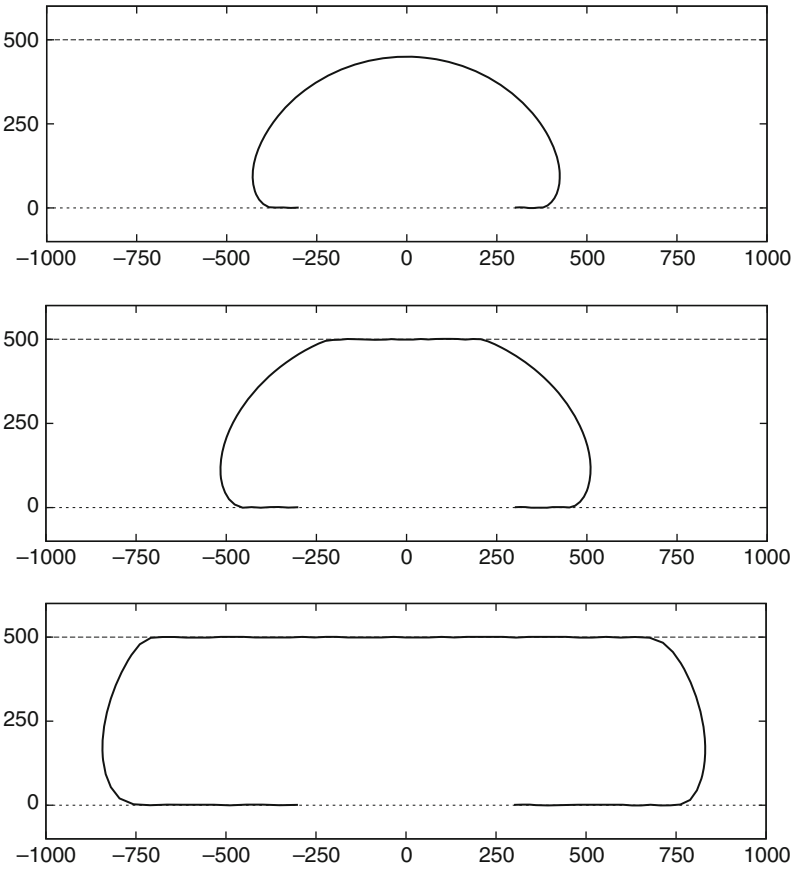


Fig. 3 Single dislocation in an infinite channel, $\tau_{app} = 40\,\text{MPa}$, $t \in (0, 0.154)$, curve discretized by $M = 200$ nodes, for $t = 0.04\,\text{s}$, $t = 0.065\,\text{s}$ and $t = 0.154\,\text{s}$

itself. At this moment, the curve splits into two parts, i.e., the dipolar loop and the dislocation line. The loop continues in expansion. The dislocation line will again undergo the same process. The Frank–Read source cannot generate unlimited number of dislocation loops because new loops interact with each other and slow down the source. The source can usually generate about 300 or 400 of dipolar loops (see [9]). Parameters of the simulation are $t \in (0, 0.29)$, $M = 400$.

Figure 3 illustrates the behavior of an open dislocation curve in an infinite channel. The channel is created by a spatially variable external force $F_C = 0.01\,\text{N}$ for $z < 0\,\text{nm}$ and $z > 500\,\text{nm}$. The curve expands upwards due to external stress $\tau_{app} = 40\,\text{MPa}$. The upper channel wall restricts its movement and the curve can therefore evolve aside only. The algorithm for curvature adjusted redistribution of points allows to rarify number of discretization points along straight parts of the dislocation and accumulate discretization points at parts with higher curvature. This

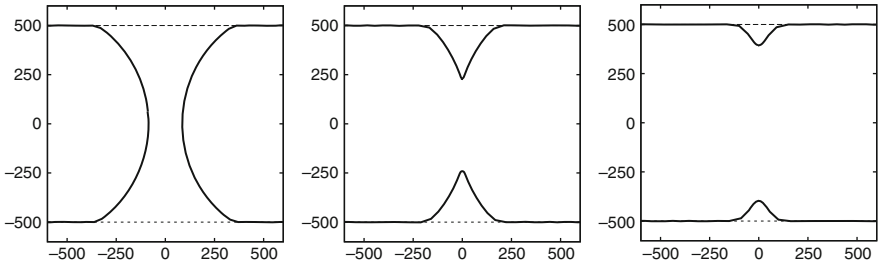


Fig. 4 Merging two dislocations in a channel, $\tau_{app} = 40\,\text{MPa}$, $F_C = 0.01\,\text{N}$ for $z < -500\,\text{nm}$ and $z > 500\,\text{nm}$, $t \in (0, 0.102)$, each curve discretized by $M = 100$ nodes, for $t = 0.0565\,\text{s}$, $t = 0.087\,\text{s}$ and $t = 0.102\,\text{s}$

results into more accurate and faster computations. The parameters of simulation are $t \in (0, 0.154)$, $M = 200$.

The simulation of cross-slip of two dislocations is shown in Fig. 4. The dislocations are moving in the channel created by a spatially variable external force $F_C = 0.01\,\text{N}$ for $z < -500\,\text{nm}$ and $z > 500\,\text{nm}$. At a certain time, they touch each other and connect. In real material, each dislocation can evolve in a different parallel plane. This case is not yet covered by the described model. Parameters of the simulation are $\tau_{app} = 40\,\text{MPa}$, $t \in (0, 0.102)$, $M = 100$.

6 Conclusion

The simulation of dislocation dynamics is important in practice as dislocations affect many material properties. Dislocation dynamics can be mathematically modelled by the mean curvature flow. We presented a method based on a parametric approach. We applied the model to situations similar to the real context including a mechanism of creating new dislocations (i.e., Frank–Read source, cross-slip, etc.). The scheme had to be improved by an algorithm for tangential redistribution of points and by an algorithm for topological changes for parametric model.

Acknowledgements This work was partly supported by the project MSM No. 6840770100 "Applied Mathematics in Technical and Physical Sciences" and by the project No. LC06052 "Nečas Center for Mathematical Modelling" of the Ministry of Education, Youth and Sport of the Czech Republic.

References

1. Beneš, M.: Phase field model of microstructure growth in solidification of pure substances. Acta Math. Univ. Comenianae 70, 123–151 (2001)
2. Beneš, M., Kratochvíl, J., Křišťan, J., Minárik, V., Pauš, P.: A parametric simulation method for discrete dislocation dynamics. Eur. J. Phys

3. Deckelnick, K., Dziuk, G.: Mean curvature flow and related topics. Front. Numer. Anal. 63–108 (2002)
4. Dziuk, G., Schmidt, A., Brillard, A., Bandle, C.: Course on mean curvature flow. Manuscript 75p., Freiburg (1994)
5. Hirth, J.P., Lothe, J.: Theory of dislocations. John Willey, New York (1982)
6. Křišťan, J., Kratochvíl, J.: Interactions of glide dislocations in a channel of a persistent slip band. Philos. Mag. 87(29), 4593–4613 (2007)
7. Křišťan, J., Kratochvíl, J., Minárik V., Beneš M.: Simulation of interacting dislocations glide in a channel of a persistent slip band. In: Modeling and Simulation in Materials Science and Engineering 17 (2009)
8. Kroupa F.: Long-range elastic field of semi-infinite dislocation dipole and of dislocation jog. Phys. Status Solidi 9, 27–32 (1965)
9. Kroupa, F.:Dislokace v pevných látkách. SNTL, Praha (1966)
10. Mikula, K., Ševčovič, D.: Evolution of plane curves driven by a nonlinear function of curvature and anisotropy. SIAM J. Appl. Math. 61(5), 1473–1501 (2001)
11. Mikula, K., Ševčovič, D.: Computational and qualitative aspects of evolution of curves driven by curvature and external force. Comput. Vis. Sci. 6(4), 211–225 (2004)
12. Minárik, V., Kratochvíl, J.: Dislocation dynamics – Analytical description of the interaction force between dipolar loops. Kybernetik 43, 841–854 (2007)
13. Minárik, V., Beneš, M., Kratochvíl, J.: Simulation of dynamical interaction between dislocations and dipolar loops. J. Appl. Physics
14. Minárik, V., Kratochvíl, J., Mikula, K., Beneš, M.: Numerical simulation of dislocation dynamics. In: Feistauer, M., Dolejší, V., Knobloch, P., Najzar, K. (eds.) Numerical Mathematics and Advanced Applications – ENUMATH 2003, Springer, New York, pp. 631–641 (2004)
15. Minárik, V., Kratochvíl, J., Mikula, K.: Numerical simulation of dislocation dynamics by means of parametric approach. In: Beneš, M., Mikyška, J., Oberhuber, T. (eds.) Proceedings of the Czech Japanese Seminar in Applied Mathematics, Faculty of Nuclear Sciences and Physical Engineering, Czech Technical University in Prague, Prague, pp. 128–138 (2005)
16. Mura, T.: Micromechanics of defects in solids. Springer, New York (1987)
17. Osher, S., Fedkiw, R.P.: Level set methods and dynamic implicit surfaces. Springer, New York (2003)
18. Pauš, P.: Numerical simulation of dislocation dynamics. In: Vajsáblová, M., Struk, P. (eds.) Proceedings of Slovak-Austrian Congress, Magia, Bratislava, pp. 45–52 (2007)
19. Pauš, P., Beneš, M.: Algorithm for topological changes of parametrically described curves. In: Handlovičová A., Frolkovič P., Mikula K., Ševčovič D. (eds.) Algoritmy 2009, Proceedings of contributed papers and posters Slovak University of Technology in Bratislava, Publishing House of STU, 2009, pp. 176–184 (2009)
20. Pauš, P., Beneš, M.: Direct approach to mean-curvature flow with topological changes, Kybernetik (2009)
21. Sedláček, R.: Viscous glide of a curved dislocation. Philos. Mag. Lett. 76(4), 275–280 (1997)
22. Sethian, J.A. : Level set methods and fast marching methods. Cambridge University Press, Cambridge (1999)
23. Ševčovič, D., Yazaki, S.: On a motion of plane curves with a curvature adjusted tangential velocity. In: http://www.iam.fmph.uniba.sk/institute/sevcovic/papers/cl39.pdf, arXiv:0711.2568, 2007

Non-Symmetric Algebraic Multigrid Preconditioners for the Bidomain Reaction–Diffusion system

Micol Pennacchio and Valeria Simoncini

Abstract We deal with the efficient solution of the so-called *bidomain* system which is possibly the most complete model for the cardiac bioelectric activity. We study the performance of a non-symmetric structured algebraic multigrid (AMG) preconditioner on the formulation generally used of the bidomain model, i.e., the one characterized by a parabolic equation coupled with an elliptic one. Our numerical results show that, for this formulation, the non-symmetric preconditioner provides the best overall performance compared with the AMG based block structured preconditioners developed in [J. Sci. Comput. 36, 391–419 (2008)]. In this paper we provide theoretical justification for the observed optimality.

1 The Bidomain Model

The excitation process in the myocardium is a complex phenomenon characterized by rapid ionic fluxes through the cellular membrane separating the intracellular and the interstitial fluid in the myocardium [8]. The *bidomain* is the most complete model for the cardiac bioelectric activity, and it consists of a non-linear Reaction–Diffusion (R–D) system of equations for the intra- and extracellular potential u_i and u_e, coupled through the transmembrane potential $v := u_i - u_e$ [10]. The nonlinearity arises through the current–voltage relationship across the membrane which is described by a set of nonlinear ODEs, see [8]. The anisotropic properties of the media are modeled by the intra- and extracellular conductivity tensors $M_i = M_i(\mathbf{x})$ and $M_e = M_e(\mathbf{x})$ that satisfy a uniform ellipticity condition [11].

M. Pennacchio (✉)
Istituto di Matematica Applicata e Tecnologie Informatiche del CNR, via Ferrata, 1, I-27100 Pavia, Italy
e-mail: micol@imati.cnr.it

V. Simoncini
Dipartimento di Matematica, Università di Bologna, Piazza di Porta S. Donato, 5, I-40127 Bologna, Italy and IMATI-CNR, Pavia, Italy
e-mail: valeria@dm.unibo.it

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_78,

1.1 (u_e, v) Formulation

The R–D system governing the cardiac electric activity may be written in various forms involving different combinations of the variables u_i, u_e, v; see, e.g., [13]. Here we deal with the formulation generally used for the numerical simulations, i.e., with a parabolic equation for the transmembrane potential v coupled with an elliptic equation for the extracellular potential u_e:
find $(v(\mathbf{x},t), u_e(\mathbf{x},t))$, $\mathbf{x} \in \Omega$, $t \in [0,T]$ *such that*

$$\begin{cases} c_m \partial_t v - \text{div } M_i \nabla v + I_{ion} = \text{div } M_i \nabla u_e + I_{app} & \text{in } \Omega \times]0,T[\\ -\text{div } M \nabla u_e = \text{div } M_i \nabla v & \text{in } \Omega \times]0,T[\\ \mathbf{n}^T M_i \nabla v = 0, \qquad \mathbf{n}^T M \nabla u_e = 0 & \text{on } \Gamma \times]0,T[\\ v(\mathbf{x},0) = 0 & \text{in } \Omega. \end{cases} \tag{1}$$

with $M = M_i + M_e$ bulk conductivity tensor. Due to the presence of different time and space scales, the numerical solution of the bidomain system represents a very intensive computational task: realistic three dimensional simulations typically yield discrete problems with millions of unknowns, and time steps of the order of 10^{-2} ms or less. To reduce the computational cost, different numerical techniques have been developed [4–6, 9, 12, 19]. Here we employ a semi-implicit method in time, that only requires the solution of linear systems at each time step and allows performing larger time steps than explicit schemes. By using a finite element discretization in space and a semi-implicit scheme in time, we get:

$$\mathscr{B}\boldsymbol{\xi}^{k+1} = \mathbf{b} \qquad \text{with} \qquad \mathscr{B} = \begin{bmatrix} C_t + A_i & A_i \\ A_i & (A_i + A_e) \end{bmatrix}, \tag{2}$$

with $\mathbf{b} = \left[C_t\, \mathbf{v}^k - I_{ion}^h(\mathbf{v}^k) + I_{app}^h; \mathbf{0}\right]$, $\mathbf{v}^k = \mathbf{u_i}^k - \mathbf{u_e}^k$, $\boldsymbol{\xi}^{k+1} = \left[\mathbf{v}^{k+1}; \mathbf{u_e}^{k+1}\right]$.

Whatever the method chosen for discretizing the problem, a huge computational effort is required to solve the associated linear system in (2) at each time step, whose conditioning considerably worsens as the problem dimension increases, resulting in an unacceptable increase in the computational costs of the whole simulation. Preconditioning is therefore mandatory. Attempts in literature have employed diagonal preconditioners [17], Symmetric Successive Over Relaxation [11], Block Jacobi preconditioners with incomplete LU factorization (ILU) [20]. General Algebraic Multigrid (AMG) preconditioning has been already applied to the bidomain system and its effectiveness when compared to other classical methods has been reported [1, 14, 21]. However in [13] we verified that the performance of AMG based preconditioner is strictly related to the formulation chosen for the bidomain system and can be improved if the structure of the linear system is exploited. In [13] we verified numerically that the best performance for (2) is obtained by a nonsymmetric preconditioner generally used for saddle point problems. In this paper we provide a theoretical justification for it. A similar preconditioner but built using a simplified version of the bidomain model is studied in [7].

2 Block Preconditioners

In the coefficient matrix $\mathscr{B}$ in (2), the (1,1) block is symmetric positive definite (SPD) while the (2,2) block is only positive semi-definite. Moreover, all matrices are square and symmetric. It is therefore natural to derive preconditioners that exploit this structure. In [13] we analyzed symmetric structured preconditioners and in particular a block diagonal preconditioner $\mathscr{P}_d$ and a block factorized preconditioner $\mathscr{P}_f$:

$$\mathscr{P}_d = \text{blockdiag}(K, D), \qquad \mathscr{P}_f = \begin{bmatrix} I & O \\ A_i K^{-1} & I \end{bmatrix} \begin{bmatrix} K & A_i \\ O & D \end{bmatrix}, \tag{3}$$

where K is an SPD approximation to the (1,1) block, while D is an SPD approximation either to the Schur complement $(A_i + A_e) - A_i (C_t + A_i)^{-1} A_i$, or to the (2,2) block $A_i + A_e$. In [13] we also experimentally verified that more general structured preconditioners may also be appealing. The following "one-sided" version of $\mathscr{P}_f$ is used for symmetric (indefinite) saddle point problems (see, e.g., [2]):

$$\mathscr{P}_M = \begin{bmatrix} K & A_i \\ O & D \end{bmatrix}.$$

If K and D coincide with the (1,1) block and the Schur complement, then

$$\mathscr{B}\mathscr{P}_{M,ex}^{-1} = \begin{bmatrix} I & O \\ A_i (C_t + A_i)^{-1} & I \end{bmatrix},$$

whose spectrum consists of the single unit eigenvalue, so that a minimal residual method such as GMRES ([16]) would converge in at most two iterations. In this case, as well as when $D = A_i + A_e$, we denote the "exact" preconditioner with $\mathscr{P}_{M,ex}$. In general, the behavior of the approximate versions of K and D is less predictable; moreover, a good approximation of the Schur complement may be very expensive to obtain. The performance of $\mathscr{P}_M$ within the indefinite saddle point context highly overcomes its nonsymmetric nature. The situation is considerably different in our context, where the original matrix is positive (semi)definite. Remarkably, however, the use of $\mathscr{P}_M$ in our 2D problem yields some interesting numerical results, cf. [13]. Here we provide an analytical justification of the good performance of $\mathscr{P}_M$. In the following we assume that A_i^{-1} stands for the pseudo-inverse whenever the matrix is singular. Singularity does not effect the analysis as all vectors are assumed to lie in the range of the considered matrices. For $K = A_i + C_t$, $D = A_i + A_e$, it can be easily verified that

$$\mathscr{B}\mathscr{P}_{M,ex}^{-1} = \begin{bmatrix} I & O \\ A_i (A_i + C_t)^{-1} & I - \mathscr{S} \end{bmatrix}, \quad \mathscr{S} = A_i (A_i + C_t)^{-1} A_i (A_i + A_e)^{-1}. \tag{4}$$

The following result shows that the spectrum of $\mathscr{B}\mathscr{P}_M^{-1}$ is bounded independently of the mesh parameter for judiciously chosen D.

Theorem 1. *With the previous notation, let $K = A_i + C_t$ and let $\mathscr{X}$ be an eigenvector matrix of $\mathscr{B}\mathscr{P}_M^{-1}$. If $D = A_i + A_e$, then*

$$\lambda_{min}(\mathscr{B}\mathscr{P}_{M,ex}^{-1}) = 1 - \mu \quad \lambda_{max}(\mathscr{B}\mathscr{P}_{M,ex}^{-1}) = 1, \quad \mathscr{X} = \begin{bmatrix} I & O \\ (A_i + A_e)A_i^{-1} & Y \end{bmatrix},$$

with $\mu \leq (1 + \lambda_{\min}(A_e, A_i))^{-1}$, μ constant independent of h, and Y eigenvector matrix of $I - \mathscr{S}$.

If instead D is such that there exist positive constants α_1, α_2 such that $\alpha_1 \mathbf{x}^T D\mathbf{x} \leq \mathbf{x}^T (A_i + A_e)\mathbf{x} \leq \alpha_2 \mathbf{x}^T D\mathbf{x}$ for all $\mathbf{x}$ in the range of $A_i + A_e$, then either $\lambda(\mathscr{B}\mathscr{P}_M^{-1}) = 1$ or $\alpha_1(1 - \mu) \leq \lambda(\mathscr{B}\mathscr{P}_M^{-1}) \leq \alpha_2$, with μ defined above. Moreover,

$$\mathscr{X} = \begin{bmatrix} I & O \\ M & Y \end{bmatrix},$$

with $M = -(G - I)^{-1} A_i (A_i + C_t)^{-1}$, $G = (A_i + A_e - A_i(C_t + A_i)^{-1} A_i) D^{-1}$ and Y eigenvector matrix of G.

Proof. We shall see that the eigenvalues of $\mathscr{S}$ are real and non-negative. From the structure of the matrix $\mathscr{B}\mathscr{P}_{M,ex}^{-1}$ it thus follows that $\lambda_{min}(\mathscr{B}\mathscr{P}_{M,ex}^{-1}) = 1 - \lambda_{max}(\mathscr{S})$ and $\lambda_{max}(\mathscr{B}\mathscr{P}_{M,ex}^{-1}) = 1$. To analyze the eigenvalues λ of $\mathscr{S}$ we consider the eigenvalue problem $A_i(A_i + C_t)^{-1} A_i (A_i + A_e)^{-1}\mathbf{x} = \lambda\mathbf{x}$, that is $A_i(A_i + C_t)^{-1} A_i \mathbf{u} = \lambda(A_i + A_e)\mathbf{u}$. Clearly, $\lambda = 0$ for $\mathbf{u} \in N(A_i) = N(A_e) = N(A_i + A_e)$. Moreover, since the matrices on both sides are SPD in the range of A_i, A_e, $\lambda \geq 0$. We can write

$$\lambda = \frac{\mathbf{u}^T A_i (A_i + C_t)^{-1} A_i \mathbf{u}}{\mathbf{u}^T (A_i + A_e)\mathbf{u}}, \quad \mathbf{u} \notin N(A_i).$$

Thanks to [13, Lemma 4.1] we obtain that $\lambda \leq (1 + \lambda_{\min}(A_e, A_i))^{-1} < 1$. Using the conductivity coefficients defined in Sect. 5 of [13], the two stiffness matrices are related as $c_1 \mathbf{v}^T A_e \mathbf{v} \leq \mathbf{v}^T A_i \mathbf{v} \leq c_2 \mathbf{v}^T A_e \mathbf{v}$ independently of the mesh. Thus, $\lambda_{\min}(A_e, A_i)$ is bounded by a quantity that only depends on the conductivity tensors of the two stiffness matrices, and not on the grid. This completes the proof for $D = A_i + A_e$. One can readily verify that $\mathscr{X}$ satisfies $\mathscr{B}\mathscr{P}_{M,ex}^{-1}\mathscr{X} = \mathscr{X}\,\text{blockdiag}(I, I - \Lambda)$, where Λ is the eigenvalues matrix of $I - \mathscr{S}$.

For general D we have

$$\mathscr{B}\mathscr{P}_M^{-1} = \begin{bmatrix} I & O \\ A_i(A_i + C_t)^{-1} & I \end{bmatrix} \begin{bmatrix} I & O \\ O & (A_i + A_e - A_i(C_t + A_i)^{-1} A_i) D^{-1} \end{bmatrix}.$$

The eigenvalues θ's of the (2,2) block in the second factor satisfy

$$\theta = \frac{\mathbf{x}^T(A_i + A_e - A_i(C_t + A_i)^{-1}A_i)\mathbf{x}}{\mathbf{x}^T D\mathbf{x}} = \frac{\mathbf{x}^T(A_i + A_e)\mathbf{x}}{\mathbf{x}^T D\mathbf{x}}\left(1 - \frac{\mathbf{x}^T A_i(C_t + A_i)^{-1}A_i\mathbf{x}}{\mathbf{x}^T(A_i + A_e)\mathbf{x}}\right) = \gamma_1\gamma_2.$$

Using the spectral equivalence of D, we have $\alpha_1 \leq \gamma_1 \leq \alpha_2$. Moreover, using the definition of μ above, $(1-\mu) \leq \gamma_2 \leq 1$, from which the result follows. The fact that the given $\mathscr{X}$ is an eigenvector matrix can be readily verified.

We observe that due to the matrix structure, we expect cond($\mathscr{X}$) to be mesh independent in the exact case ($K = A_i + C_t$, $D = A_i + A_e$). A spectral analysis when K is an approximation to the (1,1) block, that is it is not exact, is much more involved. With a convenient splitting of $\mathscr{B}$, we write

$$\begin{aligned}\mathscr{B}\mathscr{P}_M^{-1} &= \left(\begin{bmatrix} K & A_i \\ A_i & D \end{bmatrix} + \begin{bmatrix} C_t + A_i - K & O \\ O & (A_i + A_e) - D \end{bmatrix}\right)\mathscr{P}_M^{-1} \\ &= \begin{bmatrix} I & O \\ A_i K^{-1} & I - A_i K^{-1} A_i D^{-1} \end{bmatrix} + \\ &+ \begin{bmatrix} (C_t + A_i - K)K^{-1} & -(C_t + A_i - K)K^{-1}A_i D^{-1} \\ O & (A_i + A_e - D)D^{-1} \end{bmatrix} \equiv R + E.\end{aligned}$$

If K, D are spectrally equivalent to $A_i + C_t$ and $A_i + A_e$ respectively, the spectrum of R is also spectrally equivalent to that of the exactly preconditioned matrix $\mathscr{B}\mathscr{P}_{M,ex}^{-1}$ (for $K = A_i + C_t$, $D = A_i + A_e$). The matrix E represents a perturbation to the ideal case, and its size depends on the accuracy of the preconditioning blocks.

Assume that all relevant[1] eigenvalues of the (2,2) block of R are less than one, and let X be an eigenvector matrix of R; in fact, it is possible to derive a more explicit structure for X, but such a description is beyond the scope of this paper. Then we have (cf, e.g., [18])

$$|\lambda(\mathscr{B}\mathscr{P}_M^{-1}) - \lambda(R)| \leq \|X^{-1}EX\| \leq \text{cond}(X)\|E\|,$$

where cond(X) is the spectral condition number of X and $\|\cdot\|$ is the matrix norm induced by the Euclidean vector norm. Therefore, if K and D are good approximations to the corresponding blocks, then we expect the spectrum of $\mathscr{B}\mathscr{P}_M^{-1}$ not to deviate significantly from that of R, unless the eigenvector matrix X is very ill conditioned. If the condition number of the eigenvector matrix of $\mathscr{B}\mathscr{P}_M^{-1}$ is moderate (cf. Th. 1), we also expect that a nonsymmetric solver like GMRES will converge in approximately the same number of iterations as for the exact case; if K, D are

[1] That is, those associated to eigenvectors in the range of the given matrices.

Table 1 CPU time and number of iterations (in parenthesis) for: $\mathscr{P}_f$ with AMG; $\mathscr{P}_M$ with exact blocks ($\mathscr{P}_{M,ex}$) and AMG-based blocks, when using the nonsymmetric solver GMRES, FOM and the symmetric solver CG (with regularization). Here K= AMG($C_t + A_i$) and D= AMG($A_i + A_e$)

n	$\mathscr{P}_f$	$\mathscr{P}_{M,ex}$		AMG-based $\mathscr{P}_M$		
		CG	GMRES	GMRES	FOM	CG
2,705	0.41 (6)	1.1 (21)	0.38 (6)	0.15 (6)	0.27 (6)	0.32 (22)
10,657	0.88 (7)	2.98 (13)	1.95 (7)	0.5 (7)	0.52 (7)	0.64 (12)
42,305	2.82 (8)	11.85 (11)	8.92 (7)	2.15 (8)	2.17 (8)	2.07 (10)
168,577	9.92 (8)	58.11 (11)	44.83 (7)	8.99 (8)	9.06 (8)	8.27 (10)
673,025	47.47 (10)	315.49 (10)	249.95 (7)	41.02 (9)	41.32 (9)	36.21 (11)

chosen to be spectrally equivalent to the corresponding matrices, then we also expect mesh independence, as with $\mathscr{B}\mathscr{P}_{M,ex}^{-1}$. Our numerical results (cf. Table 1) confirm these considerations.

3 Numerical Results

In this section we report on our experiments with the exact and "inexact" versions of the block triangular preconditioner $\mathscr{P}_M$. We consider a square domain $\Omega = [0, 1]^2$ modeling a block of myocardium with cardiac fibers parallel to a diagonal of the square and the conductivity coefficients defined as in [13]. The meshes on Ω were built by using a Delaunay triangulation algorithm. The number of mesh nodes for each refined grid was $2n$ with $n \in \{2705, 10657, 42305, 168577, 673025\}$ and the time step $\tau = 4.10^{-2}$ ms. All experiments correspond to a typical temporal instant in the time step evolution, so that the right–hand side includes information generated during the previous time steps. All computations were performed with Matlab 7.4.0 (R2007a) on a iMac Intel Core 2 Duo 2 GbRAM 2.66 GHz and 6Mb L2 cache.

In the approximate (inexact) case, the matrices K and D are implicitly defined by applying an AMG preconditioner at each iteration to approximate the corresponding blocks; the preconditioner is built once for all at the beginning. As in [13], we reorder each block matrix of $\mathscr{B}$ by using the matlab function `symrcm`. We employ the AMG code available in the HSL library, the HSL_MI20 routine, equipped with a Matlab interface [3]. This function implements the classical (Ruge-Stüben) AMG method, as described in [15]. The code was used as a black box: Gauss–Seidel smoothing was used in all instances. The multilevel method is often built on originally singular matrices. To increase the robustness of the preconditioning strategy, in some cases we generated the preconditioner by using a shifted (nonsingular) matrix, with a shift equal to $\varepsilon^{1/2}$ and $\varepsilon \approx 10^{-16}$ the Matlab machine precision. Table 1 reports the results for the considered discretization meshes: CPU times (in seconds) and in parenthesis number of iterations are shown. A stopping tolerance of 10^{-6} was used for the residual norm. In the second column we recall the performance of the best performing preconditioner in [13] for this formulation, namely P_f with AMG for computing K and D. The subsequent two columns show the performance of

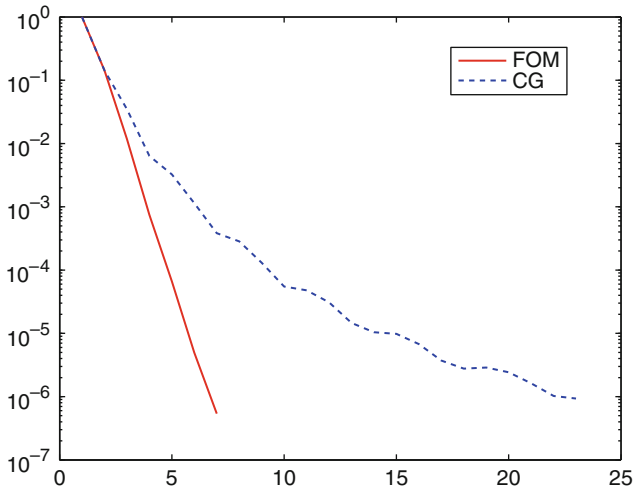


Fig. 1 Convergence history of FOM and CG on the nonsymmetric matrix $\mathscr{B}\mathscr{P}_M^{-1}$, $n = 2{,}705$

$\mathscr{P}_{M,ex}$ (with $K = A_i + C_t$ and $D = A_i + A_e$), when using either the nonsymmetric solver GMRES or CG (see below for further comments on the latter method). The last columns show the performance of $\mathscr{P}_M$ when AMG is used to build K and D as approximations to $A_i + C_t$ and $A_i + A_e$, respectively. Note that a sparse direct solver employs 26.15 s, after a proper reordering, to solve the whole system in (2) for $n = 168{,}577$ (an "out of memory" results for $n = 673{,}025$). The reported timings clearly confirm the competitiveness of the AMG-based preconditioner compared to the exact version and the direct solver. In particular, results are reported for the minimal residual method GMRES, and for both FOM (Full Orthogonalization Method) and CG [16]. If the preconditioned problem were symmetric and positive definite, then FOM and CG would be mathematically equivalent. Since $\mathscr{B}\mathscr{P}_M^{-1}$ is nonsymmetric, we expect CG to behave more poorly than FOM, which is a Galerkin-type method devised for nonsymmetric problems. It is however quite surprising that CG converges very quickly in spite of the full nonsymmetry of the problem. In Fig. 1 we display the convergence history of the methods, in terms of residual norms: the CG curve deviates from the expected one, represented by FOM, as soon as nonsymmetry is detected. However, the spectral properties are so favourable that nonsymmetry does not prevent the method from converging in just a few more iterations. In fact, CG may be viewed in this case as a (highly) truncated full orthogonalization procedure; cf. [13]. Due to the very cheap short-term recurrence, the CG timings are also very competitive; cf. the last column of Table 1.

In the exact case, our theory predicts mesh independence, and this is confirmed in the table. In addition, the use of AMG preconditioning maintains mesh independence in the inexact case, with in general a number of iterations only slightly higher than in the exact case.

References

1. T.M. Austin, M.L. Trew, A. Pullan. Solving the cardiac bidomain equations for discontinuous coefficients. IEEE Trans. Biomed. Eng. **53**, 1265–1272 (2006)
2. M. Benzi, G.H. Golub, J. Liesen. Numerical solution of saddle point problems. Acta Numerica **14**, 1–137 (2005)
3. J. Boyle, M.D. Mihajlović, J.A. Scott. HSL_MI20: an efficient AMG preconditioner. RAL Tech.Rep. RAL-TR-2007-021 (2007)
4. E. Cherry, H. Greenside, C.S. Henriquez. Efficient simulation of three-dimensional anisotropic cardiac tissue using an adaptive mesh refinement method. Chaos **13**, 853–865 (2003)
5. P. Colli Franzone, M. Pennacchio, L. Guerri. Accurate computation of electrograms in the left ventricular wall. Math. Mod. Meth. Appl. Sci.**10**, 507–538 (2000)
6. P. Colli Franzone, P. Deuflhard, B. Erdmann, J. Lang, L.F. Pavarino. Adaptivity in space and time for reaction–diffusion systems in electrocardiology. SIAM J. Sci. Comput. **28**, 942–962 (2006)
7. L. Gerardo Giorda, L. Mirabella, F. Nobile, M. Perego, A. Veneziani. A model-based block-triangular preconditioner for the bidomain system in electrocardiology. J. Comput. Phys. **228**, 3625–3639 (2009)
8. J. Keener, J. Sneyd. Mathematical physiology. Springer, New York (2009)
9. M. Pennacchio. The mortar finite element method for the cardiac "bidomain" model of extracellular potential. J. Sci. Comput. **20**, 191–210 (2004)
10. M. Pennacchio, G. Savaré, P. Colli Franzone. Multiscale modeling for the bioelectric activity of the heart. SIAM J. Math. Anal. **37**, 1333–1370 (2006)
11. M. Pennacchio, V. Simoncini. Efficient algebraic solution of reaction–diffusion systems for the cardiac excitation process. J. Comput. Appl. Math. **145**, 49–70 (2002)
12. M. Pennacchio, V. Simoncini. Substructuring preconditioners for mortar discretization of a degenerate evolution problem. J. Sci. Comput. **36**, 391–419 (2008)
13. M. Pennacchio, V. Simoncini. Algebraic multigrid preconditioners for the bidomain reaction–diffusion system. Appl. Numer. Math. **59**, 3033–3050 (2009)
14. G. Plank, M. Liebmann, R. Weber dos Santos, E.J. Vigmond, G. Haase. Algebraic multigrid preconditioner for the cardiac bidomain model. IEEE Trans. Biomed. Eng. **54**, 585–596 (2007)
15. J.W. Ruge, K. Stüben. Algebraic multigrid. In S.F. McCormick, editor, Multigrid Methods, v.3 of Frontiers in applied mathematics, pp. 73–130, SIAM (1987)
16. Y. Saad. Iterative methods for sparse linear systems. The PWS Publishing Company (1996)
17. K. Skouibine, W. Krassowska. Increasing the computational efficiency of a bidomain model of defibrillation using a time-dependent activating function. Ann. Biomed. Eng. **28**, 772–780 (2000)
18. G.W. Stewart, J-G. Sun. Matrix Perturbation Theory. Academic Press (1990)
19. J. Trangestein, C. Kim. Operator splitting and adaptive mesh refinement for the Luo–Rudy I model. J. Comput. Phys. **196**, 645–679 (2004)
20. E.J. Vigmond, F. Aguel, N.A. Trayanova. Computational techniques for solving the bidomain equations in three dimensions. IEEE Trans. Biomed. Eng. **49**, 1260–1269 (2002)
21. E.J. Vigmond, R. Weber dos Santos, A.J. Prassl, M. Deo, G. Plank. Solvers for the cardiac bidomain equations. Progr. Biophys. Mol. Biol. **96**, 3–18 (2008)

Efficiency of Shock Capturing Schemes for Burgers' Equation with Boundary Uncertainty

Per Pettersson, Qaisar Abbas, Gianluca Iaccarino, and Jan Nordström

Abstract Burgers' equation with uncertain initial and boundary conditions is approximated using a polynomial chaos expansion approach where the solution is represented as a series of stochastic, orthogonal polynomials. Even though the analytical solution is smooth, a number of discontinuities emerge in the truncated system. The solution is highly sensitive to the propagation speed of these discontinuities. High-resolution schemes are needed to accurately capture the behavior of the solution. The emergence of different scales of the chaos modes require dissipation operators to yield accurate solutions. We will compare the results using the MUSCL scheme with previously obtained results using conventional one-sided operators.

P. Pettersson (✉)
Department of Mechanical Engineering, Stanford University, 488 Escondido Mall, Stanford, CA 94305, USA
and
Department of Information Technology, Scientific Computing, Uppsala University, SE-751 05 Uppsala, Sweden
e-mail: massperp@stanford.edu

Q. Abbas
Department of Information Technology, Scientific Computing, Uppsala University, SE-751 05 Uppsala, Sweden
e-mail: qaisar.abbas@it.uu.se

G. Iaccarino
Department of Mechanical Engineering, Stanford University, 488 Escondido Mall, Stanford, CA 94305, USA
e-mail: jops@stanford.edu

J. Nordström
Department of Information Technology, Scientific Computing, Uppsala University, SE-751 05 Uppsala, Sweden
and
School of Mechanical, Industrial and Aeronautical Engineering,University of the Witvatersrand, PO WITS 2050, Johannesburg, South Africa
and
Department of Aeronautics and Systems Integration, FOI, The Swedish Defense Research Agency, SE-164 90 Stockholm, Sweden
e-mail: jan.nordstrom@it.uu.se

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_79,

1 Introduction

The inviscid Burgers' equation is investigated subject to uncertain boundary and initial conditions. The stochastic solution is represented as a polynomial chaos series, using a suitable basis of orthogonal stochastic polynomials. The stochastic Galerkin method [3] is applied and the stochastic equation is projected onto a stochastic polynomial basis yielding a system of deterministic equations for the time and space dependent coefficients of the series. The resulting system is hyperbolic and exhibit multiple discontinuities in finite time. This motivates the use of high-resolution schemes.

We investigate two different approaches. First, a central scheme with artificial local dissipation is investigated. Summation by parts operators (SBP) [2] and the simultaneous approximation term (SAT) technique [1] to impose boundary conditions weakly lead to stability. The amount of artificial dissipation should be proportional to the system eigenvalues which are generally unknown a priori. Second, we study the monotone upstream centered schemes for conservation laws (MUSCL) approach originally developed by van Leer [4]. We use the minmod limiter and Roe averages to approximate the fluxes, see [5].

Both types of schemes exhibit excellent properties of shock capturing for model problems such as the scalar Burgers' equation. However, the hyperbolic systems resulting from the stochastic Galerkin projection are considerably more demanding in several ways. The non-linearity of the problem results in poor convergence properties independent of the numerical method. Also, finer grids are needed for convergence of higher order polynomial chaos systems, increasing the computational cost.

The paper is organized as follows. The systems of equations is derived in Sect. 2, followed by an outline of the numerical methods in Sect. 3. Section 4 contains numerical experiments where the efficiency of the numerical methods are investigated. Finally, Sect. 5 contains a discussion and concluding remarks.

2 Polynomial Chaos Approximation of Burgers' Equation

The solution of a partial differential equation characterized by input uncertainty (e.g., uncertain boundary data) is expressed as the polynomial chaos expansion

$$u(x,t,\xi) = \sum_{i=0}^{\infty} u_i(x,t)\Psi_i(\xi), \tag{1}$$

where Ψ_i denotes orthogonal basis polynomials of a stochastic variable ξ. For optimal convergence in the L^2-sense, the distribution of ξ should reflect the input uncertainty. In practice, the infinite series (1), which is convergent for second order random fields (i.e., random fields with finite variance), is truncated to a finite number of terms. Polynomial chaos of order M denotes the truncated polynomial chaos series where only the first $M + 1$ terms are used.

(1) is inserted into the inviscid Burgers' equation

$$u_t + uu_x = 0,\ 0 \le x \le 1 \tag{2}$$

which yields

$$\sum_{i=0}^{\infty} \frac{\partial u_i}{\partial t}\Psi_i(\xi) + \left(\sum_{j=0}^{\infty} u_j \Psi_j(\xi)\right)\left(\sum_{i=0}^{\infty} \frac{\partial u_i}{\partial x}\Psi_i(\xi)\right) = 0. \tag{3}$$

A stochastic Galerkin projection is performed by truncating the polynomial chaos expansion to order M, multiplying (3) by $\Psi_k(\xi)$ for non-negative integers $k = 0, \ldots, M$ and integrating over the probability domain. The orthogonality of the basis polynomials (Ψ_i) then yields a system of deterministic equations. The result is a symmetric system of deterministic equations,

$$\frac{\partial u_k}{\partial t}\langle \Psi_k^2 \rangle + \sum_{i=0}^{M}\sum_{j=0}^{M} u_i \frac{\partial u_j}{\partial x}\langle \Psi_i \Psi_j \Psi_k \rangle = 0 \qquad \text{for } k = 0, 1, \ldots, M. \tag{4}$$

To simplify notation, (4) can be written in conservative matrix form as

$$Bu_t + \frac{1}{2}\frac{\partial}{\partial x}(A(u)u) = 0 \tag{5}$$

where $(B)_{jk} = \delta_{jk}\langle \Psi_j^2 \rangle$ and $(A(u))_{jk} = \sum_{i=0}^{M} u_i \langle \Psi_i \Psi_j \Psi_k \rangle$. The polynomial basis is chosen to be the set of Hermite polynomials.

We consider the Riemann problem with uncertainty of the left and right states described by the addition of a Gaussian perturbation

$$u(x, t = 0, \xi) = \begin{cases} 1 + \hat{\sigma}\xi & \text{for } x < 0.5 \\ -1 + \hat{\sigma} & \text{for } x > 0.5 \end{cases}$$

and use the time dependent analytical solution for the Cauchy problem derived in [7] with $\hat{\sigma} = 0.1$.

3 Numerical Methods

3.1 Central Differences

We approximate the first derivative with a matrix operator of the form $P^{-1}Q$ where P is a positive diagonal matrix and Q satisfies $Q + Q^T = \text{diag}(-1, 0, \ldots, 0, 1)$. For a more detailed description of the SBP technique, see [6, 9].

The system (5) is semi-discretized as

$$(I \otimes B)u_t + \frac{1}{2}(P^{-1}Q \otimes I)A_g u = (P^{-1} \otimes I)\left[(E_0 \otimes \Sigma_0)(u - g_0) + (E_n \otimes \Sigma_1)(u - g_1)\right]. \tag{6}$$

A split approach is used to show stability [8] and artificial dissipation [7] is added in the form

$$A_{2k} = -\Delta x P^{-1} \widetilde{D}_k^T B_w \widetilde{D}_k, \tag{7}$$

where $\widetilde{D}_k$ is an approximation of $(\Delta x)^k \partial^k / \partial x^k$ and B_w is a diagonal positive definite matrix.

3.2 MUSCL Scheme

The semi-discrete system is given by

$$(u_t)_i + \frac{F_{i+\frac{1}{2}} - F_{i-\frac{1}{2}}}{\Delta x} = 0 \tag{8}$$

where

$$F_{i+\frac{1}{2}} = \frac{1}{2}\left(F(u^L_{i+\frac{1}{2}}) + F(u^R_{i+\frac{1}{2}})\right) + \frac{1}{2}|A_{i+\frac{1}{2}}|\left(u^L_{i+\frac{1}{2}} - u^R_{i+\frac{1}{2}}\right) \tag{9}$$

with the absolute value of the Roe average $A_{i+\frac{1}{2}}$ given by

$$|A_{i+\frac{1}{2}}| = X\left|\Lambda(u_{i+\frac{1}{2}})\right|X^{-1} = \frac{1}{2}X\left|\Lambda(u^L_{i+\frac{1}{2}}) + \Lambda(u^R_{i+\frac{1}{2}})\right|X^{-1}. \tag{10}$$

where $\Lambda(u)$ is a diagonal matrix with the eigenvalues of $A(u)$ and X is the eigenvector matrix. The left and right states are given by

$$u^L_{i+\frac{1}{2}} = u_i + 0.5\phi(r_i)(u_{i+1} - u_i) \quad \text{and} \quad u^R_{i+\frac{1}{2}} = u_{i+1} - 0.5\phi(r_{i+1})(u_{i+2} - u_{i+1})$$

respectively. The flux limiter $\phi(r)$ is the minmod limiter. For a more detailed description of the MUSCL scheme, see e.g., [5].

4 Numerical Experiments

The true solution of the original problem (3) is qualitatively different from the solution of any truncated system after stochastic Galerkin projection. A fair measure of the efficiency of the numerical method should take this into account. The reference

Table 1 Grid convergence, $M = 1$, $T = 0.3$

m	Central SBP		MUSCL	
	$\|\varepsilon_{u_0}\|_h$	$\|\varepsilon_{u_1}\|_h$	$\|\varepsilon_{u_0}\|_h$	$\|\varepsilon_{u_1}\|_h$
51	0.1207	0.1871	0.1309	0.2100
101	0.0856	0.1330	0.0934	0.1498
201	0.0607	0.0943	0.0664	0.1067
401	0.0430	0.0667	0.0479	0.0764
801	0.0304	0.0472	0.0358	0.0554

Table 2 Grid convergence, $M = 3$, $T = 0.3$

m	Central SBP				MUSCL			
	$\|\varepsilon_{u_0}\|_h$	$\|\varepsilon_{u_1}\|_h$	$\|\varepsilon_{u_2}\|_h$	$\|\varepsilon_{u_3}\|_h$	$\|\varepsilon_{u_0}\|_h$	$\|\varepsilon_{u_1}\|_h$	$\|\varepsilon_{u_2}\|_h$	$\|\varepsilon_{u_3}\|_h$
51	0.0928	0.1001	0.1487	0.0557	0.0899	0.0950	0.1291	0.0569
101	0.0766	0.0740	0.1259	0.0510	0.0652	0.0754	0.0819	0.0336
201	0.0594	0.0551	0.0980	0.0423	0.0399	0.0469	0.0566	0.0240
401	0.0371	0.0318	0.0626	0.0281	0.0183	0.0157	0.0292	0.0128

solution used for the first order polynomial chaos is therefore the analytical solution to the system (4) with $M = 1$. The discrete error norm used for the coefficient u_i is given by

$$\|\varepsilon_{u_i}\|_h = \sqrt{\frac{\sum_{j=1}^{m}((u_i^{num})_j - (u_i^{ref})_j)^2}{m-1}}.$$

Tables 1 and 2 show the grid convergence up to third order polynomial chaos. For expansions of higher order ($M > 1$), a solution on a fine mesh ($m = 800$) is used as reference solution, since the analytical solution to the truncated system is unknown.

The number of shocks increase with the order of polynomial chaos, and requires a finer spatial mesh for convergence. As the order of polynomial chaos is increased, we expect a more accurate approximation to the original problem, before truncation of the polynomial chaos expansion. However, as shown in Table 3, the convergence is not monotone. Also, the computational cost strongly increases with the order of polynomial chaos. Accordingly, high order expansions are not necessarily desirable for these problems.

In order to understand and remedy these issues, consider the coarse grid solution of Fig. 1, where the SBP solution with artificial dissipation appears to be a more accurate approximation of the true analytical solution than the MUSCL solution. By scaling the i^{th} Hermite polynomial by $1/\sqrt{i!}$ of the MUSCL scheme, the solution approaches the SBP solution on an equal grid, Fig. 2. However, unlike the scalar cases and lowest order of polynomial chaos, this solution is not grid converged. Since these numerical solutions are solutions to the *truncated* system, we can not evaluate the efficiency of the numerical methods by comparison with the *true analytical* solutions only.

With $m = 400$ mesh points, the solutions are grid converged, but different due to the non-sharp eigenvalue estimate of the SBP approach which modifies the artificial

Table 3 PC expansion convergence, $m = 100$, $T = 0.3$

M	Central SBP		MUSCL	
	$\|\varepsilon_{\text{Exp}}\|_h$	$\|\varepsilon_{\text{Var}}\|_h$	$\|\varepsilon_{\text{Exp}}\|_h$	$\|\varepsilon_{\text{Var}}\|_h$
1	0.1428	0.1897	0.1518	0.2007
2	0.1336	0.1884	0.1614	0.1965
3	0.0263	0.0854	0.0742	0.2102
4	0.0330	0.0758	0.1079	0.2327
5	0.0198	0.0407	0.0783	0.1887
6	0.0176	0.0529	0.0733	0.1215
7	0.0267	0.0813	0.0694	0.1419

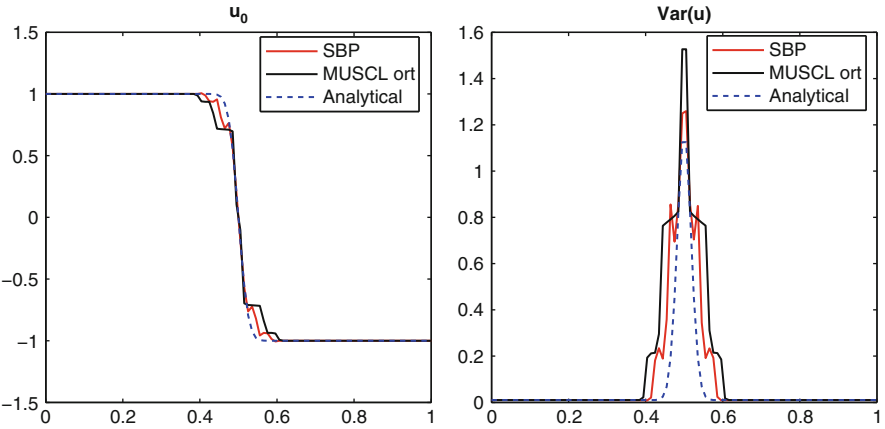


Fig. 1 Comparison SBP and MUSCL. Locally weighted dissipation. $M = 3, m = 100, T = 0.2$. Orthonormal Hermite polynomials

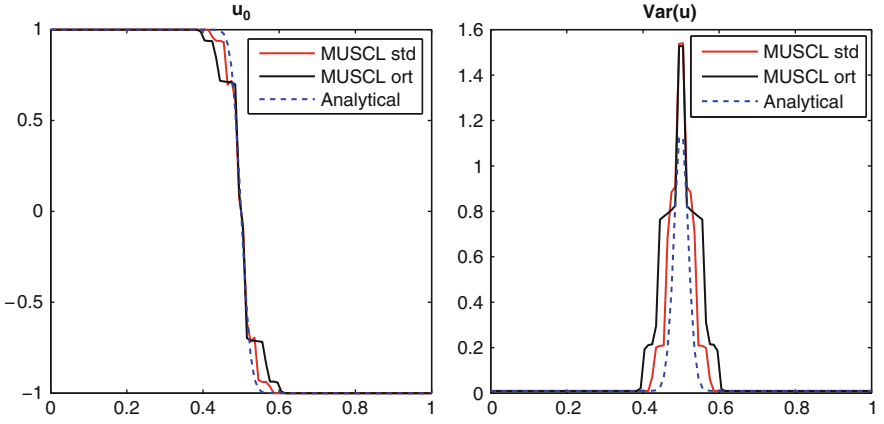


Fig. 2 $M = 3$, $m = 100$, $T = 0.2$. The label "Std" denotes a solution based on the standard probabilistic Hermite polynomials and "ort" denotes the scaling to make these polynomials orthonormal

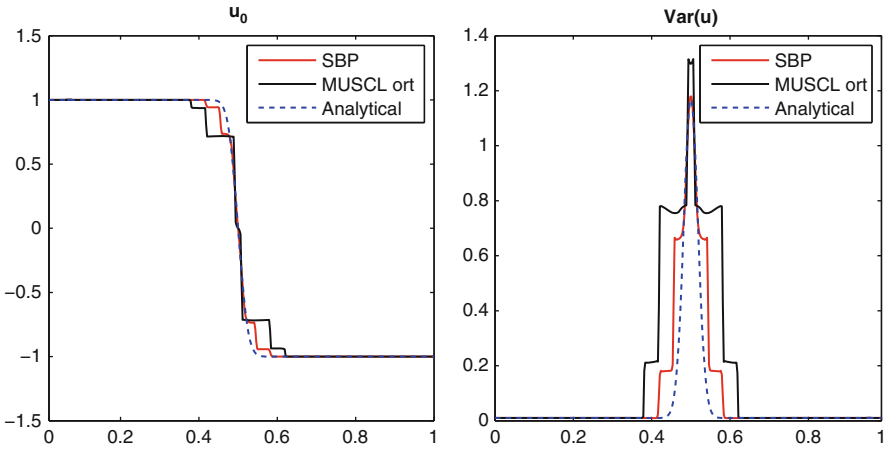


Fig. 3 Comparison SBP and MUSCL. Dissipation based on largest eigenvalues. $M = 3$, $m = 400$, $T = 0.2$

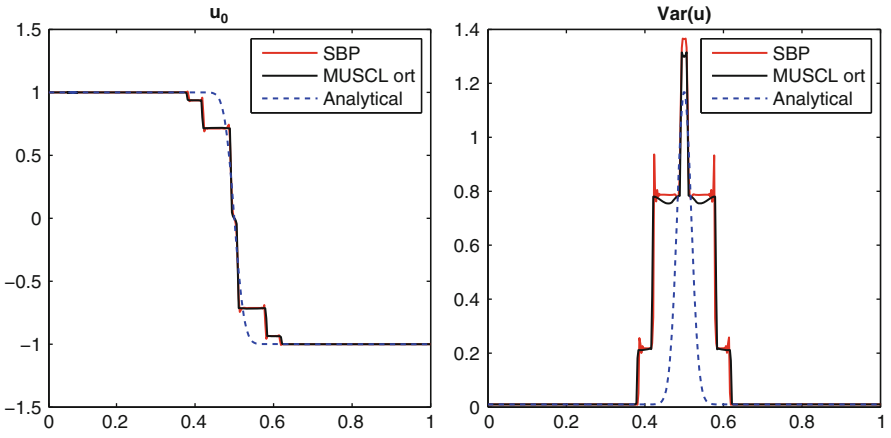


Fig. 4 Comparison SBP and MUSCL. Locally weighted dissipation. $M = 3$, $m = 400$, $T = 0.2$

dissipation and the solution, Fig. 3. By successively decreasing the amount of artificial dissipation to the point where the solution fails to converge, the SBP solution approaches the MUSCL solution, Fig. 4. The two different scalings of the basis polynomials also result in the same solution as the mesh is refined (Fig. 5).

This illustrates the fact that excessive use of artificial dissipation on a coarse mesh might appear to provide a more accurate solution to the original problem than the most accurate solution to the truncated system, given by the MUSCL scheme. This suggests that the effect of the truncation of the polynomial chaos expansion should be taken into account in the solution method.

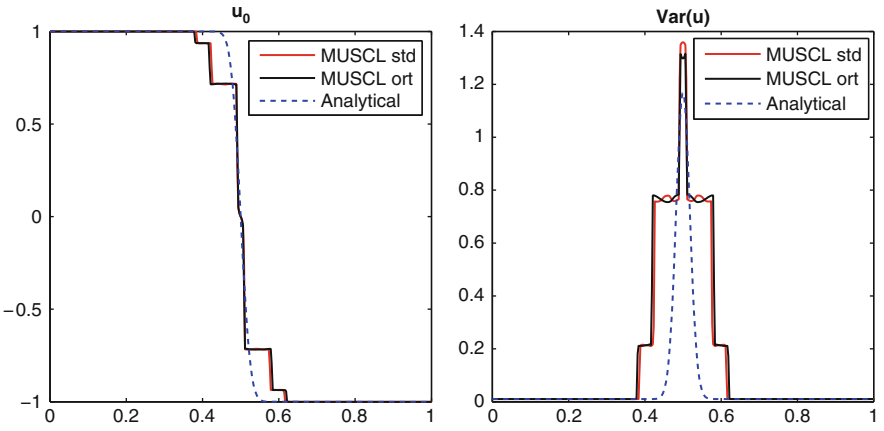


Fig. 5 $M = 3$, $m = 400$, $T = 0.2$. Standard basis and orthogonal Hermite basis

5 Conclusions

Compared to the classical deterministic Burgers' equation, where shocks are accurately captured by both the SBP method and the MUSCL scheme, high order polynomial chaos systems representing uncertain Burgers' equation are very sensitive to the choice of numerical method. The inaccurate scaling of the artificial dissipation for the central summation by parts operators often result in either oscillatory and eventually unstable schemes or inaccurate schemes with poor shock capturing properties. Therefore, the MUSCL scheme seems to be a more suitable choice of numerical method for these problems. However, the increasing number of shocks that result from higher order polynomial chaos requires finer grids and are therefore computationally expensive. Even with the MUSCL scheme, a different scaling of the basis polynomials affects the grid convergence.

References

1. Carpenter, M.H., Gottlieb, D., Abarbanel, S.: Time-stable boundary conditions for finite-difference schemes solving hyperbolic systems: methodology and application to high-order compact schemes. Journal of Computational Physics. **111**, 220–236 (1994)
2. Carpenter, M.H., Nordström, J., Gottlieb, D.: A stable and conservative interface treatment of arbitrary spatial accuracy. Journal of Computational Physics. **148**, 341–365 (1999)
3. Ghanem, R., Spanos, P.: Stochastic finite elements: A spectral approach, Springer, New York (1991)
4. van Leer, B.: Towards the ultimate conservative difference scheme. Journal of Computational Physics. **135**, 229–248 (1997)
5. LeVeque, R.J.: Finite volume methods for hyperbolic problems, Cambridge University Press, Cambridge (2002)

6. Mattsson, K., Svärd, M., Nordström, J.: Stable and accurate artificial dissipation. Journal of Scientific Computing. **21**, 57–79 (2004)
7. Pettersson, P., Iaccarino, G., Nordström, J.: Numerical analysis of the Burgers' equation in the presence of uncertainty. Journal of Computational Physics, 2009, DOI 10.1016/j.jcp.2009.08.012
8. Richtmyer, R.D., Morton, K.W., Difference methods for initial-value problems, second ed., Interscience Publishers, New York (1967)
9. Strand, B.: Summation by parts for finite difference approximations for d/dx. Journal of Computational Physics. **110**, 47–67 (1994)

FEM Techniques for the LCR Reformulation of Viscoelastic Flow Problems

A. Ouazzi, H. Damanik, J. Hron, and S. Turek

Abstract We present special numerical techniques for viscoelastic fluid flow utilizing a fully coupled monolithic multigrid finite element approach with consistent edge-oriented stabilization technique. The governing equations arise from the Navier–Stokes for the Oldroyd-B type of fluid with the help of the log-conformation reformulation to allow a wide range of Weissenberg numbers. The resulting nonlinear system consists of 6 variables for velocity, pressure and the logarithm of the conformation stress tensor in 2D. The system is discretized in time by using a fully implicit second order accurate time integrator. In each time step, we have to solve a discretized system in space employing the high order finite element triple $Q_2/P_1^{disc}/Q_2$. We utilize the discrete damped Newton method with divided differences for handling the Jacobian, and apply a geometrical multigrid solver with a special Vanka smoother to handle the linear subproblems. Local refinement can be assigned at regions of interest to reduce the computational cost. The presented methodology is implemented on the open source software package FEATFLOW (www.featflow.de) and validated for several well-known benchmark problems.

1 Introduction

The numerical simulation of polymer processing problems incorporates the most important characteristics of viscoelastic fluids. Various nonlinear differential models exist to describe their behavior, but all represent the same numerical challenges, namely the strong coupling between the velocity gradient and the elastic stress which leads to a restriction for the choice of FEM approximation spaces, besides

A. Ouazzi (✉), H. Damanik, and S. Turek
Institute for Applied Mathematics, TU Dortmund, 44227 Dortmund, Germany
e-mail: Abderrahim.Ouazzi@math.tu-dortmund.de, Hogenrich.Damanik@math.tu-dortmund.de, ture@featflow.de

J. Hron
Institute of Mathematics, Charles University, Czech Republic
e-mail: hron@karlin.mff.cuni.cz

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_80,

their hyperbolic nature which makes the numerical solution difficult. In this paper, we restrict to the Odroyd-B model, for testing the monolithic FEM approach [4].

For the Oldroyd-B model [5], the computational rheologist introduces the conformation tensor, which has the special property to be positive definite:

$$\boldsymbol{\sigma}^c = \frac{\eta_p}{\mathsf{We}}(\boldsymbol{\sigma}^p - \mathsf{I}) \tag{1}$$

It is worth to note that this tensor has an integral form with exponential expression

$$\boldsymbol{\sigma}^c(t) = \int_{\infty}^{t} \frac{1}{\mathsf{We}} \exp\left(\frac{-(t-s)}{\mathsf{We}}\right) F(s,t)F(s,t)^T ds \tag{2}$$

where $F(s,t)$ is the relative deformation gradient. Then, the set of full equations can be written as

$$\begin{cases} \rho\left(\dfrac{\partial}{\partial t} + \mathbf{u}\cdot\nabla\right)\mathbf{u} - \mathbf{div}(2\eta_s \mathsf{D}(u)) + \nabla p + \dfrac{\eta_p}{\mathsf{We}}\mathbf{div}\boldsymbol{\sigma}^c = 0, \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \mathbf{div}\mathbf{u} = 0, \\ \left(\dfrac{\partial}{\partial t} + \mathbf{u}\cdot\nabla\right)\boldsymbol{\sigma}^c - \nabla\mathbf{u}\boldsymbol{\sigma}^c - \boldsymbol{\sigma}^c(\nabla\mathbf{u})^{\mathrm{T}} + \dfrac{1}{\mathsf{We}}(\boldsymbol{\sigma}^c - \mathsf{I}) = 0 \end{cases} \tag{3}$$

where η_s and η_p are the amount of solvent and polymer contributions repectively. In [6] it is shown for 1D problems that the convection part is not able to balance the exponential growth of the stress. By introducing a new logarithmic variable, the positivity property of the conformation tensor is preserved by design. Indeed the conformation tensor is replaced by its logarithm through exact evaluation, i.e., eigenvalue computations, which leads to the **L**og **C**onformation **R**epresentation (**LCR**)

$$\psi = R\begin{pmatrix} \log\lambda_1 & 0 \\ 0 & \log\lambda_2 \end{pmatrix} R^{\mathrm{T}} \tag{4}$$

Here, $\lambda_{i=1,2}$ are the eigenvalues of the conformation tensor $\boldsymbol{\sigma}^c$ and R is the corresponding eigenvector matrix. Then, a new decomposition of the velocity gradient is introduced [5,6],

$$\nabla\mathbf{u} = \mathsf{G} + \Omega + \mathsf{N}(\boldsymbol{\sigma}^c)^{-1} \tag{5}$$

where G is a symmetric matrix which commutes with the conformation tensor, Ω is a pure rotation matrix (anti-symmetric matrix) and N is an antisymmetric matrix. Then, the constitutive laws in terms of conformation tensor $\boldsymbol{\sigma}^c$ and in terms of the log conformation tensor $\psi = \log\boldsymbol{\sigma}^c$ transform respectively into

$$\left(\frac{\partial}{\partial t} + \mathbf{u}\cdot\nabla\right)\boldsymbol{\sigma}^c - (\Omega\boldsymbol{\sigma}^c - \boldsymbol{\sigma}^c\Omega) - 2\mathsf{G}\boldsymbol{\sigma}^c = \frac{1}{\mathsf{We}}(\mathsf{I} - \boldsymbol{\sigma}^c), \tag{6}$$

and consequently with $\boldsymbol{\sigma}^c = e^{\psi}$:

$$\left(\frac{\partial}{\partial t} + \mathbf{u}\cdot\nabla\right)\psi - (\Omega\psi - \psi\Omega) - 2\mathsf{G} = \frac{1}{\mathsf{We}}(e^{-\psi} - \mathsf{I}) \tag{7}$$

Hence, the new set of equations of the LCR reformulation is written as follows:

$$\begin{cases} \rho\left(\dfrac{\partial}{\partial t} + \mathbf{u}\cdot\nabla\right)\mathbf{u} = -\nabla p + \mathbf{div}(2\eta_s \mathsf{D}(\mathbf{u})) + \dfrac{\eta_p}{\mathsf{We}}\mathbf{div}e^{\psi}, \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \mathbf{divu} = 0, \\ \left(\dfrac{\partial}{\partial t} + \mathbf{u}\cdot\nabla\right)\psi - (\Omega\psi - \psi\Omega) - 2\mathsf{G} = \dfrac{1}{\mathsf{We}}(e^{-\psi} - \mathsf{I}) \end{cases} \tag{8}$$

2 Spatial and Time Discretization

We apply implicit 2nd order time stepping methods to preserve the high accuracy and robustness in nonstationary flow simulations, for instance the Crank–Nicolson or Fractional-Step-ϑ scheme, which allow adaptive time stepping due to accuracy reasons only, but which do not depend on CFL-like restrictions. Then, the LCR equations are discretized in time as follows:

$$\begin{aligned} &\frac{\mathbf{u}^{n+1} - \mathbf{u}^n}{\triangle t} + \vartheta\left[\rho\mathbf{u}^{n+1}\cdot\nabla\mathbf{u} + \nabla p^{n+1} + 2\nabla(\eta_s\mathsf{D}(\mathbf{u}^{n+1})) + \frac{\eta_p}{\mathsf{We}}\mathbf{div}e^{\psi^{n+1}}\right] \\ &\qquad + (1-\vartheta)\left[\rho\mathbf{u}^{n}\cdot\nabla\mathbf{u} + \nabla p^{n} + 2\nabla(\eta_s\mathsf{D}(\mathbf{u}^{n})) + \frac{\eta_p}{\mathsf{We}}\mathbf{div}e^{\psi^{n}}\right] = 0 \\ &\mathbf{divu}^{n+1} = 0 \\ &\frac{\psi^{n+1} - \psi^n}{\triangle t} + \vartheta\left[\mathbf{u}^{n+1}\cdot\nabla\psi^{n+1} - (\Omega(\mathbf{u}^{n+1})\psi^{n+1} - \psi^{n+1}\Omega(\mathbf{u}^{n+1})) - 2\mathsf{G}(\mathbf{u}^{n+1})\right] \\ &\qquad + (1-\vartheta)\left[\mathbf{u}^{n}\cdot\nabla\psi^{n+1} - (\Omega(\mathbf{u}^{n})\psi^{n} - \psi^{n}\Omega(\mathbf{u}^{n})) - 2\mathsf{G}(\mathbf{u}^{n})\right] \\ &\qquad - \frac{\vartheta}{\mathsf{We}}\left[e^{-\psi^{n+1}} - \mathsf{I}\right] - \frac{1-\vartheta}{\mathsf{We}}\left[e^{-\psi^{n}} - \mathsf{I}\right] = 0 \end{aligned} \tag{9}$$

For the FEM approximation, we utilize the high order $Q_2/P_1^{disc}/Q_2$ finite element triple for discretization in space which can be applied on general meshes together with local grid refinement strategies including hanging nodes. Due to the velocity and stress coupling the choice of the velocity finite element space and the stress finite element space is subject to the LBB condition. In order to use the same finite element space for velocity as well as for the stress one has to use some stabilization techniques. Indeed, to maintain the elliptic character of the momentum equation, the jump term of the following form can be introduced [3, 7]

$$J_{\mathbf{u}}(\mathbf{u},\mathbf{v}) = \sum_{\text{edge E}} \max(\gamma_{\mathbf{u}}\eta_p h_E, \gamma^*_{\mathbf{u}} h_E^2) \int_E [\nabla \mathbf{u}] : [\nabla \mathbf{v}] ds \tag{10}$$

which relaxes the choice of the stress space even in the absence of the pure viscous contribution. Nevertheless the hyperbolic nature of the constitutive equations may require further treatment, so that similarly further jump terms for the stress may be introduced [3]:

$$J_{\psi}(\psi,\tau) = \sum_{\text{edge E}} \gamma_{\psi} h_E^2 \int_E [\nabla \psi] : [\nabla \tau] ds \tag{11}$$

Then, the discrete system reads as follows

$$\begin{pmatrix} S_{\mathbf{u}}(\mathbf{u}) & C & B \\ \widetilde{C}^T & S_{\psi}(\mathbf{u}) & 0 \\ B^T & 0 & 0 \end{pmatrix} \begin{pmatrix} \mathbf{u} \\ \psi \\ p \end{pmatrix} = \begin{pmatrix} \text{rhs}_{\mathbf{u}} \\ \text{rhs}_{\psi} \\ \text{rhs}_p \end{pmatrix} \tag{12}$$

where $S_{\mathbf{u}} = \frac{1}{\Delta t} M_{\mathbf{u}} + L_{\mathbf{u}} + K_{\mathbf{u}} + J_{\mathbf{u}}$, $S_{\psi} = \frac{1}{\Delta t} M_{\psi} + K_{\mathbf{u}} + K_{\Omega} + J_{\psi}$, $M_{\mathbf{u}}$ and M_{ψ} are mass matrices, $L_{\mathbf{u}}$ is the discrete diffusion operator, $K_{\mathbf{u}}$ the discrete convective term, K_{Ω} is the discrete operator such that $K_{\Omega}\psi = -(\Omega\psi - \psi\Omega)$, $\widetilde{C}^{\mathrm{T}} = M_{\mathsf{G}(\nabla \mathbf{u}, \sigma^c)}$, and C is the discrete matrix of $-\frac{\eta_p}{\mathsf{We}}\nabla \cdot \exp$. Furthermore, B and B^{T} are discrete analogous to the gradient and divergence operators.

3 Nonlinear and Linear Solvers

The strongly coupled system (12) is then linearized through a discrete Newton approach which results in the solution steps of the form

$$\mathbf{x}^{n+1} = \mathbf{x}^n + \omega^n \mathsf{J}^{-1}(\mathbf{x}^n)\mathsf{R}(\mathbf{x}^n)$$

where ω^n is a damping parameter. In this approach, we approximate the Jacobian $\mathsf{J} = \left[\frac{\partial \mathsf{R}(\mathbf{x}^n)}{\partial \mathbf{x}}\right]$ using divided differences

$$\left[\frac{\partial \mathsf{R}(\mathbf{x}^n)}{\partial \mathbf{x}}\right]_{ij} \approx \frac{\mathsf{R}_i(\mathbf{x}^n + \varepsilon \mathbf{e}_j) - \mathsf{R}_i(\mathbf{x}^n - \varepsilon \mathbf{e}_j)}{2\varepsilon} \tag{13}$$

with $\mathbf{x} = (\mathbf{u}, \psi, p)$, $\mathsf{R}(\mathbf{x})$ is the residual coming from the discrete problem of the system (12), and $e_i = \delta_{ij}$ is the standard Kronecker symbol. Hence, the resulting linear system is a typical saddle point problem which is solved effectively using coupled multigrid [4, 5], i.e., local Pressure Schur Complement approach as generalization of so-called Vanka smoothers which are simple iterative relaxation methods for such coupled systems of saddle point type. The smoothers are acting directly on element level and are embedded into an outer block Jacobi/Gauss-Seidel iteration. The local

character of this procedure together with a global defect-correction mechanism is crucial for this monolithic approach:

$$\begin{bmatrix} \mathbf{u}^{n+1} \\ \psi^{n+1} \\ p^{n+1} \end{bmatrix} = \begin{bmatrix} \mathbf{u}^{n} \\ \psi^{n} \\ p^{n} \end{bmatrix} + \omega^n \sum_{T \in \mathcal{T}_h^{\mathrm{T}}} \mathsf{J}_{|T}^{-1} \begin{bmatrix} \mathsf{R}_{\mathbf{u}} \\ \mathsf{R}_{\psi} \\ \mathsf{R}_{p} \end{bmatrix}_{|T} \tag{14}$$

The coarse grid discretizations are effectively done using the finite element approach, and the grid transfer operators (restriction and prolongation) are standard due to the conforming approximation. Here, the "summation" over each element $T \in \mathcal{T}_h^{\mathrm{T}}$ represents an assembling technique.

4 Numerical Examples

For prototypical numerical tests of this new approach, we consider the numerical simulation of both directly steady and nonstationary flow in a lid-driven cavity for the Oldroyd-B model. The initial condition for the stress tensor is unity and a regularized velocity boundary condition is implemented such that $\mathbf{u}(x,t) = (8(1 + \tanh\ 8(t - 0.5))x^2(1-x)^2, 0)^T$ on the top boundary while zero velocity on the rest of boundary is prescribed. For direct steady simulations the velocity profile evolves to $\mathbf{u}(x,t) = (16x^2(1-x)^2, 0)^T$ on the boundary. For the total viscosity (zero-shear viscosity), η_s and η_p are equal to 1. The simulation is performed with the mesh size $h = 1/64$ and with coarse mesh size $h = 1/4$. The time step is chosen to be $\triangle t = 0.1$ in the sense that no further improvement in kinetic energy with respect to smaller time steps could be observed. The number of cells for the corresponding computation level n is $\mathrm{Ln} = 2^{4+2n}$. We calculate the kinetic energy by $\frac{1}{2}\|\mathbf{u}_h\|^2_{L_2(\Omega)}$ and analyze the impact of jump stabilization for different We numbers. For $\mathsf{We} = 1$, the kinetic energy seems to reach a steady state as shown in Fig. 1 and it remains steady at least up to time $t = 30$. As the We number increases the kinetic energy oscillates stronger and the LCR variable becomes more spurious at time $t = 30$, see Fig. 2. Longer computation times may lead to numerical break down. EO-FEM in this case is able to relax these oscillations, thus it significantly improves numerical stability.

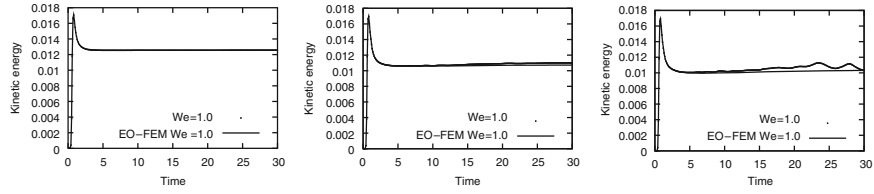


Fig. 1 Driven cavity flow: Kinetic energy until $t = 30$ for different We numbers with and without EO-FEM

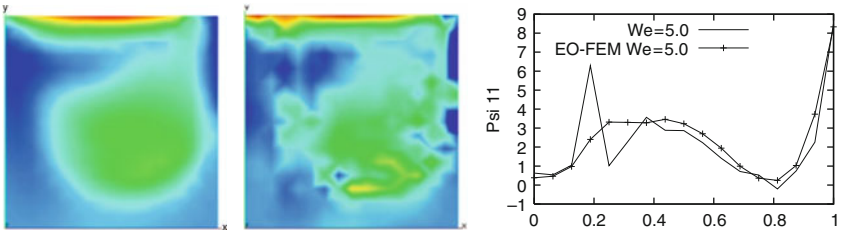


Fig. 2 Driven cavity flow: The plot of the sterss ψ_{11} with EO-FEM (*left*), without EO-FEM (*middle*) and the Cutline of ψ_{11} at $x = 0.5$, $t = 30$ with and without EO-FEM (*right*)

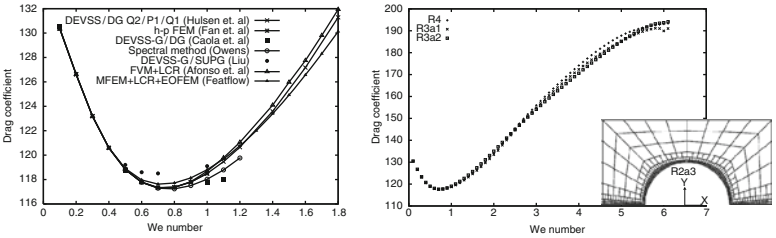


Fig. 3 Planar flow around cylinder: Drag coefficient from different authors (*left*) and for different levels for higher We with EO-FEM (*right*) and one exemplary computational mesh with local refinement

Next, we consider planar flow around cylinder and plot the drag up to $\mathsf{We} = 1.8$ in which the drag coefficients are comparable with other authors as can be seen in Fig. 3. However, it is remarkable that with the LCR formulation, results for quite high Weissenberg numbers in comparison to standard formulation can be easily obtained. While usually the maximum We number, which can be obtained by LCR, is in the range of $\mathsf{We} = 1.8$ or $\mathsf{We} = 2.0$, see [1, 6], here EO-FEM helps to go further as far as $\mathsf{We} = 6.0$. Note that this is calculated with a direct steady approach which shows the big potential of EO-FEM stabilization for viscoelastic flow. Further results can be seen in Fig. 4 which shows the stress behavior w.r.t. We numbers and different meshes (for more detail see [5]). As mentioned before, the linear subproblem is handled by a special monolithic multigrid solver. In Table 1 we show the corresponding convergence behavior in a direct steady approach with respect to the number of nonlinear iterations for increasing We numbers. Multigrid seems to be stable with respect to the mesh refinement and the nonlinearity of the problem as the number increases.

Finally, we present preliminary results for the planar 4:1 contraction problem which is one of the most well-known benchmarks for viscoelastic flow. As a current result for this configuration, we are able to reproduce the qualitative phenomenon of lip vortex growth with respect to increasing We number (Fig. 5) in which case we perform the calculations on a locally refined mesh with hanging nodes as shown in Fig. 6.

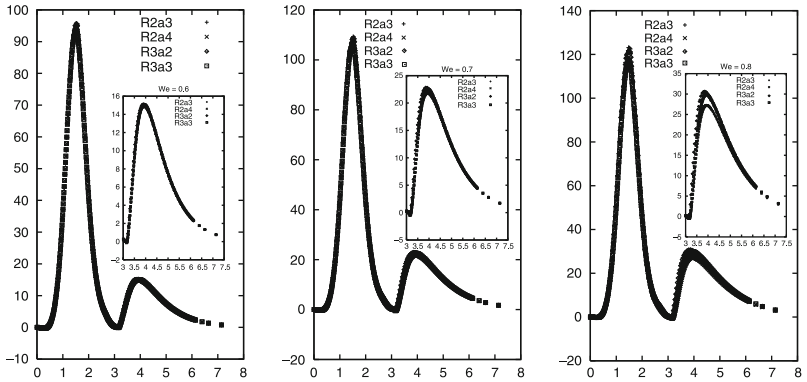


Fig. 4 Planar flow around cylinder: Normal stress convergence with local refinement for Weissenberg numbers $\mathsf{We} = 0.6$ (*left*), $\mathsf{We} = 0.7$ (*middle*) and $\mathsf{We} = 0.8$ (*right*) with the zoom of in the wake part

Table 1 Newton-multigrid behavior: Nonlinear iterations (NNL)/Average multigrid sweeps (AVMG) per nonlinear iterations for several levels refinement (Ri, i = 1,4), different We numbers and different linear tolerance parameters ϵ for planar flow around cylinder configuration

We	0.01		0.1		1.0	
ϵ	0.1	0.01	0.1	0.01	0.1	0.01
R1	9/2	5/3	10/1	7/3	14/1	10/3
R2	9/3	5/5	10/2	7/4	16/2	10/5
R3	9/3	5/6	10/3	7/5	16/2	11/5
R4	9/3	5/6	10/3	9/5	13/3	11/5

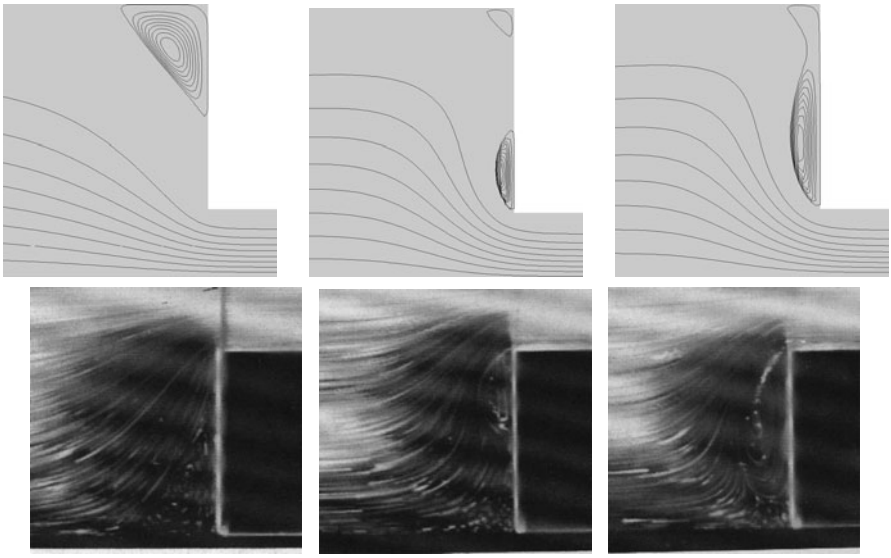


Fig. 5 Lip vortex growth for Oldroyd-B model: Numerical simulation (*top*) vs. experiment (*bottom* [2]) for lip vortex growth in a 4 to 1 contraction

Fig. 6 The planar 4:1 contraction: Computational mesh with local refinement

5 Conclusion

We have presented special numerical simulation techniques for viscoelastic flow within a monolithic finite element framework of utilizing the new LCR technique for Oldroyd-B type of fluids. Edge-oriented FEM stabilization is implemented to increase the numerical stability. Together with local refinement the method shows to be a very promising way for solving viscoelastic flow problems particularly for high We numbers. Several numerical examples of cavity flow, flow around cylinder and the growth of lip vortex in a contraction flow are also presented. Numerical stability has been significantly improved by the help of stabilization. Future work will include the implementation of LCR in other viscoelastic models together with an additional coupling of the energy equation with a viscous dissipation term, see [4], in order to be able to simulate more realistic flow problems, particulary in 3D.

Acknowledgements This work was supported by the German Research Association (DFG) through the collaborative research center SFB/TRR 30 and through the grant TU 102/21 and by the Graduate School of Production Engineering and Logistics.

References

1. Afonso, A., Oliveira, P.J., Pinho, F.T. and Alves, M.A. The log-conformation tensor approach in the finite-volume method framework, J. Non-Newt. Fluid Mech., **157** 55–65 (2009)
2. Boger, D. V. and Walters, K. Rheological phenomena in focus, Elsevier, Amsterdam (1993)
3. Bonito, A. and Burman, E. A continuous interior penalty method for viscoelastic flows, SIAM J. Sci. Comput., **30** 1156–1177 (2008)
4. Damanik, H., Hron, J., Ouazzi, A. and Turek, S. A monolithic FEM–multigrid solver for non–isothermal incompressible flow on general meshes, J. Comput. Phys., **228** 3869–3881 (2009)
5. Damanik, H., Hron, J., Ouazzi, A. and Turek, S. A monolithic FEM approach for the log–conformation reformulation (LCR) of viscoelastic flow problems, J. Non-Newt. Fluid Mech. (2010), in press: DOI 10.1016/j.jnnfm.2010.05.008
6. Hulsen, M. A., Fattal, R. and Kupferman, R. Flow of viscoelastic fluids past a cylinder at high Weissenberg number: Stabilized simulations using matrix logarithms, J. Non-Newt. Fluid Mech., **127** 27–39 (2005)
7. Turek, S. and Ouazzi, A. Unified edge–oriented stabilization of nonconforming FEM for incompressible flow problems: Numerical investigations, J. Numer. Math., **15** 299–322 (2007)

A Posteriori Estimates for Variational Inequalities

S. Repin

Abstract This paper is concerned with guaranteed and computable error bounds for approximate solutions of variational inequalities. The estimates are derived by purely functional methods. The first method is based upon methods of convex analysis and calculus of variations and the second one derives estimates with the help of certain transformations of the corresponding variational inequality. Both methods (variational and nonvariational) has been earlier developed and applied for linear problems where they lead to the same estimates [Two-sided estimates of deviation from exact solutions of uniformly elliptic equations, 2001]. In the paper, we shortly discuss variational inequalities associated with obstacle type problems and show that both methods also result in the same error majorants. The majorants are valid for any approximation from the admissible functional class and does not exploit Galerkin orthogonality, higher regularity of solutions, or a priori information on the structure of coincidence set. Also, the paper contains a concise overview of results related to similar a posteriori error estimates derived for other classes of nonlinear problems.

1 Introduction

Mathematical theory of variational inequalities was created in the second half of the twentieth century and nowadays presents an important part of nonlinear analysis related to a wide spectrum of nonlinear models in continuum mechanics (see, e.g., Duvaut and Lions [9], Friedman [11]). Numerical methods for variational inequalities were studied by many authors. At this point we first of all mention the book Glowinski, Lions, Tremolieres [14]. One of the first error estimation results was obtained by Falk [10] who derived a priori rate convergence estimates for

S. Repin
St. Petersburg Department of Steklov Institute of Mathematics, Russian Academy of Sciences, 191023, Fontanka 27, St. Petersburg
e-mail: repin@pdmi.ras.ru

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_81,

approximations of a problem with obstacles. This short note does not have space for a systematic overview of the results obtained in a posteriori error control of variational inequalities. We mention only several papers where other references can be easily found. A posteriori estimates (by residual type approaches) were derived in Ainsworth, Oden and Lee [1], Braess [5], Chen and Nochetto [8], Hoppe and Kornhuber [15], Kornhuber [16], and other authors. Pointwise estimates for problems with obstacles were obtained in Nochetto, Siebert, and Veeser [18]. An approach based on equilibration was recently suggested in Braess, Hoppe, Schöberl [6].

In this paper, we discuss a different class of methods developed for the derivation of a posteriori estimates. These methods operate on purely functional grounds without attracting such properties as Galerkin orthogonality, extra regularity of exact solutions, and superconvergence of approximate solutions. A posteriori estimates of this type provide guaranteed and computable measures of the distance (measured in the natural energy norm) between the exact solution of a BVP and any conforming approximation. Special properties of approximations (or numerical method) can be exploited later, when a concrete solution is substituted into the error majorant.

For convex variational problems the estimates (which are called *a posteriori estimates of the functional type*) has been derived in the middle of 90s (see [17, 20, 23, 24] for a systematic overview). At present, functional a posteriori estimates has been derived for problems generated by all main classes of linear differential equations. In particular, let us consider a linear elliptic problem: find $u \in V_0 + u_0$ such that

$$a(u, w) + \langle \ell, w \rangle = 0 \quad \forall w \in V_0. \tag{1}$$

Here, V_0 subspace of a Banach space V, $a(u, w) = (\mathscr{A}\Lambda u, \Lambda w)$ is a coercive bilinear form generated by a self-adjoint operator $\mathscr{A} \in \mathscr{L}(U, U)$, U is a Hilbert space with the norm $\|y\| = (y, y)^{1/2}$, V is compactly embedded in U, $\Lambda : V \to U$ is a bounded linear operator, $\langle \ell, w \rangle = (f, w) + (g, \Lambda w)$, and it is assumed that

$$c_1 \|y\|^2 \le (\mathscr{A} y, y) := \| y \|^2 \le c_2 \|y\|^2, \qquad \forall y \in U, \tag{2}$$

$$\|\Lambda w\| \ge c_3 \|w\|_V, \quad \forall w \in V_0. \tag{3}$$

For this class of elliptic problems, the general form of the error majorant is given by the theorem.

Theorem 1 ([19, 21]). *For any $v \in V_0 + u_0$ and any $y \in Q^* := \{y \in U \mid \Lambda^* y \in U\}$, the following estimate holds*

$$\| \Lambda(v - u) \| \le \| \mathscr{A}\Lambda v - y \|_* + c \, \|\ell + \Lambda^* y\|, \tag{4}$$

where $c = c^{-1} c_3^{-1}$, $\| y \|_* = (\mathscr{A}^{-1} y, y)^{1/2}$, and Λ^* is the operator adjoint to Λ.

In the simplest case

$$\Delta u + f = 0 \qquad \text{in } \Omega \quad u = u_0 \qquad \text{on } \partial\Omega,$$

we have

$$\|\nabla(v-u)\| \le \|\nabla v - y\| + C_{F\Omega}\|\operatorname{div} y + f\|, \qquad \forall\, y \in H(\Omega, \operatorname{div}),$$

where v is any function in H^1 satisfying the boundary condition and $C_{F\Omega}$ is a constant in the Friedrichs-Poincare inequality. This estimate holds for any $C \ge C_{F\Omega}$. The case $C = +\infty$ leads to the Prager–Synge estimate.

Estimate (4) is a particular form of the estimates derived in [17, 19, 21] for convex variational problems by methods of duality theory in the calculus of variations. In [23], the same estimate was obtained by a nonvariational method based on transformations of the integral relation that defines the generalized solution.

2 Variational Inequalities of Elliptic Type

Let $j : V \to \mathrm{R}$ be a given convex continuous functional. We consider the following problem: find $u \in K$ such that the inequality

$$a(u, w-u) + j(w) - j(u) \ge \langle \ell, w-u\rangle \tag{5}$$

holds for any $w \in K$, where K is a convex closed subset of V and $\ell \in V^*$. Our goal is to derive computable estimates of the difference between u and any function $v \in K$. As for linear problems, there are two methods of deriving error majorants:

- Variational method based on so-called "perturbed" problems.
- Nonvariational method based on transformations of (5).

We discuss them with the paradigm of the simplest obstacle problem. In this case,

$$K := \{v \in V_0 := \overset{\circ}{H}{}^1 \mid \phi(x) \le v(x) \le \psi(x) \text{ a.e. in } \Omega\},$$

where $\phi, \psi \in H^2(\Omega)$ are two given functions.

Exact solution of the problem meets the variational inequality

$$\int_\Omega A\nabla u \cdot \nabla(w-u)dx \ge \int_\Omega f(w-u)dx \quad \forall w \in K_{\phi\psi}$$

and generates three sets:

$$\begin{aligned}
\Omega^u_\oplus &:= \{x \in \Omega \mid u(x) = \psi(x)\} && \text{(upper coincidence set)},\\
\Omega^u_\ominus &:= \{x \in \Omega \mid u(x) = \phi(x)\} && \text{(lower coincidence set)},\\
\Omega^u_0 &:= \{x \in \Omega \mid \phi(x) < u(x) < \psi(x)\}\,.
\end{aligned}$$

Here Ω_0^u is an open set, where a solution satisfies the differential equation. We note that exact solutions of obstacle problems has essentially different properties with respect to solutions of linear problems. In particular, it is well known that u has a limited regularity (even for smooth data $u \in W_2^2(\Omega)$; in the best case scenario $u \in W_\infty^2(\Omega)$). Besides, the solution has unknown free boundaries. These facts make derivation of a posteriori estimates much more complicated.

2.1 Deriving of Error Majorants by the Variational Method

Functional type error majaorants were derived in [7,21] with the help of variational techniques. In these papers, elliptic problems with obstacles were considered. Later, it was extended to certain classes of nonlinear fluids (see [12, 13, 22] and the references therein), variational inequalities for fourth order elliptic operators [3], and elasto-plastic torsion problem [4]. In this method, the first step consists of deriving the estimate

$$\frac{1}{2} \| v-u \|^2 \leq J(v) - \inf \mathscr{P} \leq J(v) - J^*(\tau^*), \;\; J(v) := \frac{1}{2} a(v,v) + j(v) - (f,v), \tag{6}$$

where $J^* : Y^* \to I\!R$ – functional of the dual problem. However, usually variational problems generated by variational inequalities (unlike problems related to linear elliptic problems) do not have J^* representable in an explicit form. A way to overcome this difficulty is to consider the so-called *perturbed problem* for the functional

$$J_\lambda(v) := J(v) - \int\limits_\Omega \lambda \cdot (\mathbf{v} - \Phi)\, dx,$$

where $\Phi = (\phi, -\psi)$ and $\mathbf{v} = (v, -v)$. It is easy to see that

$$\sup_{\lambda \in L_\oplus} J_\lambda(v) = J(v) - \inf_{\lambda \in L_\oplus} \int\limits_\Omega \lambda \cdot (\mathbf{v} - \Phi)\, dx = \begin{cases} J(v) & \text{if} v \in \mathrm{K}_{\phi\psi} \\ +\infty & \text{if} v \notin \mathrm{K}_{\phi\psi}. \end{cases},$$

where $L_\oplus := \{(\lambda_1, \lambda_2) \mid \lambda_i \in L_2(\Omega), \lambda_i(x) \geq 0 \,\text{a.e. in}\, \Omega\}$. Now, we arrive at the problem $\mathscr{P}_\lambda$: Find $u_\lambda \in V_0$ such that

$$J_\lambda(u_\lambda) = \inf_{v \in V_0} J_\lambda(v) := inf\, \mathscr{P}_\lambda. \tag{7}$$

Since

$$\inf_{v \in V_0} J_\lambda(v) \leq \inf_{v \in \mathrm{K}_{\phi\psi}} J_\lambda(v) = \inf_{v \in \mathrm{K}_{\phi\psi}} J(v) = inf\, \mathscr{P}, \quad \forall\, \lambda \in L_\oplus,$$

we see that $\inf \mathscr{P}_\lambda \le \inf \mathscr{P}$ and, consequently,

$$\frac{1}{2} \| v - u \|^2 \le J(v) - \inf \mathscr{P}_\lambda. \tag{8}$$

Unlike (6), this estimate generates an explicitly computable upper bound (because the problem dual to $\mathscr{P}_\lambda$ has an explicit form). As a result, we have an upper bound, which is valid for any $\lambda \in L_\oplus$. By a special choice of λ we arrive at the estimate

$$\| \nabla(u - v) \| \le \left(\int_\Omega (A\nabla v \cdot \nabla v + A^{-1} y \cdot y - 2y \cdot \nabla v) dx \right)^{1/2} + C_{F\Omega} \| [f + \operatorname{div} y]_v \|, \tag{9}$$

where

$$[f + \operatorname{div} y]_v := \begin{cases} (f + \operatorname{div} y)_\ominus & \text{a.e. in } \Omega^v_\oplus, \\ f + \operatorname{div} y & \text{a.e. in } \Omega^v_0, \\ (f + \operatorname{div} y)_\oplus & \text{a.e. in } \Omega^v_\ominus, \end{cases}$$

is the generalized residual term associated with the obstacle problem.

Remark 1. Estimate (9) has a simple and easily explainable structure. However it is does not give the best upper bound, which follows if λ is selected by solving a special optimization problem. The majorant obtained with the "optimal" λ has a more complicated structure (see [17, 21, 24]) but provides a sharper error bound.

2.2 Deriving Error Majorants Directly from the Variational Inequality

Let $v \in K_{\phi\psi}$ be an approximate solution. We have

$$a(u - v, u - v) \le \int_\Omega (f(u - v) - A\nabla v \cdot \nabla(u - v)) \, dx.$$

By the integral identity

$$\int_\Omega (w \operatorname{div} y + y \cdot \nabla w) dx = 0 \quad \forall w \in V_0, \ y \in H(\Omega, \operatorname{div}),$$

we arrive at the relation

$$a(u - v, u - v) \le \int_\Omega (f + \operatorname{div} y)(u - v) dx + \int_\Omega (y - A\nabla v) \cdot \nabla(u - v) \, dx. \tag{10}$$

Since the function v is known, the sets

$$\begin{aligned}
\Omega^v_{\oplus} &:= \{x \in \Omega \mid \quad v(x) = \psi(x) \quad \}, \\
\Omega^v_{\ominus} &:= \{x \in \Omega \mid \quad v(x) = \phi(x) \quad \}, \\
\Omega^v_{0} &:= \{x \in \Omega \mid \phi(x) < v(x) < \psi(x) \quad \}.
\end{aligned}$$

are explicitly defined. Thus, the generalized residual term $[f + \mathrm{div} y]_v$ is fully defined with the help of approximate knowledge on the coincidence set contained in v. It is easy to observe that

$$\int_{\Omega} (f + \mathrm{div} y)(u - v)dx \le \int_{\Omega} [f + \mathrm{div} y]_v (u - v)dx$$

and

$$\int_{\Omega} (A\nabla v - y) \cdot \nabla w \, dx \le \| A\nabla v - y \|_* \| \nabla w \| .$$

In view of the above two relations, we obtain

$$\| \nabla(u - v) \|^2 \le \| [f + \mathrm{div} y]_v \| \| v - u \| + \| A\nabla v - y \|_* \| \nabla(v - u) \|.$$

and by Friedrichs' inequality arrive at estimate

$$\| \nabla(u - v) \| \le \| A\nabla v - y \|_* + C_{F\Omega} \| [f + \mathrm{div} y]_v \|, \tag{11}$$

which is equivalent to (9).

It is not difficult to show that the right hand side of (11) vanishes if and only if v coincides with the exact solution u. Indeed, assume that $y = A\nabla v$ and

$$\begin{aligned}
(f + \mathrm{div} y)_{\ominus} &= 0 \quad \text{a.e. in } \Omega^v_{\oplus}, \\
f + \mathrm{div} y &= 0 \quad \text{a.e. in } \Omega^v_{0}, \\
(f + \mathrm{div} y)_{\oplus} &= 0 \quad \text{a.e. in } \Omega^v_{\ominus}.
\end{aligned}$$

Then

$$\begin{aligned}
\int_{\Omega} A\nabla v \cdot \nabla(v - w)dx &= \int_{\Omega} y \cdot \nabla(v - w)dx = \int_{\Omega} (\mathrm{div} y + f)(w - v)dx + \int_{\Omega} f(v - w)dx \\
&\le \int_{\Omega^v_{\oplus}} (f + \mathrm{div} y)_{\ominus}(w - v)dx + \int_{\Omega^v_{\ominus}} (f + \mathrm{div} y)_{\oplus}(w - v)dx + IntOf(v - w)dx \\
&= \int_{\Omega} f(v - w)dx \quad \forall w \in K_{\phi\psi},
\end{aligned}$$

and we find that v satisfies the variational inequality.

Remark 2. Another form of the error majorant that contains (instead of $C_{F\Omega}$) constants in the Payne-Weinberger inequalities associated with a certain splitting of Ω into a collection of convex subdomains has been recently obtained in [25].

3 A Posteriori Estimates for Other Nonlinear Problems

Generalized Newtonian fluids present another class of practically important mathematical models related to variational inequalities. For example, stationary flow of the Bingham fluid can be reduced to the variational problem

$$\inf_{v\in V_0}\int\limits_{\Omega}\Big(\frac{\nu}{2}\mid\nabla v\mid^2+k_*\mid\nabla v\mid-fv\Big)\,dx, \tag{12}$$

where $\Omega\in\mathrm{R}^2$ is a cross-section of the pipe, ν and k_* are positive (material) constants, and f is a constant (pressure difference per length unit). In this case, the dissipative potential contains a nondifferentiable term $\psi(\xi)=k_*|\xi|$, which leads to a variational inequality of the second kind. Many other models of nonlinear viscous fluids lead to variational inequalities. For these problems, estimates of deviations from exact solutions has been obtained in (see [12, 13, 21, 22]). In particular, for the problem (12) difference between the minimizer u and a (conforming) approximation v can be estimated as follows:

$$\tfrac{\nu}{2}\,\|\nabla(v-u)\|^2\le\int\limits_{\Omega}\Big(\frac{(1+\beta)}{2\nu}(\nu\nabla v-\eta)^2+k_*\mid\nabla v\mid-\nabla v\cdot\tau\Big)dx+\mathbf{n} \tag{13}$$

$$+\left(1+\frac{1}{\beta}\right)\frac{1}{\nu}C_{\Omega}^2\|\mathrm{div}(\eta+\tau)+f\|^2, \tag{14}$$

where $\eta,\tau\in L_2(\Omega,\mathrm{R}^2)$, $\mathrm{div}(\eta+\tau)\in L_2(\Omega)$, $\mid\tau(x)\mid\le k_*$ for almost all $x\in\Omega$.

A posteriori estimates for variational inequalities generated by elliptic problems with nonlinear boundary conditions were studied in [26] and for incremental elasto-plastic models in [27], and for the Ramberg–Osgood model in [2]. Finally, we note that estimates of the above discussed type has been derived for general type convex nonlinear problems in [21] and are discussed in the books [17, 24].

Acknowledgement Supported by Academy of Finland and by DAAD program of Germany.

References

1. Ainsworth, M., Oden, J.T., Lee, C. Y.: Local a posteriori error estimators for variational inequalities. Numer. Meth. PDE, **9**, 23–33 (1993)
2. Bildhauer, M., Fuchs, M., Repin S.: A functional type a posteriori error analysis for the Ramberg–Osgood model. Z. Angew. Math. Mech., **87**, 860–876 (2007)

3. Bildhauer, M., Fuchs, M., Repin S.: Duality based a posteriori error estimates for higher order variational inequalities with power growth functionals. Ann. Acad. Sci. Fenn. Math., **33**(2), 475–490 (2008)
4. Bildhauer, M., Fuchs, M., Repin S.: The elasticplastic torsion problem: a posteriori error estimates for approximate solutions. Numer. Funct. Anal. Optim., **30**(7/8), 653-664(2009)
5. Braess, D.: A posteriori error estimators for obstacle problems – another look. Numer. Math., **101**, 415–421 (2005)
6. Braess, D., Hoppe, R., Schöberl, J.: A posteriori estimators for obstacle problems by the hypercircle method. Comput. Vis. Sci. 11 (4–6), 351–362 (2008)
7. Buss H., Repin, S.: A posteriori error estimates for boundary-value problems with obstacles. In: Numerical mathematics and advanced applications (Jyvaskyla, 1999), pp. 162–170. World Sci. Publishing, New York (2000)
8. Chen, Z., Nochetto, R.: Residual type a posteriori error estimates for elliptic obstacle problems. Numer. Math. **84**, 527–548 (2000)
9. Duvant, G., Lions, J.-L.: Les inequations en mecanique et en physique. Dunod, Paris (1972)
10. Falk, R. S.: Error estimates for the approximation of a class of variational inequalities. Math. Comput. **28**, 963–971 (1974)
11. Friedman, A.: Variational principles and free-boundary problems. Wiley, New York (1982)
12. Fuchs, M., Repin, S.: Estimates for the deviation from the exact solutions of variational problems modeling certain classes of generalized Newtonian fluids. Math. Meth. Appl. Sci., **29**, 2225–2244 (2006)
13. Fuchs, M., Repin, S.: Estimates of the deviations from the exact solutions for variational inequalities describing the stationary flow of certain viscous incompressible fluids. Mathematical Methods in the Applied Sciences. **33**(9), 1136–1147 June (2010).
14. Glovinski, R., Lions, J.-L., Trémolierès, R.: Analyse numérique des inéquations variationnelles. Dunod, Paris (1976)
15. Hoppe, R., Kornhuber, R.: Adaptive multilevel methods for obstacle problems. SIAM J. Numer. Anal., **31**, 301–323 (1994)
16. Kornhuber, R.: A posteriori error estimates for elliptic variational inequalities. Comput. Math. Appl., **31**, 49–60 (1996)
17. Neittaanmäki, P., Repin, S.: Reliable methods for computer simulation, Error control and a posteriori estimates. Elsevier, New York (2004)
18. Nochetto, R., Siebert, K., Veeser, A.: Pointwise a posteriori error control for elliptic obstacle problems. Numer. Math., **95**, 163–195 (2003)
19. Repin, S.: A posteriori estimates for approximate solutions of variational problems with strongly convex functionals. In: Problems of Mathematical Analysis, **17**, pp. 199–226 (1997) (in Russian). English translation J. Math. Sci, (New York), **97**, 4311–4328 (1999)
20. Repin, S.: A posteriori error estimates for variational problems with uniformly convex functionals. Math. Comput., **69**, 481–500 (2000)
21. Repin, S.: Estimates of deviations from exact solutions of elliptic variational inequalities. Zapiski Nauchn. Semin. V.A. Steklov Math. Inst. St.-Petersburg (POMI)., **271**, 188–203 (2000)
22. Repin, S.: Estimates of deviations for generalized Newtonian fluids. Zapiski Nauchn. Semin. V.A. Steklov Math. Inst. St.-Petersburg (POMI)., **288**, 178–203 (2002)
23. Repin, S.: Two-sided estimates of deviation from exact solutions of uniformly elliptic equations. In: Proc. St. Petersburg Math. Society, **IX**, 143–171 (2001). English translation in Amer. Math. Soc. Transl. Ser. 2, **209**, Amer. Math. Soc., Providence, RI (2003)
24. Repin, S.: A posteriori estimates for partial differential equations. Walter de Gruyter, Berlin (2008)
25. Repin, S.: Estimates of deviations from exact solutions of variational inequalities based on Payne-Weinberger inequality. J. Math. Sci., **157**(6), 874–884 (2009)
26. Repin, S., Valdman, J.: Functional a posteriori error estimates for problems with nonlinear boundary conditions. J. Numer. Math. **16**(1), 51–81 (2008)
27. Repin, S., Valdman, J.: Functional a posteriori error estimates for incremental models in elastoplasticity. Cent. Eur. J. Math., **7**(3), 506–519 (2009)

Review on Longest Edge Nested Algorithms

Maria-Cecilia Rivara

Abstract Longest edge nested algorithms for triangulation refinement produce hierarchies of quality and nested irregular triangulations as needed for adaptive finite element and multigrid methods. In addition, right-triangle bintree triangulations are special longest edge methods used for terrain modelling and visualization. We review the algorithms and their properties.

1 Iterative Longest Edge Bisection of Individual Triangles

Longest edge nested algorithms [12,13,17] are based on the mathematical properties of the longest edge bisection of triangles. The study of the bisection method began in a series of papers [1, 10, 21–23] around three decades ago. First, Rosenberg and Stenger [21] proved that the method does not degenerate the smallest angle of the triangles generated by showing that it does not decrease beyond $\sigma/2$, where σ is the smallest angle from the initial triangle. Then Kearfott [10] proved a bound on the length of the longest side of any triangle obtained. Later Stynes [22] presented a better bound for certain triangles. This bound was improved independently by Stynes [23] and Adler [1] for all triangles. From their proofs they also deduced that the number of classes of similarity of triangles generated is finite, although they give no bound. Only recently Gutierrez et al. [6] studied complexity aspects of the bisection method based on a systematic classification of triangles.

A triangle ABC is bisected by the longest edge AB, by joining the midpoint D of AB with the opposite (biggest angled) vertex C (Fig. 1). The analysis of the iterative bisection is based on the geometrical position of vertex C of triangle ABC, assuming $AB \geq BC \geq CA$. See Fig. 3, where AB represents the longest side of the hypothetical triangle, D, the midpoint of AB, M is the midpoint of AD, N is such that $AN = AB/3$, $MO \perp AB$ and $DP \perp AB$. Arcs C_1, C_2, C_3 and C_4 belong, respectively, to

M.-C. Rivara
Maria Cecilia, Department of Computer Science, University of Chile
e-mail: mcrivara@dcc.uchile

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_82, © Springer-Verlag Berlin Heidelberg 2010

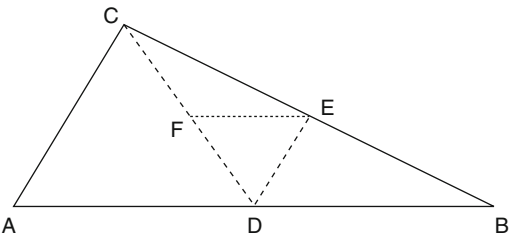


Fig. 1 Illustration of some longest edge bisections

circles $C(B, AB)$, $C(D, AD)$, $C(N, AN)$, and $C(A, AD)$. From the condition $AB \geq BC \geq CA$, it follows that vertex C of a triangle with base AB must be in the region bounded by arc C_1 and straight lines PD and AD. We partition this region into six subregions, with the property that triangles in the same subregion present similar behavior with regard to bisection by the median of the longest side, as stated in Lemma 1. Note that arc C_3 is the set of points C for which $BC = 2CD$. This separates those triangles for which the bisection of triangle DEC is performed by the edge CD implying that the generation of new non-similar triangles stops, from those for which the bisection is performed by edge CE.

Lemma 1. *Let ABC be a triangle. For the iterative process described above it holds:*

1. *If C is in region I, it generates at most four non-similar triangles.*
2. *If C is in region II, it generates at most five non-similar triangles.*
3. *If C is in region III, new $\triangle ADC$ belongs either to regions II or III. Moreover, after no more than $\lceil 5.7 \log(\frac{\pi}{6\sigma}) \rceil$ steps the only new triangle generated not similar to any previously generated belongs to region II.*
4. *If C is in region IV or V, after no more than $\lceil (\gamma - \pi/2)/\sigma \rceil$ steps, the only new triangle not similar to any previously generated has $\gamma \leq \pi/2$ (i.e., belongs to region I, II or III.)*
5. *If C is in region VI, new $\triangle ADC$ belongs to region I.*

The main theorem is stated as follows:

Theorem 1. *Let ABC a triangle, σ its smallest angle, and γ its biggest angle.*

1. *The number of steps to be executed by the bisection method until no more non-similar triangles are generated is $\mathcal{O}(\sigma^{-1})$.*
2. *If C is above arc C_3, then the number of non similar triangles generated by the bisection method is $\mathcal{O}(\log(\sigma^{-1}))$.*
3. *The number of non similar triangles generated by the bisection method is asymptotically bounded by a subexponential function of the parameter in γ/σ, i.e., it is $\mathcal{O}(a^{(\gamma/\sigma)^b})$ for any constants $a > 1$ and $b > 0$.*

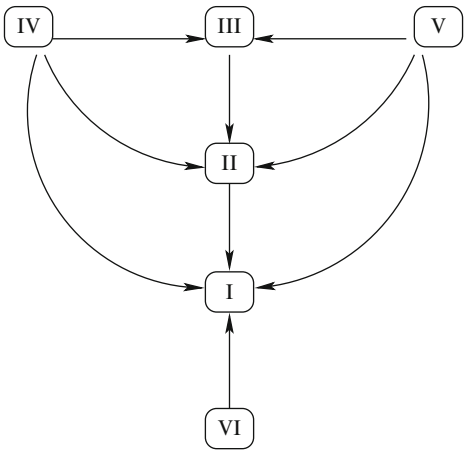


Fig. 2 A directed graph illustrating Lemma 1. *Arcs* indicate possible jumps in our proof, where the triangle being bisected recursively jumps from one region to another

The proof [6] is based on two facts: (1) In the bisection method, the new non-similar triangles generated which are not similar to any previously generated triangles follow the paths in the graph shown in Fig. 2 as proved in Lemma 1 (possibly staying in a node several steps). (2) In the counting process of the number of non-similar triangles, one only needs to examine the behavior of triangle *ADC* and, for regions IV and V, also the triangle *CDE*.

Using these results it is possible to re-prove classical results on the iterative bisection. Define triangle *ABC* as the unique triangle of level 0, and the triangles of level $i+1$ as those 2^{i+1} triangles obtained by bisecting the triangles of level i. Also define the diameter of level j as the greatest longest edge of the triangles of level j.

Theorem 2. *1. The bisection method gives $\mu_{ABC} \geq \frac{1}{2}\sigma_{ABC}$, where μ_{ABC} is the smallest angle of any triangle in any level. For triangles below arc C_2 it holds that $\mu_{ABC} = \sigma_{ABC}$, where σ_{ABC} is the smallest angle of the initial triangle.*

2. Define d_j as the diameter of triangles in level j. Then, (i) $d_5 \leq d_0/2$, i.e., after five bisection levels, the diameter of the generated mesh is no more than half of the original, and (ii) $d_j \leq c2^{-j/2}d_0$, where c is a small constant depending on the regions.

The proof of (1) is based on the fact that the only case when the smallest angle diminishes in a bisection step occurs for triangles in region I (see Fig. 3), being the worst case when $C = P$, that is, equilateral triangles. A proof of (2) uses the area of a triangle *ABC* and the fact that the area decreases exactly by half after a bisection. Then: (i) for triangles whose vertex C is below arcs C_2 or C_4, the diameter decreases by half after two levels, i.e., $d_2 \leq d_0/2$; and (ii) we use the fact we already know that, as bisection progresses, triangles "go up" the boundary of arcs C_4 and C_2. For a detailed proof see [6].

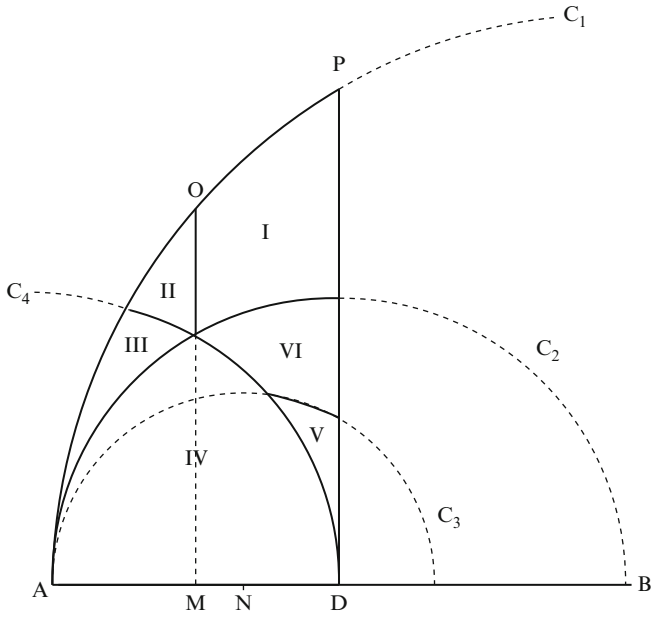


Fig. 3 Regions defining classes of triangles *ABC*. A (virtual) vertex C lying in one of the regions defines a triangle *ABC* with longest side *AB*, and greatest angle $\angle ACB$ (denoted γ)

2 Longest Edge Nested Algorithms

We consider conforming triangulations where the intersection of pairs of neighbor triangles is either a common edge or a common vertex. To simplify we consider a refinement region R and a condition over the size of the longest-edge of the triangles (a length parameter δ) to fix the desired resolution.

Definition 1. *Triangulation Refinement Problem:* given a quality and conforming triangulation (with angles greater than or equal to an angle α) of a polygonal region D, construct a locally refined, quality and conforming triangulation such that the longest edge of the triangles that intersect the refinement region R are less than δ.

The refinement area can be zero if the refinement is performed around one vertex or along a boundary side. In the adaptive finite element context, the refinement region is defined as a set of triangles S_{ref} of the current triangulation (not necessarily connected) where the error of the finite element solution is too big to be acceptable [2]. The idea is to exploit the knowledge one has of the reference triangulation for working only locally with the refinement region (and some neighboring triangles). The new points introduced in the mesh are midpoints of the longest edge of (at least) one triangle of the reference mesh or of an intermediate nested mesh.

Original pure longest edge refinement algorithm (algorithm 1 in reference [12]) deals with intermediate non conforming triangulations. Revised Lepp-bisection

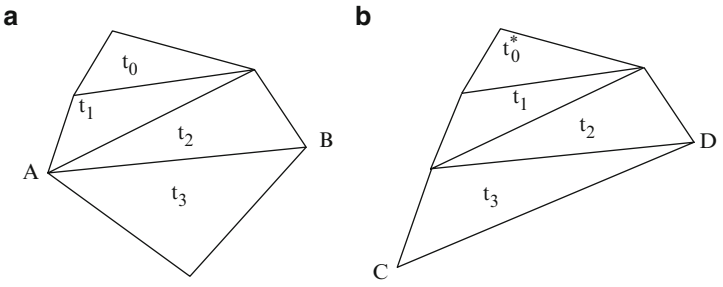


Fig. 4 (**a**) AB is an interior terminal edge shared by terminal triangles (t_2, t_3) associated to Lepp(t_0) $=\{t_0, t_1, t_2, t_3\}$; (**b**) CD is a boundary terminal edge with unique terminal triangle t_3 associated to Lepp$(t_0^*) = \{t_0^*, t_1, t_2, t_3\}$

algorithms only deal with conforming triangulations and very local refinement operations using the Lepp and terminal edge concepts [15, 16, 19].

Definition 2. An edge E is called a terminal edge in triangulation τ if E is the longest edge of every triangle that shares E, while the triangles that share E are called terminal triangles. For any triangle t_0 in τ, the longest edge propagating path of t_0, called *Lepp(t_0)*, is the ordered sequence of increasing triangles $\{t_j\}_0^{N+1}$, where t_j is the neighbor triangle on a longest edge of t_{j-1}, and longest-edge (t_j) > longest-edge (t_{j-1}), for $j = 1, ...N$. The associated terminal edge is either edge E = longest-edge(t_{N+1}) = longest-edge(t_N) if E is an interior terminal edge, or E = longest-edge (t_{N+1}) if E is a boundary edge. See Fig. 4 for an illustration.

The algorithm can be simply described as follows: for each triangle t in S_{ref}, we find Lepp(t), a pair of terminal triangles t_1, t_2 and associated terminal edge l. Then the longest edge bisection of t_1, t_2 is performed by the midpoint of l. The process is repeated until t is destroyed (refined) in the mesh. Figure 5 illustrates the point insertion process.

Lepp-Bisection Algorithm
Input : a quality triangulation, τ, and a set S_{ref} of triangles to be refined
for each t in S_{ref} **do**
  while t remains in τ **do**
    Find Lepp (t), terminal triangles t_1, t_2 and terminal edge l. Triangle t_2 can be null for boundary l.
    Select Point (P, t_1, t_2, l)
    Perform (longest edge) bisection by P of triangles t_1, t_2
    Update S_{ref}
  end while
end for

The properties of these algorithms follow from the results of Sect. [12, 17]:

Lemma 2. *(a) The iterative and arbitrary use of the algorithms only produces triangles whose smallest interior angles are always greater than or equal to* $\alpha/2$,

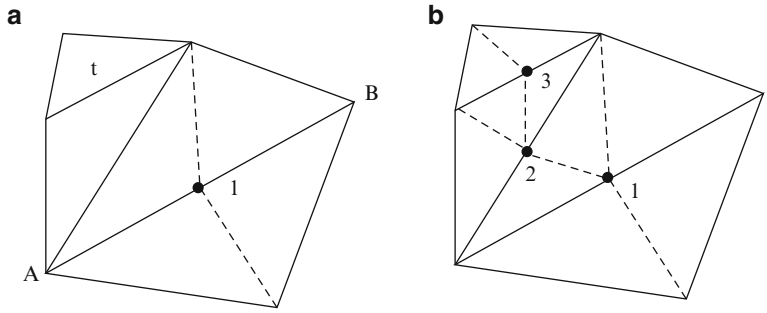


Fig. 5 (**a**) For refining triangle t, a first vertex 1 is added by bisection of the terminal triangles sharing AB. (**b**) Final triangulation obtained for refining t (points added in the creation order)

where α is the smallest interior angle of the initial triangulation. Furthermore every new triangle is similar to one of a finite number of reference triangles.
(b) Longest-edge refinement algorithms always terminate in a finite number of steps with the construction of a conforming triangulation.
(c) Any triangulation τ generated by means of the iterative use of the algorithms satisfies the following smoothness property: for any pair of side-adjacent triangles $t_1, t_2 \in \tau$ (with respective longest edges h_1, h_2) it holds that $\frac{min(h_1,h_2)}{max(h_1,h_2)} \geq k > 0$, where k depends on the smallest angle of the initial triangulation.
(d) For any triangle t, the global iterative application of the algorithm (the bisection of all the triangles in the preceding iteration) covers, in a monotonically increasing form, the area of t with triangles of region I in Fig. 3.

The algorithms have the following advantages: (1) A very local refinement operation, which guarantee that the mesh is conforming throughout the whole refinement process, is repeatedly used; (2) The algorithms are free of non-robustness issues, since they do not depend of complex computations, and the selected points are midpoints of existing previous edges; (3) Refinement / derefinement algorithms able to selectively refine and/or derefine the mesh in the course of computations [14, 18], as well 3-dimensional algorithms [11, 18] can be developed. The algorithms have been also parallelized to deal with the parallel refinement of huge meshes [3, 9].

In 3-dimensions, Lepp(t_0) has a variable number of associated terminal-edges. This is due to the fact that every tetrahedron t in Lepp(t_0) has a finite, non fixed number of neighbor tetrahedra sharing the longest edge of t. Thus in the general case, more than one of these tetrahedra has longest edge greater than the longest edge of t, which implies that the Lepp searching task is multidirectional. Note however that even when the algorithms have been successfully used in practice in 3D [11, 18], a theory on (longest edge) bisection in 3-dimensions such as that presented in [6] for 2-dimensions has not been yet developed. However, Flavio Gutierrez in [7] has proved that a finite number of non similar tetrahedra is obtained for the equilateral tetrahedron for longest edge symmetric bisection.

Finally note that longest edge algorithms applied to meshes of isosceles right triangles have great advantages: (1) The refinement algorithm produces only isosceles right-triangle meshes since this right triangle is a special type I triangle in Fig. 3. (2) The computation of the longest edge is avoided since the refinement is performed by the newest vertex. (3) A right triangle bintree hierarchy is a multiresolution representation that uses a special case of the Lepp-bisection refinement method that takes advantage of both a triangle bintree hierarchy and of the distribution of the grid data. This kind of meshes has been studied for terrain applications and real time visualization [4, 5].

Preliminary cost analysis This requires of an amortized cost analysis for the iterative use of the algorithm [17]. To simplify we consider the following two problems.

(P1) **Vertex refinement problem:** Iteratively refine the mesh around a vertex Q until the adjacent triangles have longest edge less than or equal to a parameter δ.

(P2) **Circle area refinement:** Iteratively refine the triangles that intersect a circular region R_c until every triangle in R_c has longest edge less than or equal to δ.

Lemma 3. *Fractal Property. For any vertex Q, after a finite number of iterations to repeatedly refine each triangle of vertex Q, a fixed angle molecule is obtained (the angles of vertex Q are not partitioned if the refinement follows). In addition further refinement around Q reproduces the same geometry.*

Lemma 4. *(a) For solving (P1), a finite number of points N needs to be added to the mesh, by longest edge bisection of pairs of terminal triangles, where* $N < K(Log(L/\delta))$, *K is a constant such that* $K = 2\pi/\alpha$, α *is smallest angle in the initial mesh and L is the longest interior distance in the polygonal geometry D measured over the smallest rectangle that contains D.*
(b) For solving (P2), finite number of points N_i *and* N_e *need to be respectively added in the interior and the exterior of* R_c *where* $N_i < K_1((\frac{r}{\delta})^2)$, $N_e < K_2(\frac{r}{\delta})Log(\frac{L}{\delta})$, *L is equal to the longest distance from the boundary of* R_c *to the boundary of D, r is the radius of* R_c, *and the constants are* $K_1 = 4\pi$ *and* $K_2 = 2\pi$.

Proof. (Sketch) (a) Consider the worst case where a vertex Q is shared by $2\pi/\alpha$ triangles and assume that one of these triangles has longest edge E of vertex Q and length equal to L. Then E and its sons of vertex Q are iteratively refined by binary partition until an edge son of vertex Q and longest edge less than or equal to δ is obtained. This implies that $Log(L/\delta)$ points are introduced [17]. Then the number of points inserted in the mesh is roughly bounded as: $N < \frac{2\pi}{\alpha}Log(L/\delta)$

(b) To bound N_i assume that right isosceles triangles of longest edge equal to δ (and area $\delta^2/4$) are generated inside R_c. This implies that the area of R_c is covered by $4\pi r^2/\delta^2$ triangles and $N_i < K_i(r/\delta)^2$ with $K_i = 4\pi$. For computing a bound on N_e note that the perimeter $2\pi r$ of R_c is covered by $2\pi r/\delta$ points, and since radially at most $Log(L/\delta)$ points are inserted, it follows that $N_e < 2\pi(r/\delta)Log(L/\delta)$

Lemma 5. *(a) For solving (P1), the algorithm is linear in N defined in Lemma 8.*
(b) For solving (P2), the algorithm is linear in $(N_i + N_e)$, *the number of points inserted in the mesh. In addition if* $r >> \delta$, *then the algorithm is linear in* N_i.

References

1. Adler, A.: On the bisection method for triangles. Math. Comp. **40**, 571–574 (1983)
2. Babuska, I., Zienkiewicz, O.C., Gago, J., Oliveira, E.R. de A. (Eds.): Accuracy estimates and adaptive refinements in finite element computations. Wiley, New York (1986)
3. Castaños, J.G., Savage, J.E.: Pared: a framework for the adaptive solution of pdes. In: 8th IEEE International Symposium on High Performance Distributed Computing (1999)
4. Evans, W., Kirpatrick, D., Townsend, G.: Right-triangulated irregular networks. Algorithmica **30**, 264–286 (2001)
5. Gerstner, T.: Multiresolution compression and visualization of global topographic data. GeoInformatica **7**, 7–32 (2003)
6. Gutierrez, C., Gutierrez, F., Rivara, M.C.: Complexity on the bisection method. Theoret. Comput. Sci. **382**, 131–138 (2007)
7. Gutierrez, F.: On the longest edge bisection of the regular tetrahedron. Personal Communication (2003)
8. Jones, M.T., Plassman, P.E.: Computational results for parallel unstructured mesh computations. Comput. Syst. Eng. **5**, 297–309 (1994)
9. Jones, M.T., Plassmann, E.: Adaptive refinement of unstructured finite element meshes. Finite Elem. Anal. Des. **25**, 41–60 (1997)
10. Kearfott, B.: A proof of convergence and an error bound for the method of bisection in R^n. Math. Comp. **32**, 1147–1153 (1978)
11. Muthukrishnan, S.N., Shiakolas, P.S., Nambiar, R.V., Lawrence K.L.: Simple algorithm for adaptative refinement of three-dimensional finite element tetrahedral meshes. AIAA J. **33**, 928–932 (1995)
12. Rivara, M.C.: Algorithms for refining triangular grids suitable for adaptive and multigrid techniques. Int. J. Numer. Meth. Eng. **20**, 745–756 (1984)
13. Rivara, M.C.: Design and data structure for fully adaptive, multigrid finite-element software. ACM Trans. Math. Software, **10**, 242–264 (1984)
14. Rivara, M.C.: Selective refinement/derefinement algorithms for sequences of nested triangulations. Int. J. Numer. Meth. Eng. **28**, 2889–2906 (1989)
15. Rivara, M.C.: New mathematical tools and techniques for the refinement and/or improvement of unstructured triangulations. Proc. 5th Int. Meshing Roundatable, Pittsburgh, 77–86 (1996)
16. Rivara, M.C.: New longest-edge algorithms for the refinement and/or improvement of unstructured triangulations. Int. J. Numer. Meth. Eng. **40**, 3313–3324 (1997)
17. Rivara, M.C.: Lepp-bisection algorithms, applications and mathematical properties. Appl. Numer. Math. **59**, 2218–2235 (2009)
18. Rivara M.C., Levin, C.: A 3D refinement algorithm suitable for adaptive and multigrid techniques. Commun. Appl. Numer. Meth. **8**, 281–290 (1992)
19. Rivara, M.C., Palma, M.: New LEPP algorithms for quality polygon and volume triangulation: Implementation issues and practical behavior, In: Trends in Unstructured Mesh Generation, Canann, S., Saigal, S. (Eds.), AMD 220, 1–8 (1997)
20. Rivara, M.C., Hitschfeld, N., Simpson, R.B.: Terminal edges Delaunay (small angle based) algorithm for the quality triangulation problem. Comput. Aided Des. **33**, 263–277 (2001)
21. Rosenberg, I.G., Stenger F.: A lower bound on the angles of triangles constructed by bisecting the longest side. Math. Comp. **29**, 390–395 (1975)
22. Stynes, M.: On faster convergence of the bisection nethod for certain triangles. Math. Comp. **33**, 1195–1202 (1979)
23. Stynes, M.: On faster convergence of the bisection method for all triangles. Math. Comp. **35**, 1995–1201 (1980)

Simulation of Spray Painting in Automotive Industry

Robert Rundqvist, Andreas Mark, Björn Andersson, Anders Ålund, Fredrik Edelvik, Sebastian Tafuri, and Johan S Carlson

Abstract Paint and surface treatment processes in the car paint shop are to a large extent automated and performed by robots. Having access to tools that incorporate the flexibility of robotic path planning with fast and efficient simulation of the processes is important to reduce the time required for introduction of new car models, reduce the environmental impact and increase the quality. The combination of high physical complexity, large moving geometries, and demands on near real time results constitutes a big challenge. We have developed an immersed boundary octree flow solver, IBOFlow, based on algorithms for coupled simulations of multiphase and free surface flows, electromagnetic fields, and particle tracing. The solver is included in an in-house package for automatic path planning, IPS. The major improvement of computational speed compared to other approaches is partly due to the use of grid-free methods which in addition simplifies preprocessing.

1 Introduction

Industrial car painting is a highly automated and in many aspects efficient process. To improve efficiency further accurate prediction and optimization through simulation of the key paint shop processes are required. The combination of high physical complexity, large moving geometries, and demands on near real time simulation results constitutes a big challenge. The main processes where accurate modelling can substantially improve efficiency are spray painting, sealing, electro coating and oven curing.

In this work, spray painting is considered. In spray painting paint primer, colour layers and clear coating are applied through either classical Pneumatic spray guns or using the more recent Electrostatic Rotary Bell Sprayer (ERBS) technique. The focus here is on the ERBS technique, where paint is injected at the centre of a

R. Rundqvist (✉), A. Mark, B. Andersson, A. Ålund, F. Edelvik, S. Tafuri, and J.S. Carlson
Fraunhofer-Chalmers Centre, Göteborg, Sweden
e-mail: robert.rundqvist@fcc.chalmers.se

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_83,

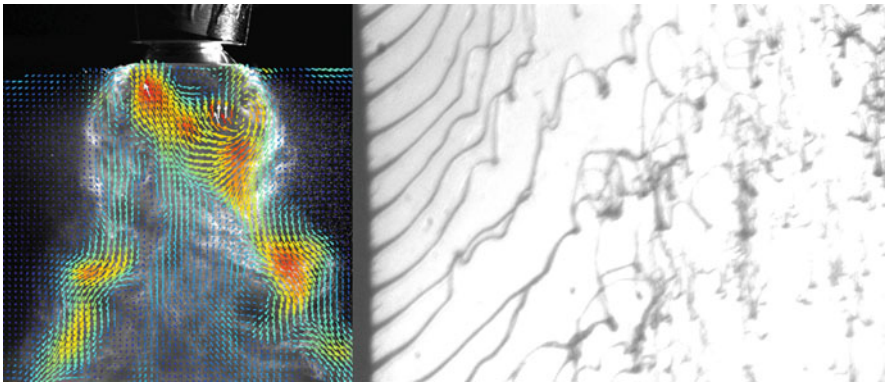


Fig. 1 Applicator and atomization: The *left* part of the figure shows a PIV measurement of the instantaneous droplet velocity field superimposed on an image of the spray. The ERSB applicator is visible on *top*, the target plate is located *below* and just outside the picture. The *right* hand part of the figure is a shadow image of the atomization of paint; at the leftmost side of this picture the edge of the bell cup is visible. From the bell cup ligaments of liquid paint emanate and are broken up into droplets. (Courtesy of Volvo Car Corporation)

rotating bell; the paint forms a film on the bottom side of the bell and is atomized at the edge of the bell. The droplets are normally charged eletrostatically and driven towards the target car body both by shaping air surrounding the rotating bell and by a potential difference in the order of 50–100 kV between paint applicator and target. Software modules for prediction of the electrostatic field and computation of electrostatic forces on the droplets have been developed, but in this paper the focus will be on the non-charged application of paint. A schematic image of the applicator and a shadow image of the paint atomization are shown in Fig. 1.

There exist a few approaches to the simulation problem of electro-static spray paints. Elwood and Braslaw laid out the fundamental physics of the ERBS problem in 1998 [1], where the flow moderately close to the bell was resolved using a finite element approach. There was however no detailed resolution neither of a target nor of the behavior close to the bell. The latter was treated by starting the simulations at a small distance from the bell, using experimental input as boundary conditions. Huang and Lai [2], and Im et al. [3] have since then deepened the understanding of the physics involved and of possible modelling approaches. In their work they present thorough simulations of the transfer process, but only towards a flat target and without any robotic motion.

Ye et al. [4–7] have made extensive modelling work on the electrostatically coupled problems of ERBS and powder coating devices. The simulation model builds on the commercial flow solver Fluent, and incorporates advanced geometries as well as the most relevant physical phenomena present in the real paint shop. The method however lacks in flexibility with respect to defining robot paths and applicator conditions; the computational time which is in the order of a week for a single paint stroke is also a drawback.

The driving force for the development described in this paper is the demand that the resulting model package should be flexible enough to operate automatically and fast enough to run on a desktop machine. Meshing and remeshing of the fluid volume must either be completely automatic or avoided. The model solution must also be able to cope with general geometries moving in transient processes.

2 Modeling

The solution package consists of a particle tracer and a flow solver. Atomization parameters, mainly droplet size and velocity at the applicator, serve as inlet conditions and are determined from user set process conditions such as applicator rotation speed, shaping air flow and paint flow. Paint droplets are traced from the applicator to the target using the particle tracing routines, which are two-way coupled to the flow solver.

2.1 Atomization Model

The region close to the applicator is strongly turbulent and heavily laden with paint droplets. The physics in this region is complex and time-consuming to resolve, but at the same time the flow in this region is relatively independent of the conditions surrounding the applicator. A good approximation of the local conditions is obtainable just by considering the applicator settings and disregarding the far field such as conditions in the spray booth and position of applicator with respect to the target geometry. In this work, droplet size distribution has been measured as a function of paint flow and bell rotation speed using shadow imaging techniques. Droplet and air velocity distributions close to the applicator have been measured as a function of shaping air flow, paint flow and bell rotation speed. These measurements have been used to build approximating functions for a span of operating conditions; that is sets of the parameters: paint flow, air flow and bell rotation speed.

2.2 Flow Solver

IBOFlow is an incompressible finite-volume based fluid flow solver. The fluid flow is governed by the Navier–Stokes' Equations,

$$\frac{\partial u_j}{\partial x_j} = 0 \tag{1a}$$

$$\rho_f \frac{\partial}{\partial t}(u_i) + \rho_f u_j \frac{\partial u_i}{\partial x_j} = -\frac{\partial p}{\partial x_i} + \frac{\partial}{\partial x_j}\left(\mu \frac{\partial u_i}{\partial x_j}\right) + \rho_f g_i, \tag{1b}$$

where u_i is the fluid velocity, ρ_f is the density, p is the pressure, μ represents viscosity of the fluid, and g_i is the gravitational acceleration. The velocity and pressure fields are coupled with the SIMPLEC [8] method and discretized on a Cartesian octree grid that can be dynamically refined and coarsened, enabling grid refinement to move with moving objects with almost no extra computational cost. The variables are stored in a co-located configuration and pressure weighted flux interpolation [9] is employed to prevent pressure oscillations.

Moving and interacting arbitrary bodies (robots or cars) inside the fluid are handled by the mirroring immersed boundary method [10]. The method models the presence of the bodies by an immersed boundary condition, which mirrors the velocity field over the boundary of the body such that the fluid exactly follows the surface of the body. As a result, a fictitious velocity field inside the body is developed, which is excluded in the continuity equation to ensure zero mass flux over the boundary. The method facilitates the treatment of moving and interacting objects in fluid flows and simplify the meshing procedure by only requiring surface descriptions of the flow boundaries to run a simulation.

2.3 Particle Tracer and Thickness Integration

As it would be too computationally expensive to resolve the flow on the droplet level, the interaction between air flow and fluid droplet is described using one ODE per particle. This ODE is determined by considering the droplets as point-like objects and approximating the fluid forces by including only drag, added mass and gravity/buoyancy from the Basset–Boussinesq–Oseen (BBO) equation [11],

$$\rho_p \frac{du_{pi}}{dt} = \left(\rho_p - \rho_f\right) g_i - \frac{18\mu}{d_p^2}\left(u_{pi}\left(t\right) - u_i\Big|_{x_p(t)}\right) - \frac{1}{2}\rho_f \frac{d}{dt}\left(u_{pi}\left(t\right) - u_i\Big|_{x_p(t)}\right) \tag{2}$$

where ρ_p is the particle density, u_{pi} is the particle velocity, and d_p is the particle diameter. This is deemed sufficient as these forces are strongly dominating in this flow scenario. The BBO equation is discretized in time through a Runge–Kutta method of order four and five. The solution of the fluid and the particle flow is two-way coupled, meaning that in all fluid cells the droplet forces are evenly distributed over the volume and fed back into the momentum equation for the continuous phase. To save computational effort, the solution is only computed on a representative distribution of paint droplets. That is, each droplet traced is multiplied by a cloud factor in the coupling to the air flow solution and in the paint thickness integration.

As the source terms from the particles in the momentum equations are distributed over the computational cells, there is a connection between the spatial resolution of fluid phase in terms of cell size and of the particle phase in terms of the cloud factor. If the cloud factor is too small, there will be unnecessarily many particles in each cell, leading to long simulation times. If the cloud factor is too large, on the other

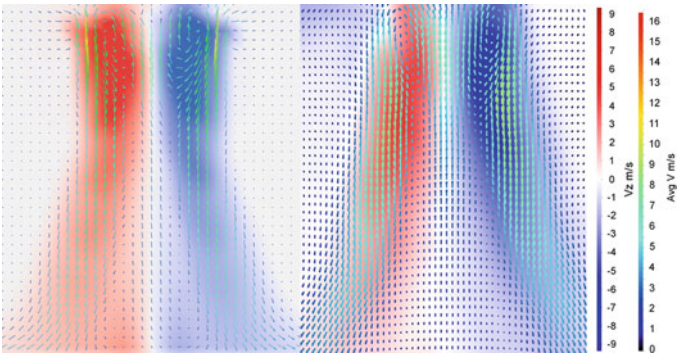


Fig. 2 Static velocity field: To the *left* is a cutting plane of the simulated air and paint velocity field and to the *right* is the same cutting plane measured experimentally. In both cases the out of plane velocity is colored in *red* and *blue*, and the in-plane velocity is illustrated with *arrows*. The scales are the same in both plots

hand, the fluid flow resolution will be too high with respect to the momentum source terms, which will then be fluctuating between higher values in the cells where there are numerical particles present and zero where there are no particles present.

3 Results

The flow solution of paint and air was first compared directly to PIV measurements of a static applicator positioned over a flat test plate. Compared velocity fields for a plane just underneath the applicator can be seen in Fig. 2.

The required grid resolution level was determined in a grid refinement study, where the air velocity field across the centreline was compared for the different solutions. As can be seen in Fig. 3, reasonable grid independence for the flow solution above the test plate is obtained at around 100,000 grid cells.

Resolution independence is required not only in the size of the grid cells, but also in the number of computational droplets. This is verified by comparing paint thickness profiles for different values of the cloud factor, seen in Fig. 4. The results indicate that a cloud factor of around 400 is sufficient for good accuracy.

The main comparison with experiments has been done on flat test plates. These plates measure 200 by 600 mm and are commonly used to check applicator settings and paint film build up in the factory paint shop. The plates are painted using a single paint stroke and the film thickness is measured along the long centreline of the test plate, across the applicator movement direction. Illustration of the test set-up is shown in Fig. 5. Experimental and numerical results for three different applicator settings are compared in Fig. 6. The agreement between simulation and experiment is good, both in the case of normal process parameter settings and in

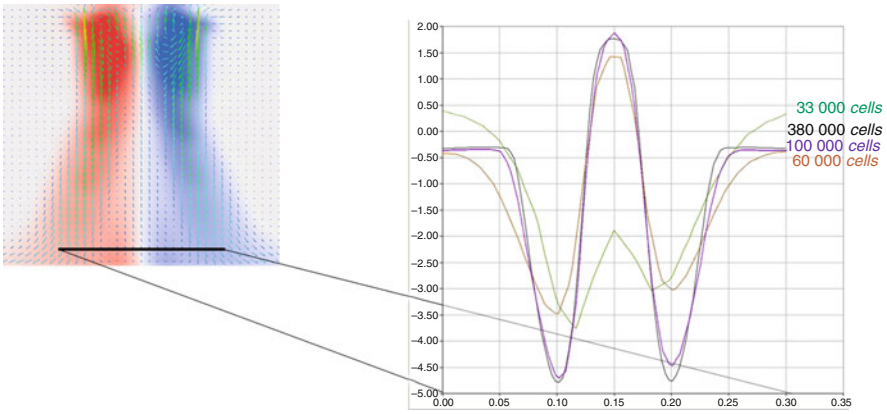


Fig. 3 Grid dependency: Comparison of simulation results for different levels of resolution. The downward air velocity is compared at a level of 20 mm above the target, evaluated at the 600 mm long centre line of the target plate. The resolution level corresponding to 100,000 cells is a good compromise of speed and accuracy

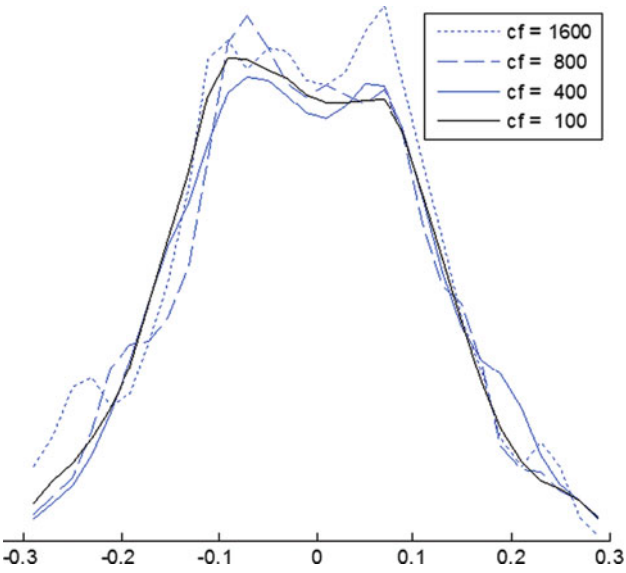


Fig. 4 Cloud factor dependency: Comparison of simulation results for different settings of the cloud factor. The curves used for the comparison is the film thickness along the 600 mm long centre line of the target plate, evaluated after simulating a standard paint stroke with the ERSB

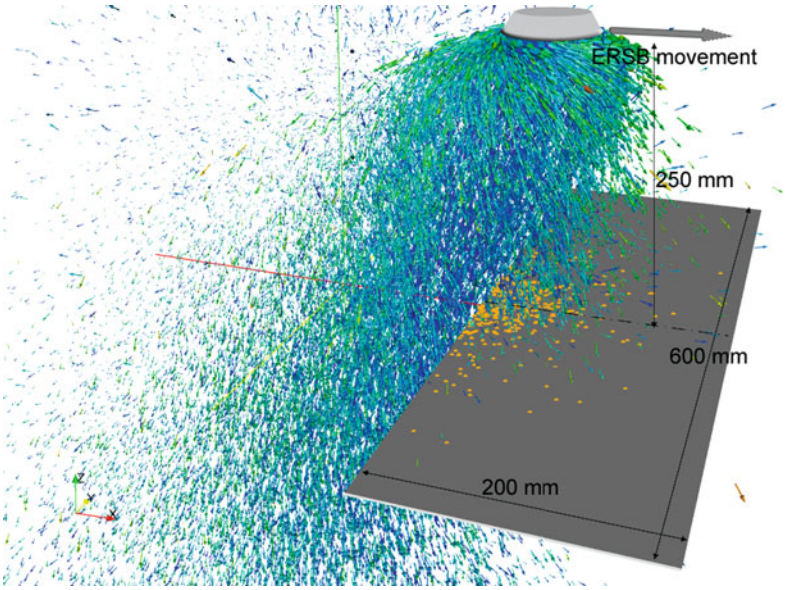


Fig. 5 Test plate geometry: Illustration of the test geometry and the standard test plate. Shown in the picture is also instantaneous velocities and impact points for simulated paint droplets

the robustness test performed with increased flow and rotation speed. This indicates that the necessary model simplifications described in Sect. 2 are justified.

Figure 7 shows an example of simulated painting in IPS Virtual Paint, where IBOFlow has been integrated in the path planning tool, IPS. In this case, three ERBS applicators with high voltage switched on simultaneously paint the trunk lid of a car. From the resulting colouring of the trunk lid it can be seen that some adjustment of the applicator paths is needed to ensure full coverage of the target – for instance a small region surrounding the hole for the rear lights is not covered with paint. Simulation time for the 6.0 s of physical time required to reach the end result was 90 min using a single processor on a standard laptop computer. Validating the virtual painting of this type of geometries is difficult, but qualitative comparison with real car painting is good, both in the film thickness taken from the end result and in the general behavior of the paint brush which can be observed during the process simulation.

4 Conclusions

From the results it can clearly be concluded that fast and accurate simulation of a complex industrial process like spray painting of cars is possible. Agreement with experimental data for a test plate is excellent when operating under normal

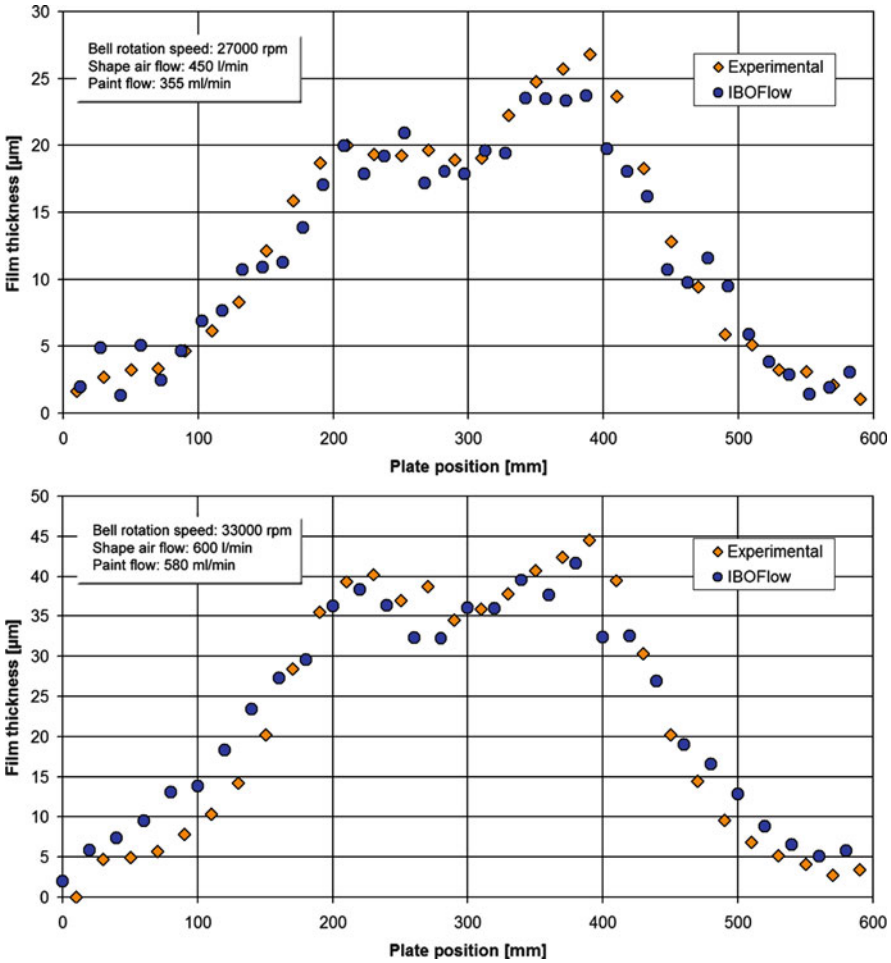


Fig. 6 Thickness profile comparison: Comparison between experiments and simulation of the paint thickness at the centre line of the test plate. The correspondence is exceptionally good in the standard test case and good in the case with increased flow

process conditions. The computational time for a full scale industrial application like painting the trunk lid of a car is in the order of an hour. Therefore the simulation tool can be useful for paint prediction and trouble shooting in the design stage as well as in off-line programming of paint robots. Future work includes simulation of the paint droplet break-up process and further comparisons with experiments using more complex geometries and more parameter combinations.

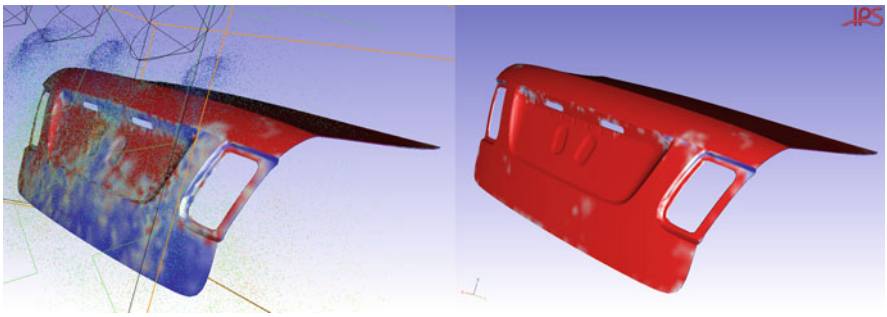


Fig. 7 Complex simulation: Painting of a car trunk lid using three applicators. Colors on the geometry surface show the paint film thickness with *blue* being un-painted and *red* indicating a film thickness of at least 100 μm. The *left* hand image shows the paint process in action after 3.0 s of the simulation, and the *right* hand image shows the result after the full 6 s (CAD geometry courtesy of Saab Automobile AB)

Acknowledgements This work was supported by the Swedish Governmental Agency for Innovation Systems, VINNOVA, through the FFI Sustainable Production Technology program. The authors are also grateful for the valuable support of geometry models and experimental measurement data from our industrial collaborating partners Volvo Car Corporation and Saab Automobile AB.

References

1. Elwood, K.R.J., Braslaw, J.: A finite-element model for an electrostatic bell sprayer. J. Electrostat. **45**, 1–23 (1998)
2. Huang, H., Lai, M.-C., Meredith, W.: Simulation of Spray Transport from Rotary Cup Atomizer using KIVA-3V. In: 10th Int. KIVA Users Group Meeting, Detroit (2000)
3. Im, K.-S., Lai, M.-C., Yu, S.-T.J., Matheson Jr, R.R.: Simulation of spray transfer processes in electrostatic rotary bell sprayer. J. Fluids Eng. **126**, 449–456 (2004)
4. Ye, Q., Scheibe, A.: Unsteady numerical simulation of electrostatic spray-painting processes with moving atomizer. In: 13th Int. Coating Science Tech. Symp., Denver (2006)
5. Ye, Q., Domnick, J., Scheibe, A.: Numerical simulation of spray painting in the automotive industry. In: Euro. Auto. CFD Conf., Bingen (2003)
6. Domnick, J., Scheibe, A., Ye, Q.: The simulation of the electrostatic spray painting process with high-speed rotary bell atomizers. Part I: Direct charging. Part. Part. Syst. Char. **22**, 141–150 (2005)
7. Domnick, J., Scheibe, A., Ye, Q.: Numerical Simulation of the Behaviour of Sprays Produced by High Speed-Rotary Bells with External Charging. In: ILASS-Europe, Zurich, Bingen (2001) Rizzi, A.A., Choset, H.: Paint deposition modeling for trajectory planning on automotive surfaces. IEEE Trans. Auto. Sci. Eng. **2**, 381–392 (2005)
8. Van Doormaal, J.P., Raithby, G.D.: Enhancements of the SIMPLE method for predicting incompressible fluid flows. Numer. Heat Tran. **7**, 147–163 (1984)
9. Rhie, C.M., Chow, W.L.: Numerical study of the turbulent flow past an airfoil with trailing edge separation. AIAA J. **21**, 1527–1532 (1983)
10. Mark, A., van Wachem, B.G.M.: Derivation and valdiation of a novel implicit second-order accurate immersed boundary method. J. Comput. Phys. **227**, 6660–6680 (2008)
11. Maxey, M.R., Riley, J.J.: Equation of motion for a small rigid sphere in a nonuniform flow. Phys. Fluids **26**(4), 883–889 (1983)

Numerical Simulation of the Electrohydrodynamic Generation of Droplets by the Boundary Element Method

P. Sarmah, A. Glière, and J.-L. Reboud

Abstract A numerical simulation of the formation of droplets from an electrified capillary using the Boundary Element Method (BEM) is presented. An incompressible and perfectly conducting liquid is injected from a capillary into a dynamically inactive and insulating gas. Assuming an irrotational liquid flow, the problem consists in coupling the BEM resolution of two Laplace equations: for the velocity potential inside the fluid domain and for the electric potential in the capacitor gas gap. The motion of the free surface is determined by the Bernoulli's equation resulting from the normal stress balance on the free surface.

1 Introduction

In electrospray devices, tiny droplets of analyte are ejected at the tip of a capillary in the presence of an electric field. This electrohydrodynamic method of production of droplets is widely used as it often constitutes the input stage of mass spectrometers. It can also be found in various other domains, such as spray coating, inkjet printing or spraying of agricultural chemicals. The liquid is injected through a capillary at a low flow rate in the presence of a strong electric field. Due to the contrast in conductivity and permittivity between the liquid and the surrounding gas, charges build-up on the interface and electric stress appears. Depending of the balance between inertia, viscosity, electric stress and surface tension, several flow regimes, such as dripping and cone-jet modes, can be encountered [4].

In recent years, the increasing demand in biological sample analysis has been accompanied by a trend towards electrospray systems miniaturization. In this context,

A. Glière (✉) and P. Sarmah
CEA, LETI, MINATEC, Grenoble, France
e-mail: alain.gliere@cea.fr, pranjit.sarmah@cea.fr

J.-L. Reboud
G2Elab (CNRS, Grenoble INP, UJF), Grenoble, France
e-mail: jean-luc.reboud@ujf-grenoble.fr

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_84, © Springer-Verlag Berlin Heidelberg 2010

numerical simulation tools are of prime interest to acquire an in-depth understanding of the involved physical phenomena, which helps afterwards improving the micro devices design. In particular, studying the influence of the liquid flow rate, the solvent composition and the nozzle geometry is necessary.

Within the past decades many researchers have adopted the Boundary Element Method (BEM) to address problems involving severe geometrical deformations of free surfaces [2, 5]. In the BEM, the governing partial differential equations are transformed into boundary integral equations. One of the distinct features of the method is that only the bounding surface of the considered domain has to be discretized [1]. The computational cost is thus in principle reduced and, of utmost importance in our context, interfacial effects can be easily and accurately incorporated.

A numerical simulation of the formation of droplets from an electrified capillary using the BEM is presented here. The model has been applied to the simulation of an electrospray nozzle used for laboratory mass spectrometry.

2 Mathematical Formulation

A thin metal capillary (with radius R and vanishingly small wall thickness) is protruded a distance H_1 from the center of the top plate of a circular parallel-plate capacitor. The bottom plate of the capacitor is grounded. The top plate and the capillary are held at a potential U_0 above ground. A dynamically inactive, insulating ambient gas of permittivity ϵ is kept between the two plates of the capacitor. An incompressible and perfectly conductive liquid, of density ρ and dynamic viscosity μ, is injected through the upper end of the capillary with a constant flow rate. The flow rate is given by $\pi R^2 \tilde{v}$, where $\tilde{v}$ is the dimensional average inlet velocity. The liquid-gas interface is submitted to surface tension σ. The capillary and the capacitor share a common axis of symmetry in the direction of gravity. The capillary radius R, capillary time scale $\sqrt{\rho R^3/\sigma}$ and top-plate potential U_0 are used, as characteristic length, time and electric potential, to non-dimensionalise the equations printed in the rest of the text. The axisymmetric geometry of the EHD spraying setup is presented in the Fig. 1. It is bounded by a cylindrical surface S_∞ of radius $R_\infty \gg 1$.

The fluid inside the domain $V(t)$, consisting of the interior of the drop and the capillary, is considered to undergo irrotational motion. Thus the fluid velocity can be expressed as the gradient of a scalar potential ϕ, $\mathbf{u} = -\nabla\phi$. It follows from the mass continuity equation for incompressible fluids that the velocity potential obeys Laplace's equation

$$\nabla^2\phi = 0 \tag{1}$$

The fluid is considered as perfectly conducting and is thus equipotential. The electric field is given in the ambient gas domain $V'(t)$ by the gradient of the scalar potential ϕ_e, $\mathbf{E} = -\nabla\phi_e$. Thus, using the Maxwell–Gauss equation, the electric potential is

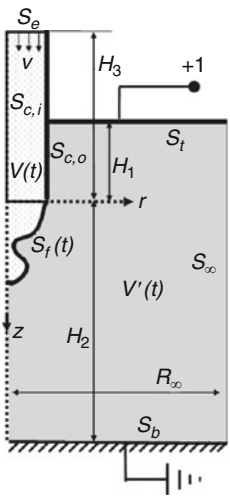


Fig. 1 Geometrical model of the EHD spraying setup

governed by Laplace's equation

$$\nabla^2 \phi_e = 0 \tag{2}$$

The fluid dynamics equation (1) is solved subject to the boundary conditions

$$\mathbf{n}.\nabla\phi = -v \ \text{ on } S_e \ ; \quad \mathbf{n}.\nabla\phi = 0 \ \text{ on } S_{c,i} \tag{3}$$

The initial shape of the droplet is assumed to be a spherical cap. The initial value of the free surface potential is derived from mass conservation of the fluid domain

$$\phi = \frac{1}{2}(H_3 + z) \ \text{ on } S_f(0) \tag{4}$$

The boundary conditions for the electrical equation (2) are given by

$$\phi_e = 1 \ \text{ on } S_t \cup S_{c,o} \cup S_f(t) \ ; \quad \phi_e = 0 \ \text{ on } S_b \ ; \quad \mathbf{n}.\nabla\phi_e = 0 \ \text{ on } S_\infty \tag{5}$$

3 Boundary Element Method Formulation

The mathematical basis of the numerical method used in this paper is described in details by Canot et al. [3]. The boundary integral representation of Laplace's equation for axisymmetric problems is [1]

$$\phi(\mathbf{x}_0) = -\frac{2\pi}{\alpha}\int_S q(\mathbf{x})G^{AX}(\mathbf{x},\ \mathbf{x}_0)r(\mathbf{x})ds + \frac{2\pi}{\alpha}\int_C \phi(\mathbf{x})(\mathbf{n}.\nabla G^{AX})(\mathbf{x},\ \mathbf{x}_0)r(\mathbf{x})ds \tag{6}$$

where $q = \mathbf{n}.\nabla\phi$, α is the aperture angle and G^{AX} is the axisymmetric Green's function. The pole $\mathbf{x_0}$ lies on the boundary S. Alternating the location of the evaluation point $\mathbf{x_0}$ of (6) on a discrete domain provides a linear set of equations relating ϕ (or ϕ_e) and q (or $q_e = \mathbf{n}.\nabla\phi_e$) values at all nodes on the boundary. Linear order boundary elements are used to approximate the potentials and normal fluxes. Boundary integrals are performed using gaussian quadrature with 4–24 points along each linear segments. The number of points of the gaussian quadrature increases as the source point comes closer to the evaluation point. The interface is assumed to be material, so the nodes on the free surface are "tracked" along their instantaneous velocity vector [3]:

$$\frac{Dz}{Dt} = \frac{\partial\phi}{\partial z}, \quad \frac{Dr}{Dt} = \frac{\partial\phi}{\partial r} \tag{7}$$

The transient evolution of the velocity potential is obtained by combining Bernoulli's equation with the normal momentum balance at the interface:

$$\frac{D\phi}{D\,t} = \kappa + \frac{1}{2}|u|^2 + N_e(\mathbf{n}.\nabla\phi_e)^2 + \frac{2}{R_e}\frac{\partial^2\phi}{\partial\, n^2} + Bz \tag{8}$$

The different terms in the right hand side of the non-dimensional equation (8) respectively stand for surface tension, inertia, normal electric stress [6], normal viscous stress [5] and gravitation. In this equation, κ is the dimensionless total curvature, $N_e = \epsilon U_0^2/2R\sigma$ is the electric Bond number, $R_e = \sqrt{\rho\sigma R}/\mu$ is the Reynolds number, $B = gR^2\rho/\sigma$ is the gravitational Bond number. The numerical simulation of electrohydrodynamic spraying is formulated as a transient free boundary problem consisting of two types of calculations:

1. The evolution problem is successively divided into tiny time steps $\triangle t$. At a fixed instant t we solve the Laplace equations (1) and (2) to obtain the corresponding normal component of the electric field and the normal and tangential components of the velocity using the BEM formulation.
2. Equations (7) and (8) are integrated using fourth order explicit Runge–Kutta method to determine the new velocity potential value and interface position at the following instant $(t + \triangle t)$. A Runge–Kutta integration step consists of four BEM solutions for both Laplace equations.

To maintain the stability in the free surface flow simulation, a linear stability algorithm, based on normal modes of the free-surface linearized perturbations has been used. Following Canot et al. [3], an optimal time step is determined at each iteration, in an analytic closed form, by investigation of the spectral radius of the 2×2 amplification matrix.

In order to improve the computational accuracy, a variable number of nodes are unevenly redistributed on the free surface at each time steps in accordance with several criteria like (i) high concentration of nodes at places where the interface curvature is important or where the free surface approaches close to the axis of symmetry, (ii) adaption of the lengths of each elements to the gradient of the

velocity potential and (iii) preservation of a bounded ratio of lengths between two neighboring elements [2].

The free surface is unstable near the end of the capillary and has the tendency to deform faster. Thus an initial high concentration of nodes is necessary to represent the free surface curvature smoothly in this region. Arithmetic progression based redistribution is employed at the solid capillary boundary to keep the node distribution uniform near the capillary tip. In this process, the length of the first element of the boundary $S_{c,o}$ is considered to be equal to the neighboring element of the free surface $S_f(t)$. Then the length of elements in $S_{c,o}$ is gradually increased towards S_e with a constant arithmetic mean value such that the length of the last element of $S_{c,o}$ is equal to the length of the first element of S_e. In summery, the computational algorithm is as follows:

1. Determine the initial geometry and the initial boundary conditions
2. Redistribute the nodes on the free surface $S_f(t)$
3. Redistribute the nodes on the capillary $S_{c,o}$ using arithmetic progression
4. Choose the appropriate time step using time stability algorithm
5. Solve Laplace equations (1) and (2)
6. Advance nodal positions and velocity potential integrating equation (7) and (8)
7. Continue steps 2–6 using the updated free surface geometry and potentials

4 Results

The current BEM based model of EHD spraying has been validated with existing literature, where the Finite Element Method is used for the simulation [6]. In this problem a capillary of radius 1.26 cm, where the fluid is injected with a flow rate of 5 ml/min is used. In Table 1 different non-dimensional parameters like break up location Z_d, primary drop volume V_1, limiting drop length L_d and characteristic time of break up t_d are compared for two values of the electric Bond number $N_e = 10$ and $N_e = 17$ (Z_D and L_D are defined in Fig. 2). It is found that the present algorithm has a good agreement with the published results of Notz and Basaran. Series of simulations have been carried out to determine the effect of flow rate Q and applied electric potential U_0 in the process of EHD spraying. A capillary of radius 30 μm is considered, where a fluid of surface tension 0.073 N m^{-1}, viscosity 10^{-3} Pa.s and density 1,000 kg m^{-3} is injected with a flow rate in the range

Table 1 Validation of the model with results of Notz et al. [6] for $H_1 = 0.0$, $B = 0.35$ and $v = 0.05$

Parameter	BEM ($N_e = 10$)	Notz et al. ($N_e = 10$)	BEM ($N_e = 17$)	Notz et al. ($N_e = 17$)
Z_d	3.21	3.54	4.29	4.3
V_1	9.73	9.91	7.61	7.7
L_d	6.53	6.45	7.84	7.65
t_d	65.7	65.28	54.1	54.3

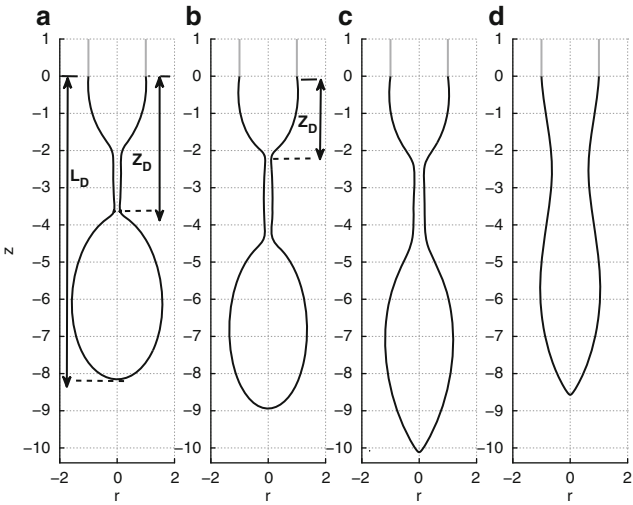


Fig. 2 Dripping mode at the flow rate 20 μl/min and applied electric potential (**a**) 1,300 kV m^{-1}, (**b**) 1,315 kV m^{-1}, (**c**) 1,320 kV m^{-1} and (**d**) 1,325 kV m^{-1}

of [5–80 μl/min]. The non-dimensional distance between the lower electrode and the capillary tip is $H_2 = 167$ and the capillary is protruded a distance $H_1 = 23$ from the upper electrode. Four different EHD spraying modes are observed: *dripping*, *micro-dripping*, *cone-jet* and *multi-jet*. In Fig. 2a–d different types of dripping modes are presented.

In micro scale, the gravitational force has no effect in the process of droplet formation ($B \simeq 10^{-4}$), thus it is only governed by electrostatic pressure, capillary pressure and inertial force. In the process, a neck like shape forms, connected to a spherical mass of liquid at its downstream. The increased capillary pressure at the thin neck gives rise to a large pressure gradient in both the upstream and downstream parts of the neck and flow reversal occurs at the upstream part, near the top end of the neck. The fluid leaving the bottom and the top ends of the neck is accelerated by the capillary pressure, while the presence of inertial force of the inlet fluid velocity creates deceleration near the top end of the neck. When the fluid is injected at high flow rate (cf. Fig. 3a for flow rate 60 μl/min), the neck drains faster at the bottom end and always breaks there. This behavior changes at lower flow rate: when the flow rate is reduced, the deceleration caused by the inertial force in the top end of the neck is also reduced. The presence of stabilizing electrostatic pressure lowers the destabilizing capillary pressure near the bottom end of the neck. Thus, the drop breaks at the top end of the neck when the applied electric field crosses a critical value (cf. Fig. 3a for flow rate 20 μl/min and electric field higher than 1,312 kV m^{-1}).

The dripping mode is observed for applied electric field ≤1,330 kV m^{-1} at flow rate between 5 μl/min and 80 μl/min. The dripping mode is characterized by the regular emission of drops of the same size, the drop radius being of the same order

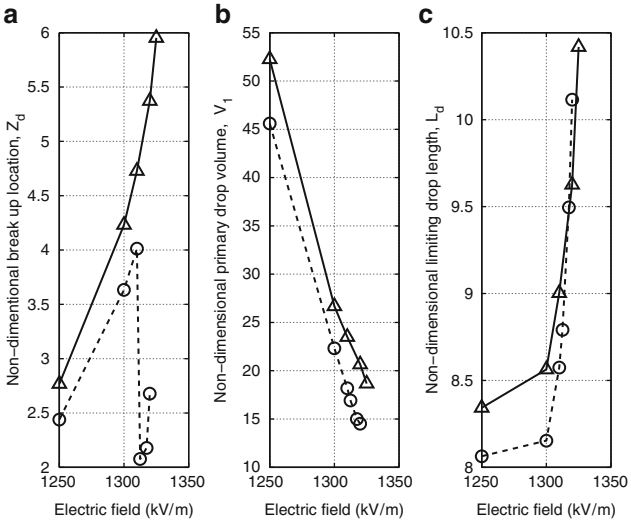


Fig. 3 Variation of different parameters with applied electric field and flow rate (– Δ – 60 μl/min and – o – 20 μl/min), (**a**) break up location (Z_d), (**b**) primary drop volume (V_1), (**c**) limiting drop length (L_d)

as that of the capillary radius. The electric field strength is highest near the tip of the capillary. When the electrostatic pressure increases, the axial acceleration near the tip of the meniscus increases. It elongates the drop in the axial direction and the curvature increases. Thus, the tip of the primary droplet takes a thick jet like shape and produces small droplets just before the detachment of the primary droplet (cf. Fig. 2c–d). The specific case of dripping mode of Fig. 2c is termed by some authors as *dripping+sibbling* mode and that of Fig. 2d is termed as *intermediate cone jet* mode or *jetting* mode [6]. These two modes have been observed for applied electric field in the narrow range between 1,320 kV m^{-1} and 1,330 kV m^{-1}.

We now switch to the study of the drop volume V_1 and the limiting drop length L_D. The electrostatic pressure creates an axial acceleration of the drop near its tip and the thread. As a result, the thread and the drop elongates in the axial direction and the curvature of the drop increases. As a result of the increase in curvature, the size of the primary drop decreases with the increase in applied electric field (cf. Fig. 3b). Also, the limiting length of the drop increases as the length of the thread and aspect ratio of the primary drop increases (cf. Fig. 3c). In Fig. 3b it is seen that at the same applied electric field the primary droplet volume increases with the increase in flow rate as the increased inertial force delays the flow reversal, allowing more fluid to enter inside the drop.

The next mode, observed at the larger values of the applied electric field and at low flow rates, is the microdripping mode (cf. Fig. 4a). It is characterized by regular emission of a small drop formed at the tip of the stable ellipsoidal meniscus. The

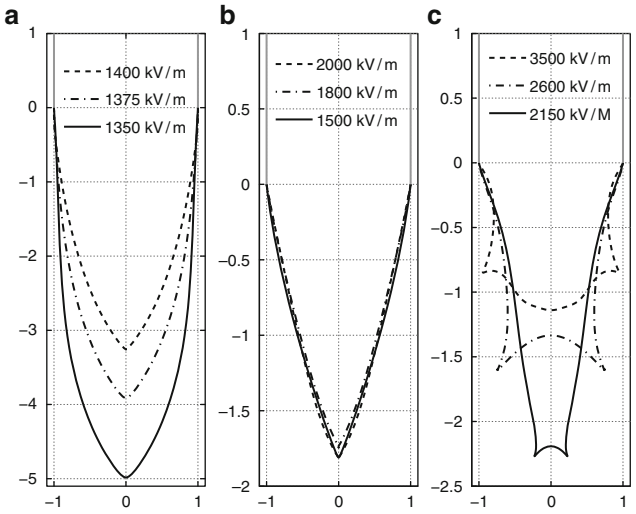


Fig. 4 Different modes at flow rate 10 μl/min, (**a**) Microdripping mode, (**b**) Cone-jet mode, (**c**) Multijet mode

drop radius is about an order of magnitude smaller than the capillary radius. This mode is observed for applied electric field between 1,350 kV m^{-1} and 1,400 kV m^{-1} at flow rate $\leq$40 μl/min. The cone-jet mode is observed on Fig. 4b for applied electric field between 1,500 kV m^{-1} and 2,000 kV m^{-1}. For perfectly conductive fluid, the jet formation regime is limited to the apex of the meniscus and an almost static equilibrium of forces exists at each point [4]. Thus, we are not able to simulate the continuous jet from the apex of the cone due to the absence of tangential electric field. This mode is characterized by the independency of surface charge to the applied electric field and the meniscus profiles remain almost unchanged with the increase of applied electric field. The multi-jet mode (cf. Fig. 4c) is obtained for electric field greater than 2,000 kV m^{-1}. In this case, the meniscus becomes unstable due to the presence of high electrostatic pressure and multiple jets emerge all around the axis of symmetry.

5 Conclusion

A numerical model has been developed to study the Electrohydrodynamic spraying process using the Boundary Element Method. The EHD break up modes observed with the increase in electric field are respectively the dripping, microdripping, cone-jet and multi-jet modes. Microdripping mode is only observed at low flow rates. Consistent with the experimental evidence, we observe that the thread breakup location switches from its top end to its bottom end while the flow rate is decreased and

a critical electric field is reached. To the best of our knowledge the work presented here is the first attempt to simulate the EHD spraying process at micro scale by the BEM.

References

1. Brebbia, C.A., Telles, J.C.F., Wrobel, L.C.: Boundary element techniques – Theory and applications in engineering. Springer, Berlin (1984)
2. Canot, E., Georgescu, S., Achard, J.L.: Bursting air bubble at a free surface – Regridding influence on the interface evolution. In: Workshop on Numerical Simulations for Fluid Mechanics and Magnetic Liquids. Timisoara, Romania (2001)
3. Canot, E., Davoust, L., El Hammoumi, M., Lachkar, D., Blake, J.R.: Numerical simulation of the buoyancy-driven bouncing of a 2-D bubble at a horizontal wall. Theoretical and Computational Fluid Dynamics. **17**(1), 51–72 (2003)
4. Cloupeau, M., Prunet-Foch, B.: Electrohydrodynamic spraying functioning modes – a critical-review. Journal of Aerosol Science. **25**(6), 1021–1036 (1994)
5. Georgescu, S.C., Achard, J.L., Canot, E.: Jet drops ejection in bursting gas bubble processes. European Journal of Mechanics B-Fluids. **21**(2), 265–280 (2002)
6. Notz, P.K., Basaran, O.A.: Dynamics of drop formation in an electric field. Journal of Colloid and Interface Science. **213**(1), 218–237(1999)

A General Pricing Technique Based on Theta-Calculus and Sparse Grids

Stefanie Schraufstetter and Janos Benk

Abstract In [An Introduction to Theta-calculus (2005)], Dirnstorfer introduced the Theta-notation for modeling financial contracts consistently by a sequence of operators. This easy-to-use modeling for financial engineers together with Monte Carlo methods is already applied successfully for option pricing. We combined the idea of Theta-calculus with an approach based on partial differential equations (PDE) to get a higher accuracy. In this paper, we give a short introduction to Theta-calculus and deduce the resulting pricing algorithm that is – in contrast to common PDE based pricing techniques – general and independent from the type of product. With the use of sparse grids, this method also works for higher dimensional problems. Thus, the approach allows an easy access to the numerical pricing of various types of multi-dimensional problems.

1 Introduction

There exist different methods for pricing financial products. The most widespread one is the Monte Carlo method which is very robust and can be applied straightforward. On the other side, the Monte Carlo method only has a low convergence rate, which makes the method inaccurate in the higher-dimensional case. One other type of pricing method, which does not have this disadvantage, is based on a partial differential equation that describes the price change of the financial contract. This underlying PDE problem has to be modeled for each type of product separately. For example, a European option can be modeled simply with a Black-Scholes PDE whereas from the American option an obstacle problem arises, and in case of the Asian option an additional integral term has to be added to the PDE.

Our aim is to develop a toolbox with that different types of options and other financial products such as swaps can be priced simply due to automated pricing

S. Schraufstetter (✉) and J. Benk
Technische Universität München, Boltzmannstr. 3, 85748 Garching
e-mail: schraufs@in.tum.de, benk@in.tum.de

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_85, © Springer-Verlag Berlin Heidelberg 2010

algorithms. The user has just to model the financial product in a script language which is easier to formulate and more natural than the formulation of the mathematical problem itself. For this purpose, we use the script language *ThetaML* which was introduced by Dirnstorfer [2] and is already used for automated option pricing with the Monte Carlo method. To handle the curse of dimensionality in case of multi-dimensional problems, we suggest the sparse grid approach that uses a hierarchical basis and reduces the number of degrees of freedom tremendously.

2 Option Pricing with Theta-Calculus

The main idea of our approach is the splitting of the option pricing problem into a structural and a stochastic model as it is illustrated in Fig. 1. The structural model is formed by the script language *ThetaML* , whereas the stochastic model, a Brownian motion, e.g., is defined separately and, thus, independently from the structural model. In this section, we will focus on the structural model and the evaluation process that results from this.

2.1 Modeling Options with ThetaML

The script language *ThetaML* [2] is a notation for stochastic processes and is compatible with common computer algebra systems. It provides an explicit operator-based representation of financial products. For every activity, *ThetaML* defines an operator $O_i : (\mathbb{R}^n \to \mathbb{R}^m) \to (\mathbb{R}^n \to \mathbb{R}^m)$. These operators can be combined to an operator sequence that has to be evaluated from the right to the left:

$$O_1 O_2 := f \mapsto O_1(O_2(f)). \tag{1}$$

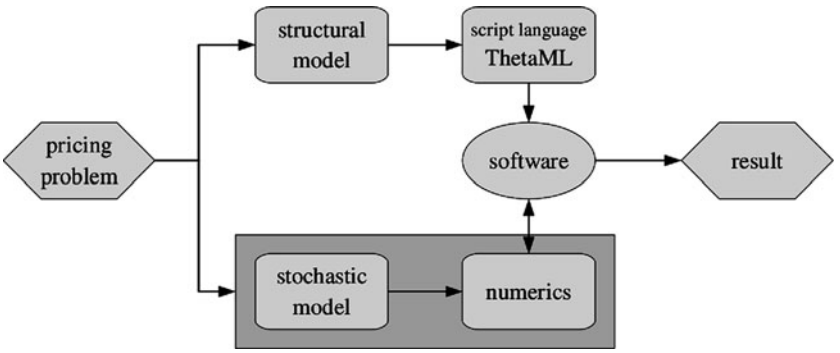


Fig. 1 The idea of option pricing with Theta calculus: The problem is split into a structural and a stochastic model

ThetaML is mainly based on three elementary effects: *waiting*, *transacting*, and *deciding*.

- *Waiting* is modeled with the *Markov process operator*

$$\Theta^{\Delta t} V(x) := E\left[V(X(t+\Delta t))|X(t)=x\right] \tag{2}$$

  that returns the expectation value of a function V, depending on a stochastic process $X(t)$, after a time step Δt without any activity.
- The effect *transacting* can be described with the *transaction operator* $\curlywedge$ that modifies the argument x of a function $V(x)$ by

$$\overset{x}{\underset{f(x)}{\curlywedge}} V(x) := V(f(x)). \tag{3}$$

  With this, we can model for example dividend payouts.
- Finally, the *decision operator*

$$\begin{array}{ccc} & \overset{C}{\diagup} & O_1 \\ & \diagdown & \end{array} := \begin{cases} O_1, & \text{where condition } C \text{ is fulfilled,} \\ O_2, & \text{otherwise,} \end{cases} \tag{4}$$

  models the effect of *deciding* between O_1 and O_2, depending on a condition C. In the following, for simplicity, we restrict C to be a condition that depends only on the function values.

Providing this calculus, we can formulate for example the price V of an Asian option with n samples and strike K simply with the operator sequence

$$V = \overset{a}{\underset{0}{\curlywedge}} \left(\Theta^{\frac{1}{n}} \overset{a}{\underset{a+S}{\curlywedge}} \right)^n \max\{a/n - K, 0\}. \tag{5}$$

Here, a is an auxiliary variable for the computation of the average stock value. First, it is set to zero by the operation $\curlywedge \overset{a}{0}$. After that, it sums up the stock value S, that underlies a stochastic process, after every time step $\Delta t = \frac{1}{n}$ by $\curlywedge \overset{a}{a+S}$. Finally, the payoff function is evaluated.

It is also possible to model constraints with the use of the decision operator. In case of an American option that allows its exercise during the whole time period, the corresponding operator sequence can be written as a *loop-inf-construct*

$$\lim_{n\to\infty} \left(\Theta^{\frac{T}{n}} \begin{array}{ccc} & \overset{\max\{K-S,0\}}{\diagup} & \max\{K-S,0\} \\ & \diagdown & \\ & & . \end{array} \right)^n . \tag{6}$$

If values fall below the payoff $\max\{K - S, 0\}$, the payoff is returned. For all other values, we go on with waiting without doing any other operation. This corresponds to an operator Θ^T with a constraint $V > \max\{K - S, 0\}$. For examples, see [2].

2.2 Operator-Based Option Pricing

From the operator-based representation that we introduced in the previous section, we can now deduce the pricing algorithm. The calculations will be performed on a grid and in a backward manner, similarly as it is done when solving a pricing problem in form of a common time-dependent boundary value problem. Thus, the operator-based algorithm splits up into two parts: the forward estimation of the domain size and the backward calculation.

First, we start with the forward estimation of the domain on which we will later do the backward calculation. For this, we have to consider for every operator of the operator sequence its modification to the grid that is done when applying the operator to the current grid: In case of a Markov process operator, the domain has to be enlarged in all dimensions that underlie a stochastic process. The formula for the extension of the domain depends on the stochastic process. In case of a geometric Brownian motion, the size of the grid in dimension of a stochastic variable $S(t)$ at time t is given by $[S_{\min}, S_{\max}]$, where

$$S_{\min,\max}(t) := S(0) \exp\left((\mu - 0.5\sigma^2)t \pm C \cdot \sigma\sqrt{t}\right). \tag{7}$$

Here, C denotes an extension factor that defines how much the grid is enlarged. The time t corresponds to the total time of all Markov process operators considered up to now. Of course, it is also possible to choose a different stochastic process, e.g., a jump diffusion process.

The transaction operator $\nearrow^{x}_{f(x)} V(x)$ distorts the grid depending on the function f, since it modifies the axis values but keeps the function values. For example, in case of the Asian option (5), the transaction operator $\nearrow^{a}_{a+S} V(a, S)$ shifts the grid with its function values along the S-axis.

Since the decision operator only modifies the function values but not an axis variable, there will be no changes of the domain.

After having finished the estimation of the domain size at maturity T, we start with the backward calculation. First of all, we initialize the grid with a function, typically the payoff function. After that, we apply the operators in a backward manner to the grid: When applying a Markov process operator, we have to compute one backward time-step of the underlying stochastic process. In the case of the geometric Brownian motion, this corresponds to a backward time-step of the Black-Scholes equation. The solution process of this PDE will be discussed in Sect. 3.

When applying the transaction operator $\nearrow^{x}_{f(x)} V(x)$, we have to invert the operation of the forward estimation. In case of the Asian option, we must set $a := a - S$

for the backward application of the operator $\nearrow_{a+S}^{a} V(a, S)$. The inversion of $f(x)$ when setting the grid values of the next grid with values $V^t(x)$ can be avoided by applying the transaction operator to the coordinates of this next grid and then evaluating the previous grid $V^{t+\Delta t}$ at this point:

$$V^t(x) = \nearrow_{f(x)}^{x} V^{t+\Delta t}(x) = V^{t+\Delta t}(f(x)). \tag{8}$$

The decision operator can not be inverted in general. We will restrict in this paper only to the special case of the loop-inf-construct of Sect. 2.1. For evaluation, the loop-inf-construct is discretized in time by neglecting the limit and setting n to a sufficiently large value. For example, in case of the American option (6), we get

$$\left(\Theta^{\frac{T}{n}} \leq \begin{matrix} \max\{K-S,0\} \\ \nearrow \\ \\ \searrow \\ . \end{matrix} \begin{matrix} \\ \max\{K-S,0\} \\ \\ \\ \end{matrix} \right)^n . \tag{9}$$

In the context of this loop-inf-construct, the decision operator needs not to be inverted since it is a constraint thats hold during the whole time interval. Thus, the direct application of the operator to the function is essential.

3 Solving the PDE with Sparse Grids

When evaluating the operator sequence introduced in Sect. 2, for every Markov process operator Θ^T, we have to solve the Black–Scholes PDE in a given time interval $[0, T]$. The d-dimensional Black–Scholes equation with backward time $\tau := T - t$ is defined by

$$\frac{\partial V}{\partial \tau} - \frac{1}{2} \sum_{i,j=1}^{d} \sigma_i \sigma_j \rho_{ij} S_i S_j \frac{\partial^2 V}{\partial S_i \partial S_j} - \sum_{i=1}^{d} \mu_i S_i \frac{\partial V}{\partial S_i} + rV = 0, \tag{10}$$

where S_i corresponds to a stock value, the coefficients σ_i denote the volatilities, ρ_{ij} the asset correlations, μ_i the drifts and r the risk-free interest rate.

Most pricing problems are multi-dimensional problems. These problems suffer from the curse of dimensionality since we have $O(n^d)$ unknowns in case of a d-dimensional full Cartesian grid with a partitioning of every dimension into n subintervals. This is reflected in the storage as well as in the computing time and consequently limits the number of dimensions that can be computed. To reduce the number of unknowns and, thus, to get rid of the curse of dimensionality, we use sparse grids [1] that reduce the number of unknowns to $O(n(\log n)^{d-1})$, but obtain almost comparably results.

To get a sparse grid solution, we apply the combination technique [4]. This allows us to compute a sparse grid solution in space easily by combining several solutions

on smaller regular grids. On the regular grids, the Black–Scholes equation (10) is currently solved with the finite difference method. Time integration is done with the well-known Crank–Nicolson scheme.

There exist several methods [5,6] that apply transformations in order to get more efficient and more stable numerical algorithms. But, usually, these approaches are only suitable for special types of options or stochastic models, respectively. Since our aim is to develop a general, automated method for pricing a wide range of financial contracts, we do not take any advantage of these approaches.

Since the combination technique does not allow any local adaptivity, we are currently working on a sparse grid solver that solves the PDE directly in the hierarchical function space. With this approach, adaptivity can be realized, too. This is important, since the non-smoothness of the payoff function should be resolved more accurately. But to give some first results of our approach with Theta calculus, a solver based on the combination technique is sufficient.

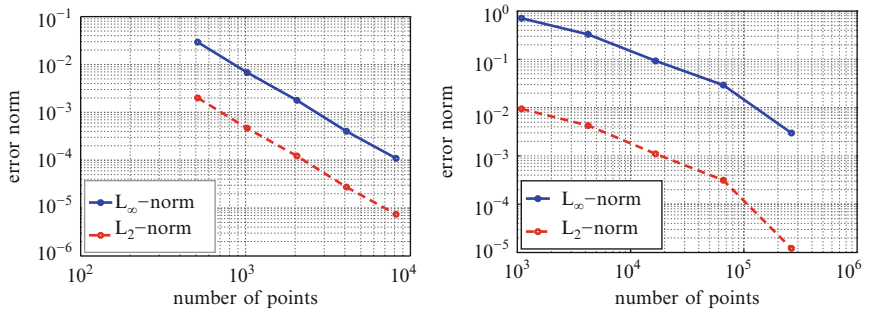
4 Numerical Examples

In the following, we give some examples that were computed with the presented general pricing approach. Every example was computed for different resolutions of the grid. For error estimation, we take the solution on the grid with the highest resolution as reference solution. To assure convergence not only pointwise, we start the calculations on an initial grid instead of a single point with the initial values. This grid is created by adding $\pm 1\%$ to the initial values. On this domain, the L_2 and L_∞ norm of the relative error are considered for every example.

4.1 American Put Option 1D

The first example is a one-dimensional American put option with a maturity of one year. The underlying is assumed to be a geometric Brownian motion with initial value 100, a volatility of 40% and a drift and a risk free interest rate of both 5%. The strike of the option is at-the-money and, thus, equals to 100. The corresponding operator sequence was already given in (9).

Figure 2a shows the resulting relative errors. In both the L_∞- and the L_2-norm, we observe a convergence rate of 2.0. We also compared the calculated prices with those of an accurate PDE method [3] that led to a relative difference of the order 10^{-6}. This shows that our approach with Theta calculus, where constraints are modeled with a simple loop-inf-construct, works well.



(a) American option (b) Asian option

Fig. 2 The L_2 and L_∞ error for the American and the Asian option priced with S = K = 100, $\sigma = 40\%$, μ = r = 5%, and T=1

4.2 Asian Call Option 2D

In the next example, that is given by an Asian option, we demonstrate the use of the transaction operator. For our test calculations, we took monthly sampled values and the stochastic process of the previous example. The corresponding operator sequence was already presented in (5) with $n = 12$. Since S as well as the average a are stochastic variables in (5), a two-dimensional grid in S and a is generated during the automated pricing process.

This example was computed with full grids since the non-adaptive sparse grid of the combination technique does not resolve the kink of the non-smooth payoff function as well and, thus, produces worse results. Due to the modular modeling, the different discretization techniques can be exchanged very easily.

The resulted prices and the relative error that is shown in Fig. 2b confirm the convergence of our method also for this type of option. The rate of convergence is more volatile with an average value of 2.4 in the L_2 and 2.0 in the L_∞ norm.

4.3 Currency Swap 3D

Finally, we consider a currency swap that is defined by the operator sequence

$$V = \overset{c}{\underset{0}{\curlywedge}} \left(\Theta^{\frac{1}{12}} \overset{c}{\underset{c+f}{\curlywedge}} \right)^{12} c + (e^{-r_{dom}} - \frac{1}{FX} \cdot e^{-r_{for}}), \tag{11}$$

where $f = \frac{1}{12}(0.04 - \frac{1}{FX} \cdot 0.02)$. For the domestic and the foreign interest as well as the FX rate, we assume correlated Brownian motions without any drift.

The values of all parameters are stated in Fig. 3 which shows the results of the calculations. We tried both full Cartesian and sparse grids on the 3-dimensional

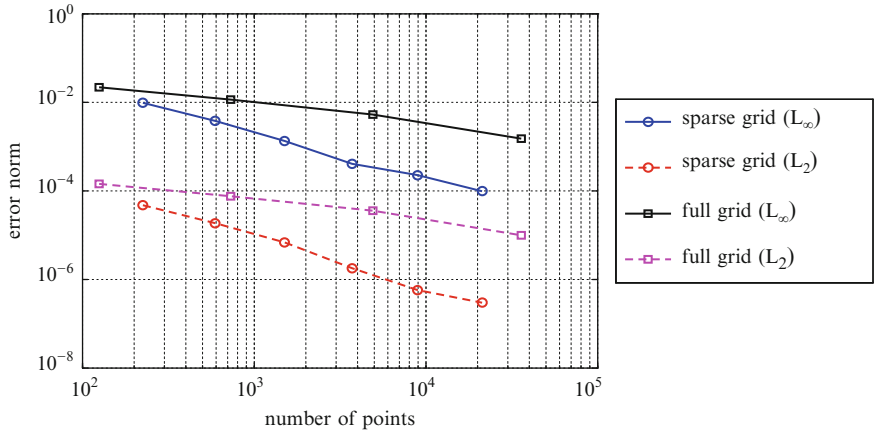


Fig. 3 The L_2 and L_∞ error for a currency swap with FX = 1.4 , $r_{dom} = 2\%$, $r_{for} = 1\%$, $\sigma_{FX} = 0.05$, $\sigma_{dom} = 0.001$, $\sigma_{for} = 0.0015$, $\rho_{for,dom} = 0.5$, and T = 1

domain. The rates of convergence are comparable for both grids but in case of the sparse grid much less points are needed. The general low convergence rate of 1.3–1.4 is supposed to come from the exponential payoff function.

5 Conclusion

In this paper, we presented a general financial pricing method. It was demonstrated how financial contracts can be represented by a sequence of basic operations and priced with this. We showed that our operator based approach works well and leads to accurate results, comparable to those of conventional PDE solvers. The last example also showed that the sparse grid approach is efficient and well-suited and for higher dimensional problems. However, in case of non-smooth functions, the standard combination technique does not resolve the singularities well enough. Thus, we currently work on sophisticated combination techniques and adaptive sparse grids in order to deal also with more general functions in higher dimensions.

References

1. Bungartz, H.-J., Griebel, M.: Sparse grids. Acta numerica **13**, 147–269 (2004)
2. Dirnstorfer, S.: An Introduction to Theta-Calculus. SSRN eLibrary (2005)
3. Forsyth, P. A., Vetzal, K. R.: Quadratic convergence of a penalty method for valuing American options. SIAM J. Sci. Comp. **23**, 2095–2122 (2002)

4. Griebel, M., Schneider, M., Zenger, C.: A combination technique for the solution of sparse grid problems. In: de Groen, P., Beauwens, R. (eds) Iterative Methods in Linear Algebra, pp. 263–281. IMACS (1992)
5. Mertens, T.: Option pricing with sparse grids. Computing in Economics and Finance, 449 (2005)
6. Reisinger, C.: Numerische Methoden für hochdimensionale parabolische Gleichungen am Beispiel von Optionspreisaufgaben. PhD thesis, Universität Heidelberg (2004)

A Posteriori Error Estimation in Mixed Finite Element Methods for Signorini's Problem

Andreas Schröder

Abstract This paper presents a posteriori error estimates for Signorini's problem which is discretized via a mixed finite element approach. The error control relies on the estimation of the discretization error of an auxiliary problem given as a variational equation. The resulting error estimates capture the discretization error of the auxiliary problem, the geometrical error and the error given by the complementary condition. The estimates are applied within adaptive finite element schemes. Numerical results confirm the applicability of the theoretical findings.

1 Introduction

The aim in this paper is to derive error estimates for mixed finite element discretization schemes for Signorini's problem, which plays an import role in mechanical engineering, cf. [6, 7, 14]. The mixed discretization is based on an approach introduced by Haslinger et al. in [8–11, 13]. A saddle point formulation is used where the geometrical contact condition is captured by a Lagrange multiplier. The constraint for the Lagrange multiplier is a sign condition and, therefore, simpler than the original contact condition. However, the multiplier is an additional variable which also has to be discretized. The discretization approach is originally developed for lower-order finite elements. However, it can be extended to higher-order finite elements, cf. [17].

Modern discretization schemes usually include a posteriori error control and adaptivity. To derive an error estimation, we seize a suggestion in [4] for the obstacle problem, where a certain auxiliary problem is considered. We will extend this approach to Signorini's problem and, in particular, to the discretization schemes given by the mixed variational formulation. We obtain error bounds which capture the discretization error of the auxiliary problem, the geometrical error and the error

A. Schröder
Department of Mathematics, Humboldt-Universität zu Berlin, 10099 Berlin, Germany
e-mail: andreas.schroeder@mathematik.hu-berlin.de

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_86,

given by the complementary condition. Furthermore, we apply the estimates within adaptive schemes.

A posteriori error estimates based on the primal, non-mixed formulation are proposed in [2, 5, 18] for the obstacle problem and in [12] for Signorini's problem. In [20,21], estimates for mixed formulations are introduced for the mortar approach. In [16], similar techniques of this work are applied to a simplified Signorini problem. In particular higher-order finite elements are discussed. Furthermore, the results can be applied to time-dependent problems, cf. [3].

2 Signorini's Problem

Signorini's problem describes the deformation of a material body which gets in contact with a rigid foundation. The body is represented by a domain $\Omega \subset \mathbb{R}^k$, $k \in \{2,3\}$, with a sufficiently smooth boundary $\Gamma := \partial\Omega$ and is clamped at a boundary part which is represented by a closed set $\Gamma_D \subset \Gamma$ with positive measure. The boundary part of the body which possibly gets in contact with the foundation is described by an open set Γ_C. We assume that $\overline{\Gamma}_C \subsetneq \Gamma\backslash\Gamma_D$ and $\Gamma_N := \Gamma\backslash(\Gamma_D \cup \overline{\Gamma}_C)$. Volume and surface forces act on the body. They are described by functions $f \in L^2(\Omega;\mathbb{R}^k)$ and $q \in L^2(\Gamma_N;\mathbb{R}^k)$. The resulting deformation is described by a displacement field $v \in H^1(\Omega;\mathbb{R}^k)$ with linearized strain tensor $\varepsilon(v) := \frac{1}{2}(\nabla v + (\nabla v))^\top$. The stress tensor describing a linear-elastic material law is defined as $\sigma(v)_{ij} := \mathscr{C}_{ijkl}\varepsilon(v)_{kl}$, where $\mathscr{C}_{ijkl} \in L^\infty(\Omega)$ with $\mathscr{C}_{ijkl} = \mathscr{C}_{jilk} = \mathscr{C}_{klij}$ and $\mathscr{C}_{ijkl}\tau_{ij}\tau_{kl} \geq \kappa\tau_{ij}^2$ for all $\tau \in L^2(\Omega;\mathbb{R}^{k\times k})$ with $\tau_{ij} = \tau_{ji}$ and a constant $\kappa > 0$. We set $H^1_D(\Omega) := \{v \in H^1(\Omega;\mathbb{R}^k) \mid \gamma_{|\Gamma_D}(v_i) = 0,\ i = 1,\dots,k\}$ for the trace operator $\gamma \in L(H^1(\Omega), L^2(\Gamma))$ and define $(\sigma_n(u))_i := \sigma_{ij}(u)n_j$, $u_n := u_i n_i$, $\sigma_{nn}(u) := \sigma_{ij}(u)n_i n_j$, $\sigma_{nt}(u) := \sigma_n(u) - \sigma_{nn}(u)n$ with outer normal n. Signorini's problem is thus to find a displacement field u such that

$$\begin{aligned} -\operatorname{div}\sigma(u) &= f \text{ in } \Omega, \\ u &= 0 \text{ on } \Gamma_D, \\ \sigma_n(u) &= q \text{ on } \Gamma_N, \\ u_n - g \leq 0,\ \sigma_{nn}(u) \leq 0,\ \sigma_{nn}(u)(u_n - g) &= 0,\ \sigma_{nt}(u) = 0 \text{ on } \Gamma_C. \end{aligned}$$

Here, the function $g \in H^{1/2}(\Gamma_C)$ is the usual linearized gap function describing the surface of the rigid foundation, cf. [14].

In this paper, the following notational conventions are used. The space $H^{-1/2}(\Gamma_C)$ denotes the topological dual space of $H^{1/2}(\Gamma_C)$ with norms $\|\cdot\|_{-1/2,\Gamma_C}$ and $\|\cdot\|_{1/2,\Gamma_C}$, respectively. Let $(\cdot,\cdot)_{0,\omega}$, $(\cdot,\cdot)_{0,\Gamma'}$ be the usual L^2-scalar products on $\omega \subset \Omega$ and $\Gamma' \subset \Gamma$, respectively, for vector and matrix-valued functions. We define $\|v\|^2_{0,\omega} := (v,v)_{0,\omega}$ and omit the subscript ω whenever $\omega = \Omega$. Moreover, we state the energy norm $\|v\|^2 := (\sigma(v),\varepsilon(v))_0$, which is equivalent to the usual norm $\|\cdot\|_1$ in $H^1(\Omega;\mathbb{R}^k)$ due to Korn's inequality. We define $\gamma_N \in L(H^1_D(\Omega), L^2(\Gamma_N,\mathbb{R}^k))$

as $\gamma_N(v)_i = \gamma_{|\Gamma_N}(v_i)$ and $\gamma_{Cn} \in L(H^1_D(\Omega), H^{1/2}(\Gamma_C))$ as $\gamma_{Cn}(v) := \gamma_{|\Gamma_C}(v_i)n_i$ which is surjective due to the assumptions on Γ_C, cf. [14]. Furthermore, we define the norm $\|\cdot\|'_{1/2,\Gamma_C}$ by $\|w\|'_{1/2,\Gamma_C} := \inf_{v\in H^1(\Omega,\Gamma_D),\gamma_{Cn}(v)=w}\|v\|$, which is equivalent to the $\|\cdot\|_{1/2,\Gamma_C}$-norm. The negative part v_- of a function v is defined as $v_-(x) := v(x)$ if $v(x) \le 0$, $v_- := 0$ otherwise.

3 Mixed Variational Formulation of Signorini's Problem and Its Discretization

It is well-known, that the solution of Signorini's problem u is also a solution $u \in K := \{v \in H^1_D(\Omega) \mid \gamma_{Cn}(v) \le g\}$ of the variational inequality

$$(\sigma(u), \varepsilon(v-u))_0 \ge (f, v-u)_0 + (q, \gamma_N(v-u))_{0,\Gamma_N}$$

for all $v \in K$. The inequality above is fulfilled if and only if u is a minimizer of the functional $E(v) := \frac{1}{2}(\sigma(v), \varepsilon(v))_0 - (f, v)_0 - (q, \gamma_N(v))_{0,\Gamma_N}$ in K. The functional E is strictly convex, continuous and coercive due to Cauchy's and Korn's inequalities. This implies the existence of a unique minimizer u.
Given the Lagrange functional $\mathscr{L}(v,\mu) := E(v) + \langle \mu, \gamma_{Cn}(v) - g\rangle$ on $H^1_D(\Omega) \times H^{-1/2}_+(\Gamma_C)$, the Hahn–Banach theorem yields

$$E(u) = \inf_{v\in H^1_D(\Omega)} \sup_{\mu\in H^{-1/2}_+(\Gamma_C)} \mathscr{L}(v,\mu). \tag{1}$$

for $H^{1/2}_+(\Gamma_C) := \{w \in H^{1/2}(\Gamma_C) \mid w \ge 0\}$ and $H^{-1/2}_+(\Gamma_C) := \{\mu \in H^{-1/2}(\Gamma_C) \mid \forall w \in H^{1/2}_+(\Gamma_C) : \langle \mu, w\rangle \ge 0\}$. Thus, u is a minimizer of E, whenever $(u,\lambda) \in H^1_D(\Omega) \times H^{-1/2}_+(\Gamma_C)$ is a saddle point of $\mathscr{L}$. The existence of a unique saddle point is guaranteed, if there exists a constant $\alpha > 0$ such that the inf-sup condition $\alpha\|\mu\|_{-1/2,\Gamma_C} \le \sup_{v\in H^1_D(\Omega),\|v\|_1=1}\langle \mu, \gamma_{Cn}(v)\rangle$ holds for all $\mu \in H^{-1/2}(\Gamma_C)$, cf. [14]. In fact, it follows from the closed range theorem and the surjectivity of γ_{Cn}, that the inf-sup condition is valid. Due to the stationary condition, $(u,\lambda) \in H^1_D(\Omega) \times H^{-1/2}_+(\Gamma_C)$ is a saddle point of $\mathscr{L}$, if and only if it fulfills the mixed variational formulation

$$\begin{aligned} &(\sigma(u), \varepsilon(v))_0 = (f, v)_0 + (q, \gamma_N(v))_0 - \langle \lambda, \gamma_{Cn}(v)\rangle, \\ &\langle \mu - \lambda, \gamma_{Cn}(u) - g\rangle \le 0 \end{aligned} \tag{2}$$

for all $v \in H^1_D(\Omega)$ and $\mu \in H^{-1/2}_+(\Gamma_C)$.

A finite element discretization based on quadrangles or hexahedrons is given in the following way: Let $\mathscr{T}_h$ and $\mathscr{T}_{C,H}$ be finite element meshes of Ω and Γ_C with mesh sizes h and H, respectively. Furthermore, let $\Psi_T : [-1,1]^k \to T \in \mathscr{T}_h$,

$\Psi_{C,T} : [-1,1]^{k-1} \to T \in \mathscr{T}_{C,H}$ be bijective transformations. The space of bilinear or trilinear functions on the reference element $[-1,1]^k$ is denoted by Q_k^1. We set $V_h := \{v \in H_D^1(\Omega) \mid \forall T \in \mathscr{T}_h : \ v_{|T} \circ \Psi_T \in (Q_k^1)^k\}$ and $M_H := \{\mu_H \in L^2(\Gamma_C) \mid \forall T \in \mathscr{T}_{C,H} : \ \mu_{H|T} \equiv \text{constant}\}$. For $M_H^+ := \{\mu_H \in M_H \mid \mu_H \geq 0\}$ the discrete problem is to find $(u_h, \lambda_H) \in V_h \times M_H^+$ such that

$$\begin{aligned} &(\sigma(u_h), \varepsilon(v_h))_0 = (f, v_h)_0 + (q, \gamma_N(v_h))_{0,\Gamma_N} - (\lambda_H, \gamma_{Cn}(v_h))_{0,\Gamma_C}, \\ &(\mu_H - \lambda_H, \gamma_{Cn}(u_h) - g)_{0,\Gamma_C} \leq 0 \end{aligned} \tag{3}$$

for all $v_h \in V_h$ and all $\mu_H \in M_H^+$. To ensure the existence of a unique solution of (3), we have to verify a discrete version of the inf-sup condition. To guarantee the discretization scheme to be stable, the corresponding constant has to be independent of h and H. This can be achieved by using meshes $\mathscr{T}_h$ and $\mathscr{T}_{C,H}$ which imply sufficiently small quotients h/H for $T \in \mathscr{T}_h$, $T_C \in \mathscr{T}_{C,H}$ and $T \subset T_C$, cf. [13]. In our implementation, we ensure $h/H \leq 0.5$, using hierarchical meshes with $\mathscr{T}_{C,H}$ being sufficiently coarser than $\mathscr{T}_h$.

4 Reliable A Posteriori Error Estimates

The basic idea for the estimation of $\|u - u_h\|$ is to consider the following auxiliary problem: Find $u_0 \in H_D^1(\Omega)$ such that

$$(\sigma(u_0), \varepsilon(v))_0 = (f, v)_0 + (q, \gamma_N(v))_{0,\Gamma_N} - (\lambda_H, \gamma_{Cn}(v))_{0,\Gamma_C} \tag{4}$$

for all $v \in H_D^1(\Omega)$. Obviously, the solution u_0 of (4) exists and is unique. Moreover, u_h is a finite element solution of (4). We will show that $\|u - u_h\| \lesssim \|u_0 - u_h\| + \mathscr{R}$ where $\mathscr{R}$ are some remainder terms given below. Here, $\lesssim$ abbreviates $\leq$ up to some constant which is independent of h and H. The idea is to use an arbitrary error estimator η_0 for problem (4) and to set $\eta := \eta_0 + \mathscr{R}$. We then obtain $\|u - u_h\| \lesssim \eta$. In principle, each error estimator known from the literature of variational equations can be used, see [1, 19] for an overview.
In the following, we will make use of the inequalities,

$$ab \leq \epsilon a^2 + \frac{1}{4\epsilon} b^2 \qquad \text{for } a, b \in \mathbb{R},\ \epsilon > 0, \tag{5}$$

$$(a+b)^2 \leq 2a^2 + 2b^2 \qquad \text{for } a, b \in \mathbb{R}, \tag{6}$$

$$x \leq a + b^{1/2} \qquad \text{for } x, a, b > 0,\ x^2 \leq ax + b. \tag{7}$$

Lemma 1. *There holds*

$$\|u - u_h\|^2 \leq \|u_0 - u_h\| \|u - u_h\| + \langle \lambda, \gamma_{Cn}(u_h) - g \rangle.$$

Proof. Since $0, 2\lambda \in H_+^{-1/2}(\Gamma_C)$ and $0, 2\lambda_H \in M_H^+$, we have $\langle \lambda, \gamma_{Cn}(u) - g\rangle = (\lambda_H, \gamma_{Cn}(u_h) - g)_{0,\Gamma_C} = 0$. Furthermore, there holds $(\lambda_H, \gamma_{Cn}(u) - g)_{0,\Gamma_C} \leq 0$. Using Cauchy's inequality, we obtain

$$\begin{aligned}\|u - u_h\|^2 &= (\sigma(u - u_0), \varepsilon(u - u_h))_0 + (\sigma(u_0 - u_h), \varepsilon(u - u_h))_0 \\ &\leq (\lambda_H, \gamma_{Cn}(u - u_h))_{0,\Gamma_C} - \langle \lambda, \gamma_{Cn}(u - u_h)\rangle + \|u_0 - u_h\|\,\|u - u_h\| \\ &= (\lambda_H, \gamma_{Cn}(u) - g)_{0,\Gamma_C} - \langle \lambda, g - \gamma_{Cn}(u_h)\rangle + \|u_0 - u_h\|\,\|u - u_h\| \\ &\leq \langle \lambda, \gamma_{Cn}(u_h) - g\rangle + \|u_0 - u_h\|\|u - u_h\|.\end{aligned}$$

Theorem 1. *Let $\epsilon > 0$, thus*

$$\|u - u_h\| \leq (1 + \epsilon)\|u_0 - u_h\| + (1 + \frac{1}{4\epsilon})\|(g - \gamma_{Cn}(u_h))_-\|'_{1/2,\Gamma_C} + |(\lambda_H, (g - \gamma_{Cn}(u_h))_-)_{0,\Gamma_C}|^{1/2}.$$

Proof. Let $d \in W := \{v \in H_D^1(\Omega) \mid \gamma_{Cn}(v) = (g - \gamma_{Cn}(u_h))_-\}$ with $\|d\| = \inf_{v\in W} \|v\|$. Thus, we have $\|d\| = \|(g - \gamma_{Cn}(u_h))_-\|'_{1/2,\Gamma_C}$. Moreover, there holds $g - \gamma_{Cn}(u_h) - \gamma_{Cn}(d) = g - \gamma_{Cn}(u_h) - (g - \gamma_{Cn}(u_h))_- \geq 0$ on Γ_C and therefore $g - \gamma_{Cn}(u_h) - \gamma_{Cn}(d) \in H_+^{1/2}(\Gamma_C)$. Hence, we obtain

$$\begin{aligned}\langle \lambda, \gamma_{Cn}(u_h) - g\rangle &= -\langle \lambda, g - \gamma_{Cn}(u_h) - \gamma_{Cn}(d)\rangle - \langle \lambda, \gamma_{Cn}(d)\rangle \\ &\leq (\sigma(u), \varepsilon(d))_0 - (f, d)_0 - (q, \gamma_N(d))_{0,\Gamma_N} \\ &= (\sigma(u - u_h), \varepsilon(d))_0 + (\sigma(u_h), \varepsilon(d))_0 - (f, d)_0 - (q, \gamma_N(d))_{0,\Gamma_N} \\ &\leq \|u - u_h\|\|d\| + (\sigma(u_h - u_0), \varepsilon(d))_0 - (\lambda_H, \gamma_{Cn}(d))_{0,\Gamma_C} \\ &\leq \|u - u_h\|\|(g - \gamma_{Cn}(u_h))_-\|'_{1/2,\Gamma_C} \\ &\quad + \|u_0 - u_h\|\|(g - \gamma_{Cn}(u_h))_-\|'_{1/2,\Gamma_C} \\ &\quad + |(\lambda_H, (g - \gamma_{Cn}(u_h))_-)_{0,\Gamma_C}|.\end{aligned}$$

Consequently, Lemma 1 implies

$$\begin{aligned}\|u - u_h\|^2 &\leq \|u_0 - u_h\|\,\|u - u_h\| + \langle \lambda, \gamma_{Cn}(u_h) - g\rangle \\ &\leq \|u - u_h\|(\|u_0 - u_h\| + \|(g - \gamma_{Cn}(u_h))_-\|'_{1/2,\Gamma_C}) + \\ &\quad \|u_0 - u_h\|\|(g - \gamma_{Cn}(u_h))_-\|'_{1/2,\Gamma_C} + |(\lambda_H, (g - \gamma_{Cn}(u_h))_-)_{0,\Gamma_C}|.\end{aligned}$$

The application of (5) and (7) yields

$$\begin{aligned}\|u - u_h\| &\leq \|u_0 - u_h\| + \|(g - \gamma_{Cn}(u_h))_-\|'_{1/2,\Gamma_C} \\ &\quad + (\|u_0 - u_h\|\|(g - \gamma_{Cn}(u_h))_-\|'_{1/2,\Gamma_C} + |(\lambda_H, (g - \gamma_{Cn}(u_h))_-)_{0,\Gamma_C}|)^{1/2} \\ &\leq (1 + \epsilon)\|u_0 - u_h\| + (1 + \frac{1}{4\epsilon})\|(g - \gamma_{Cn}(u_h))_-\|'_{1/2,\Gamma_C} \\ &\quad + |(\lambda_H, (g - \gamma_{Cn}(u_h))_-)_{0,\Gamma_C}|^{1/2}.\end{aligned}$$

Corollary 1. *Let $\eta_0 > 0$ with $\|u - u_h\| \lesssim \eta_0$ and*

$$\eta^2 := \eta_0^2 + \|(g - \gamma_{Cn}(u_h))_-\|^2_{1/2,\Gamma_C} + |(\lambda_H, (g - \gamma_{Cn}(u_h))_-)_{0,\Gamma_C}|. \quad (8)$$

Thus, there holds $\|u - u_h\| \lesssim \eta$.

Proof. Theorem 1, (6) and the equivalence of $\|\cdot\|_{1/2,\Gamma_C}$ and $\|\cdot\|'_{1/2,\Gamma_C}$ yield the assertion.

Remark 1. The terms in the error estimate of Corollary 1 correspond to typical error sources in Signorini's problem: $\|(g - \gamma_{Cn}(u_h))_-\|_{1/2,\Gamma_C}$ measures the error in the geometrical contact condition and $|(\lambda_H, (g - \gamma_{Cn}(u_h))_-)_{0,\Gamma_C}|$ describes the error in the complementary condition.

Remark 2. To calculate η in (8) we have to determine $\|(g - \gamma_{Cn}(u_h))_-\|_{1/2,\Gamma_C}$. Since $\gamma_{Cn}(u_h)$ is piecewise polynomial, we have $(g - \gamma_{Cn}(u_h))_- \in H^1(\Gamma_C)$ for $g \in H^1(\Gamma_C)$. By interpolation results, we get $\|(g - \gamma_{Cn}(u_h))_-\|^2_{1/2,\Gamma_C} \lesssim \|(g - \gamma(u_h))_-\|_{0,\Gamma_C} \|(g - \gamma(u_h))_-\|_{1,\Gamma_C}$, cf. [15, Theorem 7.7.].

Corollary 2. *Let the assumptions of Corollary 1 be fulfilled. Hence, there holds*

$$\|u - u_h\| + \|\lambda - \lambda_H\|_{-1/2,\Gamma_C} \lesssim \eta.$$

Proof. Using $\|\lambda - \lambda_H\|_{-1/2,\Gamma_C} \lesssim \|u - u_0\|$, cf. [16], we obtain

$$\begin{aligned} \|u - u_h\| + \|\lambda - \lambda_H\|_{-1/2,\Gamma_C} &\lesssim \|u - u_h\| + \|u - u_0\| \\ &\lesssim 2\|u - u_h\| + \|u_0 - u_h\| \lesssim \eta + \eta_0 \lesssim \eta. \end{aligned}$$

Remark 3. Since we do not use specific properties of quadrangles or hexahedrons, all results are also valid for discretizations based on triangles or tetrahedrons.

5 Numerical Results

In our numerical experiments, we study Signorini's problem with $\Omega := (-1, 1)^2$, $\Gamma_C := (-1, 1) \times \{-1\}$, $\Gamma_D := [-1, 1] \times \{1\}$, $f := 0$ and $q := 0$. The rigid foundation is given by $\{(x_1, (1 - x_1^2)^{1/2} - 1.85) \in \mathbb{R}^2 \mid x_1 \in [-1, 1]\}$. We use Hooke's law for plain stress with Young's modulus $E := 70kN/mm^2$ and Poisson's number $\nu := 0.33$. In Fig. 1a, the deformation caused by the contact with the rigid foundation is shown. Furthermore, the von-Mises-stress $\sigma_v := (\sigma_{11} + \sigma_{22} - \sigma_{11}\sigma_{22} + 3\sigma_{12}^2)^{1/2}$ is depicted. We see high stress concentrations at the contact zone. An adaptive mesh is shown in Fig. 1b. We use a standard residual error estimator η_0, which is defined by $\eta_0^2 := \sum_{T \in \mathcal{T}_h} (h_T^2 R_{0,T}^2 + \sum_{e \in \mathcal{E}_T} h_e R_{0,e}^2)$ with

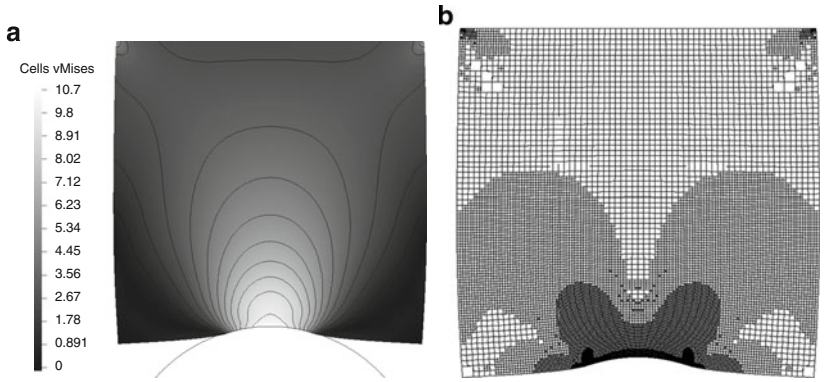
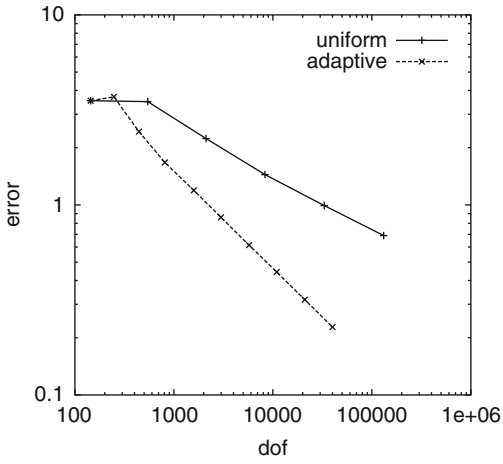


Fig. 1 **(a)** Solution u of Signorini's problem with an obstacle function g, **(b)** adaptive mesh

Fig. 2 Convergence rates



$$R_{0,T} := \|f + \operatorname{div}\sigma(u_h)\|_{0,T},\ T \in \mathscr{T}_h,$$

$$R_{0,e} := \begin{cases} \frac{1}{2}\|[\sigma_n(u_h)]\|_{0,e}, & e \in \mathscr{E}^\circ, \\ \|\sigma_n(u_h) - q\|_{0,e}, & e \in \mathscr{E}_N \\ \|\sigma_{nn}(u_h) + \lambda_H\|_{0,e} + \|\sigma_{nt}(u_h)\|_{0,e}, & e \in \mathscr{E}_C, \end{cases}$$

where $\mathscr{E}_T$ is the set of edges of $T \in \mathscr{T}_h$, $\mathscr{E}^0$ contains the internal edges and $\mathscr{E}_N$ and $\mathscr{E}_C$ the edges on Γ_N and Γ_C, respectively. As usual, $[\cdot]_e$ denotes the jump across an edge $e \in \mathscr{E}^\circ$.

In the adaptive mesh, we find local refinements towards both ends of the contact zone and towards two end points of the dirichlet boundary part Γ_D. Moreover, there are local refinements within the contact zone.

In Fig. 2, the estimated errors obtained by adaptive and uniform refinements are depicted. As the diagram shows, the estimated convergence rate is nearly $\mathscr{O}(h^{1/2})$

for uniform refinements which corresponds to a priori results, cf. [13]. For adaptive refinements, we obtain an optimal algebraic convergence rate $\mathcal{O}(h)$.

References

1. Ainsworth, M., Oden, J.T.: A posteriori error estimation in finite element analysis. Pure and Applied Mathematics. Wiley, Chichester (2000)
2. Bartels, S., Carstensen, C.: Averaging techniques yield reliable a posteriori finite element error control for obstacle problems. Numer. Math. **99**(2), 225–249 (2004)
3. Blum, H., Rademacher, A., Schröder, A.: Space adaptive finite element methods for dynamic signorini problems. Comput. Mech. **44**(4), 481–491 (2009)
4. Braess, D.: A posteriori error estimators for obstacle problems – another look. Numer. Math. **101**(3), 415–421 (2005)
5. Chen, Z., Nochetto, R.H.: Residual type a posteriori error estimates for elliptic obstacle problems. Numer. Math. **84**(4), 527–548 (2000)
6. Glowinski, R.: Numerical methods for nonlinear variational problems. Springer Series in Computational Physics. Springer, New York (1984)
7. Glowinski, R., Lions, J.L., Trémolierès, R.: Numerical analysis of variational inequalities. Studies in Mathematics and its Applications. North-Holland, Amsterdam (1981)
8. Haslinger, J.: Mixed formulation of elliptic variational inequalities and its approximation. Appl. Mat. **26**, 462–475 (1981)
9. Haslinger, J., Hlavacek, I.: Approximation of the signorini problem with friction by a mixed finite element method. J. Math. Anal. Appl. **86**, 99–122 (1982)
10. Haslinger, J., Lovisek, J.: Mixed variational formulation of unilateral problems. Commentat. Math. Univ. Carol. **21**, 231–246 (1980)
11. Haslinger, J., Sassi, T.: Mixed finite element approximation of 3d contact problems with given friction: error analysis and numerical realization. Math. Mod. Numer. Anal. **38**, 563–578 (2004)
12. Hild, P., Nicaise, S.: A posteriori error estimations of residual type for Signorini's problem. Numer. Math. **101**(3), 523–549 (2005)
13. Hlávaček, I., Haslinger, J., Nečas, J., Lovíšek, J.: Solution of variational inequalities in mechanics. Applied Mathematical Sciences. Springer, New York (1988)
14. Kikuchi, N., Oden, J.: Contact problems in elasticity: A study of variational inequalities and finite element methods. SIAM Studies in Applied Mathematics. SIAM, Society for Industrial and Applied Mathematics, Philadelphia (1988)
15. Lions, J.L., Magenes, E.: Non-homogeneous boundary value problems and applications. Vol. I. Translated from the French by P. Kenneth. Die Grundlehren der mathematischen Wissenschaften. Springer, New York (1972)
16. Schröder, A.: Error control in h- and hp-adaptive fem for signorini's problem. J. Numer. Math. **17**(4), 299–318 (2009)
17. Schröder, A.: Mixed finite element methods of higher-order for model contact problems. Humboldt Universität zu Berlin, Institute of Mathematics, Preprint 09-16, submitted to SINUM (2009)
18. Veeser, A.: Efficient and reliable a posteriori error estimators for elliptic obstacle problems. SIAM J. Numer. Anal. **39**(1), 146–167 (2001)
19. Verfürth, R.: A review of a posteriori error estimation and adaptive mesh-refinement techniques. Wiley-Teubner Series Advances in Numerical Mathematics. Wiley, Chichester (1996)
20. Weiss, A., Wohlmuth, B.: A posteriori error estimator and error control for contact problems. Math. Comp. **78**(267), 1237–1267 (2009)
21. Wohlmuth, B.I.: An a posteriori error estimator for two-body contact problems on non-matching meshes. J. Sci. Comput. **33**(1), 25–45 (2007)

Solution of an Inverse Problem for a 2-D Turbulent Flow Around an Airfoil

Jan Šimák and Jaroslav Pelant

Abstract The presented method is intended for a solution of an airfoil design inverse problem. It is capable of suggesting an airfoil shape corresponding to a given pressure distribution on its surface. The method is an extension of a method presented earlier. Using the $k-\omega$ turbulence model it can handle a turbulent boundary layer which improves its applicability. The method is aimed to a subsonic flow, the angle of attack is one of the results of the method. The method is based on the use of an approximate inverse operator coupled with the Navier–Stokes equations equipped with the turbulence model. The equations describing the flow are solved using an implicit finite volume method, the linearized system is solved by the GMRES method. Numerical results are also presented.

1 Introduction

This chapter concerns a numerical method for a solution of an airfoil design inverse problem. The main idea is to construct an inexact inversion, a mapping between a pressure distribution and an airfoil shape. This mapping is used in an iterative process until the desired airfoil is obtained. The method is useful in cases where a specific pressure distribution is desired. The method is aimed to a subsonic flow, the angle of attack is one of the results of the method. The method presented here is an extension of methods described in previous works [2, 4, 5]. The original method deals with an inviscid flow, later with a laminar viscous flow. The current method deals with a turbulent viscous flow described by the Navier–Stokes equations equipped with the $k-\omega$ turbulence model, which improves its applicability. In the following text a short description of the method is given.

J. Šimák (✉) and J. Pelant
Aeronautical Research and Test Institute, Beranových 130, 199 05 Prague – Letňany, Czech Republic
e-mail: simak@vzlu.cz, pelant@vzlu.cz

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_87,

2 Inverse Problem

As was said above, the goal of the method is to find an airfoil shape corresponding to a prescribed pressure distribution. The method utilizes the possibility of derivation of an approximate inversion. Thus the problem could be written as:

Find a pressure pseudo-distribution p such that

$$\mathbf{P}\,\mathbf{L}(p) = f, \tag{1}$$

where f is the given pressure distribution, $\mathbf{P}$ *is an operator representing a solution of a flow problem and finally* $\mathbf{L}$ *represents an approximate inversion to* $\mathbf{P}$.

This problem is solved by the method of successive iterations, where the solution of (1) is a limit of the sequence

$$\{p_k\}_{k=0}^{\infty}\,,\ p_{k+1} = p_k + \alpha\,(f - \mathbf{P}\,\mathbf{L}p_k)\,. \tag{2}$$

The parameter α is a positive real number chosen such that the sequence converges. According to experiences from numerical results, the choice $\alpha = 0.6$ is sufficient to achieve convergence in most cases. Lower values ensure better convergence but also increase the number of iterations.

The approximate inverse operator $\mathbf{L}$ is derived using the thin airfoil theory. The details of this can be found in [2]. In short, the airfoil is composed of a mean camber line and a thickness function, mathematically written

$$\begin{aligned}\psi_1(x) &= x \pm t(x)\frac{s'(x)}{\sqrt{1+s'^2(x)}}\,,\\ \psi_2(x) &= s(x) \mp t(x)\frac{1}{\sqrt{1+s'^2(x)}}\,,\quad x \in \langle 0,1\rangle\,.\end{aligned} \tag{3}$$

In this notation the airfoil coordinates are denoted by (ψ_1, ψ_2), the upper sign is for the upper part of the airfoil and the bottom sign for the bottom part. The functions $s(x)$ (= the mean camber line) and $t(x)$ (= the thickness function) are expressed using the following integrals

$$\begin{aligned}s(x) = {} & \frac{x}{2\pi}\int_0^1 \big(u_{\mathrm{up}}(\xi) - u_{\mathrm{lo}}(\xi)\big)\ln\left|\frac{1-\xi}{\xi}\right| d\xi\,-\\ & -\frac{1}{2\pi}\int_0^1 \big(u_{\mathrm{up}}(\xi) - u_{\mathrm{lo}}(\xi)\big)\ln\left|\frac{x-\xi}{\xi}\right| d\xi,\end{aligned} \tag{4}$$

$$t(x) = \frac{1}{\pi}\int_0^1 \left(\frac{u_{\mathrm{up}}(\xi) + u_{\mathrm{lo}}(\xi)}{2} - 1\right)\ln\left|\frac{1+\sqrt{(\xi - x\xi)/(x - x\xi)}}{1-\sqrt{(\xi - x\xi)/(x - x\xi)}}\right| d\xi. \tag{5}$$

The variables x and ξ run along the normalized chord line, given by the interval $\langle 0, 1\rangle$. The symbols u_{up} and u_{lo} represent a velocity distribution on the airfoil

surface, normalized by the free stream velocity. These functions need not represent a physically relevant distribution, they are related to the sequence (2) instead. Hence, similar to the notation used in (1), they can be called pseudo-distributions. The transformation between the pressure pseudo-distribution and the velocity pseudo-distribution, in the formula denoted as $u(x)$, is the following,

$$\left(\frac{u(x)}{u_\infty}\right)^2 = \frac{2/M_\infty^2 + \gamma - 1}{\gamma - 1}\left(1 - \left(\frac{p(x)}{p_0}\right)^{(\gamma-1)/\gamma}\right). \tag{6}$$

The symbol p_0 is the pressure at zero velocity, M_∞ is the Mach number in the free stream, u_∞ is the free stream velocity and γ is the Poisson adiabatic constant.

From the construction of the airfoil coordinates, it is clear that u_{up} and u_{lo} need to be functions. Since the chord line doesn't need to connect the leftmost point with the rightmost one, simply taking the x-coordinate of a point on the surface doesn't satisfy this requirement, in general. From that reason the distribution is assumed along the mean camber line, with the x-coordinate as the leading variable x. The mean camber line is evaluated in each iteration, so the only additional expense is the inversion of the mapping (3).

3 Flow Problem

The viscous compressible flow around an airfoil is described by the system of the Navier–Stokes equations. Since most of the flow in real situations is turbulent, the laminar model seems insufficient. To improve the quality of the predicted flow and also the stability of the method, a $k-\omega$ model of turbulence is included (see [1,6]).

Model of the Flow

The system of the equations can be written in a vector form

$$\frac{\partial \mathbf{w}}{\partial t} + \sum_{j=1}^{2} \frac{\partial \mathbf{F}_j(\mathbf{w})}{\partial x_j} = \sum_{j=1}^{2} \frac{\partial \mathbf{R}_j(\mathbf{w}, \nabla \mathbf{w})}{\partial x_j} + \mathbf{S}(\mathbf{w}, \nabla \mathbf{w}), \tag{7}$$

where

$$\mathbf{w} = (\varrho, \varrho v_1, \varrho v_2, E, \varrho k, \varrho \omega)^T, \tag{8}$$

$$\mathbf{F}_j(\mathbf{w}) = \left(\varrho v_j, \varrho v_1 v_j + \delta_{1j} p, \varrho v_2 v_j + \delta_{2j} p, (E + p) v_j, \varrho k v_j, \varrho \omega v_j\right)^T, \tag{9}$$

$$\mathbf{R}_j(\mathbf{w}, \nabla\mathbf{w}) = \Big(0, \tau_{j1}, \tau_{j2}, \tau_{j1}v_1 + \tau_{j2}v_2 + \left(\frac{\mu}{P_r} + \frac{\mu_T}{P_{rT}}\right)\gamma\frac{\partial e}{\partial x_j}, (\mu + \sigma_k\mu_T)\frac{\partial k}{\partial x_j}, (\mu + \sigma_\omega\mu_T)\frac{\partial \omega}{\partial x_j}\Big)^T, \tag{10}$$

$$\mathbf{S}(\mathbf{w}, \nabla\mathbf{w}) = \big(0, 0, 0, 0, P_k - \beta^*\varrho\omega k, P_\omega - \beta\varrho\omega^2 + C_D\big)^T. \tag{11}$$

Using the common notation, ϱ denotes a density, p is a pressure, v_1, v_2 are velocity components, E is an energy, k is a turbulent kinetic energy and finally ω is a specific turbulent dissipation. The symbol P_r represents the Prandtl number (the subscript T denotes the turbulence). The viscosity coefficient μ is evaluated using the Sutherland's formula. The symbol μ_T denotes an eddy viscosity coefficient, which is given by the formula

$$\mu_T = \frac{\varrho k}{\omega}\,.$$

The stress tensor in the Navier–Stokes equations is given by relations

$$\begin{aligned}
\tau_{11} &= (\mu + \mu_T)\left(\frac{4}{3}\frac{\partial v_1}{\partial x_1} - \frac{2}{3}\frac{\partial v_2}{\partial x_2}\right) - \frac{2\varrho k}{3},\\
\tau_{22} &= (\mu + \mu_T)\left(-\frac{2}{3}\frac{\partial v_1}{\partial x_1} + \frac{4}{3}\frac{\partial v_2}{\partial x_2}\right) - \frac{2\varrho k}{3},\\
\tau_{12} = \tau_{21} &= (\mu + \mu_T)\left(\frac{\partial v_1}{\partial x_2} + \frac{\partial v_2}{\partial x_1}\right).
\end{aligned}$$

The production of turbulence P_k and the production of dissipation P_ω are expressed as

$$\begin{aligned}
P_k &= \overline{\tau}_{11}\frac{\partial v_1}{\partial x_1} + \overline{\tau}_{12}\left(\frac{\partial v_1}{\partial x_2} + \frac{\partial v_2}{\partial x_1}\right) + \overline{\tau}_{22}\frac{\partial v_2}{\partial x_2},\\
P_\omega &= \alpha_\omega\omega\frac{P_k}{k},
\end{aligned}$$

where $\overline{\tau}_{ij} = \tau_{ij}$ setting $\mu = 0$. Finally, the cross-diffusion term C_D is given by the relation

$$C_D = \sigma_D\frac{\varrho}{\omega}\max\left(\frac{\partial k}{\partial x_1}\frac{\partial \omega}{\partial x_1} + \frac{\partial k}{\partial x_2}\frac{\partial \omega}{\partial x_2}, 0\right).$$

The turbulence model is closed by parameters $\beta^* = 0.09$, $\beta = 5\beta^*/6$, $\alpha_\omega = \beta/\beta^* - \sigma_\omega\kappa^2/\sqrt{\beta^*}$ (where $\kappa = 0.41$ is the von Kármán constant), $\sigma_k = 2/3$, $\sigma_\omega = 0.5$ and $\sigma_D = 0.5$. This choice of parameters resolves the dependence of the $k - \omega$ model on the free stream values [1].

If the turbulent kinetic energy k is set to zero, the turbulence model has no influence upon the Navier–Stokes equations and the laminar model is described.

In the assumed problem, three types of boundary conditions occur: a condition on a wall, a condition at an inlet boundary and a condition at an outlet boundary.

Due to the viscosity, a zero velocity together with a zero turbulent kinetic energy are prescribed on the wall and also a static temperature is prescribed there. The value of the specific turbulent dissipation ω is obtained by the formula

$$\omega_{wall} = \frac{120\mu}{\varrho y_c^2} ,$$

where y_c is the distance between the wall and the centre of a cell in the first row. At the inlet part of the boundary, the velocity vector (v_1, v_2), the density ϱ, turbulent energy k and dissipation ω are prescribed. At the outlet part of the boundary, the static pressure p is prescribed. The other variables are evaluated from values inside the domain. The values of k and ω at the inlet boundary are values of the free stream and are given in the form of a turbulent intensity I and a viscosity ratio $\mathrm{Re}_T = \mu_T/\mu$.

Numerical Treatment

The system of equations mentioned above is solved by the implicit finite volume method. The variables are normalized using critical values of the density, velocity and pressure. The resulting dimensionless system has the same form as the original one and thus no modification to the system is needed. The computational domain is discretized by a structured quadrilateral C-type mesh.

Since the coupling between the equations describing the flow and the equations describing the turbulence is only by the viscous terms, it is possible to solve the problem in two parts [6]. In the first part (continuity equation, momentum equations, energy equation), the variables k and ω are assumed time independent. Similarly, in the second part ($k - \omega$ equations), the variables p, ϱ, v_1, v_2 are held constant in time and the system of two equations is solved with respect to the unknowns k and ω. These systems can be solved independently of each other. This approach reduces computational costs and allows to easily modify a laminar solver into a turbulent one.

The linearized system of algebraic equations is solved by the GMRES method (using software *SPARSKIT2* [3]). The convective terms $\mathbf{F}_i$ are evaluated using the Osher-Solomon numerical flux in the case of the flow part and by the Vijayasundaram numerical flux in the turbulent part. A higher order reconstruction based on the Van Albada limiter is also implemented. The numerical evaluation of a gradient on an edge of two cells is based on values in centres of the six neighbouring cells.

Since a suitable angle of attack has to be found in order to satisfy the condition on the position of the stagnation point on the leading edge, the airfoil is rotated round a chosen point.

4 Numerical Examples

The first example shows an asymmetric case. The method was examined on the NACA4412 airfoil. The Reynolds and Mach numbers are $Re = 6 \cdot 10^6$ and $M_\infty = 0.6$. Because of the restriction on the stagnation point, the angle of attack is set $\alpha_\infty = 1.8°$. The obtained airfoil together with the obtained pressure distribution after 40 iterations is in Fig. 1a. The next Fig. 1b shows a difference between the original and obtained airfoil and also a distribution of the error between the computed and prescribed pressure. A convergence history of the L^2-norm of error $\| p - f \|_2$ is shown in Fig. 2.

The second example is an airfoil shape resulting from a by-hand prescribed distribution. The Mach number $M_\infty = 0.7$ and Reynolds number $Re \approx 15 \cdot 10^6$. The pressure distribution after 30 iterations together with the resulting shape are in Fig. 3a. The angle of attack is $\alpha_\infty = 0.82°$. The error along the chord is in Fig. 3b. The results are quite good, although the pressure is a little bit distorted on the trailing edge. The high error near the leading edge is partly from the steep gradient of

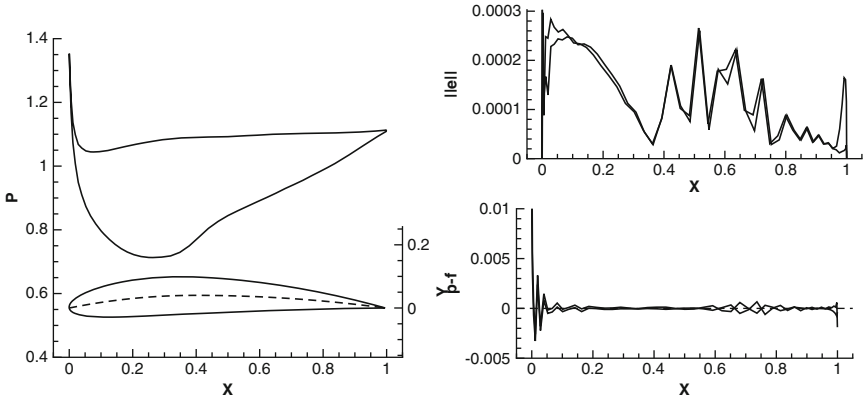


Fig. 1 Example 1: (a) resulting airfoil and pressure distribution (normalized); (**b**) error along the chord, $\|e\| = \|\psi(x) - \psi_{NACA4412}(x)\|$ (*upper*), difference between the obtained and prescribed pressure along the chord (*lower*)

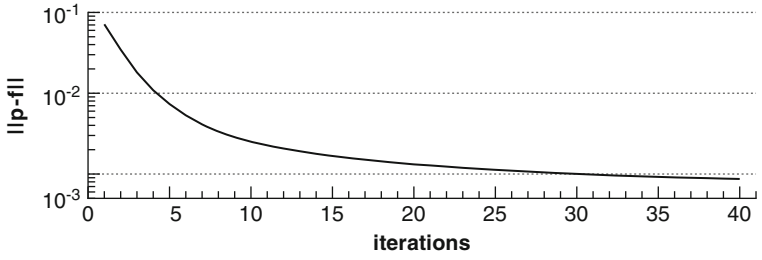


Fig. 2 Example 1: convergence history of an error $\| p - f \|_2$

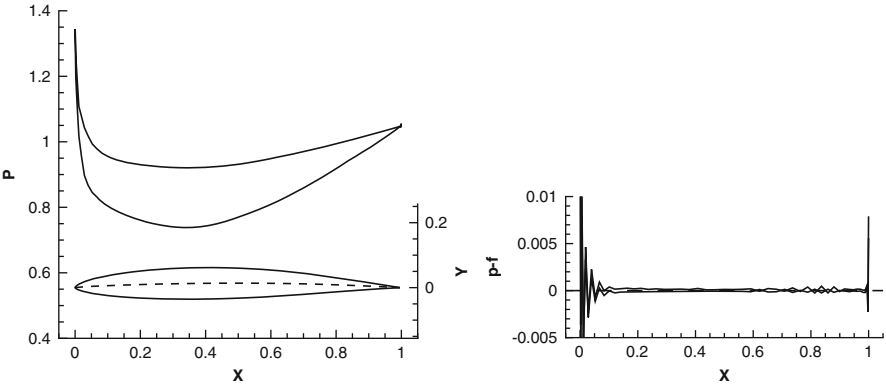


Fig. 3 Example 2: **(a)** Resulting airfoil and pressure distribution (normalized); **(b)** distribution of the error along the chord

the pressure and partly from the method itself. The error of the approximate inverse operator has the highest values right here.

5 Conclusion

An extension of a numerical method for a solution of an inverse problem for a flow around an airfoil was described. This extension is suitable for a design based on a prescribed pressure distribution in the case of a 2D turbulent viscous flow. It was shown that the inverse operator can deal with turbulent viscous flow without the necessity of a correction based on the boundary layer. It is necessary to prescribed the distribution with care because it is possible to happen that the solution doesn't exist. Thus the method can be used to modify existing airfoil with some knowledge. Since the angle of attack is one of the results, the method is applicable in the cases, where a specific angle is not demanded.

Acknowledgements This work was supported by the Grant MSM 0001066902 of the Ministry of Education, Youth and Sports of the Czech Republic. The authors acknowledge this support.

References

1. Kok, J.C.: Resolving the dependence on freestream values for the $k - \omega$ turbulence model. AIAA Journal **38**(7), 1292–1295 (2000)
2. Pelant, J.: Inverse Problem for Two-dimensional Flow Around a Profile. Report Z–69, VZLÚ, Prague (1998)
3. Saad, Y.: Iterative Methods for Sparse Linear Systems, second edn. SIAM (2003)

4. Šimák, J., Pelant, J.: A contractive operator solution of an airfoil design inverse problem. PAMM **7**, 2100023–2100024 (2007) doi:10.1002/pamm.200700136
5. Šimák, J., Pelant, J.: Solution of an Airfoil Design Problem with Respect to a Given Pressure Distribution for a Viscous Laminar Flow. Report R–4186, VZLÚ, Prague (2007)
6. Wilcox, D.C.: Turbulence Modeling for CFD, second edn. DCW Industries Inc (1998)

On Skew-Symmetric Splitting and Entropy Conservation Schemes for the Euler Equations

Björn Sjögreen and H.C. Yee

Abstract The Tadmor type of entropy conservation formulation for the Euler equations and various skew-symmetric splittings of the inviscid flux derivatives are discussed. Numerical stability of high order central and Padé type (centered compact) spatial discretization is enhanced through the application of these formulations. Numerical test on a 2-D vortex convection problem indicates that the stability and accuracy of these formulations using the same high order central spatial discretization are similar for vortex travel up to a few periods. For two to three times longer time integrations, their corresponding stability and accuracy behaviors are very different. The goal of this work is to improve treatment of nonlinear instabilities and to minimize the use of numerical dissipation in numerical simulations of shock-free compressible turbulence and turbulence with shocks.

1 Introduction

Many high resolution numerical schemes for the simulation of turbulence with shocks consist of employing primarily a high order accurate central or Padé (centered compact) spatial discretization in the entire computational domain, and activating a shock-capturing scheme through a flow sensor only in the neighborhood of shocks and in the regions of spurious high frequency oscillations. One example is the filter schemes developed in [8, 9, 14–16]. The objective of this paper is to investigate the stability and accuracy behavior of high order spatial central schemes in conjunction with the use of the various skew-symmetric splittings of the inviscid flux derivative or the Tadmor type of entropy conservation formulation for the

B. Sjögreen (✉)
Lawrence Livermore National Laboratory, Livermore CA94550
e-mail: sjogreen2@llnl.gov

H.C. Yee
NASA Ames Research Center, Moffett Field, CA94035
e-mail: Helen.M.Yee@nasa.gov

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_88,

Euler equations. The flow solutions studied here will therefore be assumed to have a smooth solution.

Due to nonlinear instabilities, solving highly coupled nonlinear conservation laws by spatial centered difference or Padé approximations does not usually lead to a stable method, even when the solution is smooth. Ways to stabilize such methods are to add high order numerical dissipation, or to employ a high pass filter to the solution after each time step. However, in long time integrations and compressible turbulent simulations even small amounts of numerical dissipation can be amplified over time, leading to, e.g., smearing of turbulence fluctuations to un-recognizable forms. An approach to minimize the use of numerical dissipation is to apply these schemes to the split form of the flux derivatives to improve nonlinear stability of the simulation. To understand how this works, consider the the scalar Burgers' equation, $u_t + f_x = 0$ with $f = u^2/2$. The flux derivative can be split into the equivalent form $f_x = \frac{1}{2} f_x + \frac{1}{2} \frac{\partial f}{\partial u} \frac{\partial u}{\partial x}$. For simplicity of discussion, we discretize the split form by a second-order central scheme

$$\frac{d}{dt} u_j + \frac{1}{2} u_j D_0 u_j + \frac{1}{4} D_0 u_j^2 = 0. \tag{1}$$

The grid is uniform, $x_j = (j-1)\Delta x$, with grid spacing Δx, and $u_j(t)$ is an approximation of the solution $u(x,t)$, at the grid point x_j. The centered difference operator is $D_0 u_j = (u_{j+1} - u_{j-1})/2\Delta x$. Linearization of (1) around a smooth and bounded solution $\hat{u}_j(t)$ leads to the equation

$$\frac{d}{dt} e_j + \frac{1}{2} e_j D_0 \hat{u}_j + \frac{1}{2} \hat{u}_j D_0 e_j + \frac{1}{2} D_0 \hat{u}_j e_j = 0$$

for the small perturbation e_j. In the scalar product $(u,v)_h = \Delta x \sum_j u_j v_j$ and norm $||u||_h^2 = (u,u)_h$, we obtain

$$\frac{1}{2} \frac{d}{dt} ||e(t)||_h^2 = (e, e_t)_h = -\frac{1}{2}(e, e D_0 \hat{u})_h - \frac{1}{2}(e, \hat{u} D_0 e)_h - \frac{1}{2}(e, D_0 \hat{u} e)_h.$$

The summation by parts property $(u, D_0 v)_h = -(D_0 u, v)_h$ eliminates the last two terms to give

$$\frac{1}{2} \frac{d}{dt} ||e(t)||_h^2 = -\frac{1}{2}(e, e D_0 \hat{u}) \le C(e,e)_h$$

for a constant C that depends on the maximum spatial derivative of $\hat{u}$. Gronwall's lemma gives the standard well-posedness estimate,

$$||e|| \le K_1 e^{K_2 t}$$

for constants K_1 and K_2. Consequently, the linearization of (1) is L^2 stable. Strang's theorem, see [7], states that if the solution is smooth, and if the method is pth order accurate and smooth, and has an L^2 stable linearization, then the numerical

solution converges with pth order convergence rate. We have thus proved that the split method (1) is convergent as long as no shocks form.

However, because the convergence is up to a fixed time as the grid is refined, it does not necessarily imply that the method is suitable for long time integration. Furthermore, other splittings are possible with different weights on the conservative and non-conservative terms in (1) that do not directly lead to a well-posedness estimate, but turned out to work equally well in numerical experiments. In the absence of better mathematical tools, numerical investigations to assess various possible schemes will be necessary.

In our previous work, non-conservative entropy splitting [17] turned out to be stable and accurate, but when mixed with shock capturing schemes, non-conservative effects sometimes make shocks move with incorrect speeds. As conservative alternatives to entropy splitting, we will consider here the skew-symmetric splitting of Ducros et al. [2, 3] and the entropy conservative formulation of Tadmor [11, 12]. Section 2 describes these methods for the Euler equations of compressible gas dynamics. Section 3 reports results from numerical experiments comparing the entropy split scheme, the skew-split scheme, and the entropy conserving scheme. The discussion concentrates of high order central schemes. Padé type of spatial discretizations will not be discussed due to lack of space.

2 Non-Dissipative Schemes

For ease of presentation we will describe non-dissipative schemes applied to the compressible Euler equations in one space dimension. The generalization to three space dimensions is straightforward. The Euler equations are

$$\mathbf{u}_t + \mathbf{f}(\mathbf{u})_x = \mathbf{0}, \tag{2}$$

where $\mathbf{u} = (\rho,\ \rho u,\ e)$ and $\mathbf{f}(\mathbf{u}) = (\rho u,\ \rho u^2 + p,\ u(e + p))$. The dependent variables are density ρ, momentum ρu, and total energy e. The pressure is $p = (\gamma - 1)(e - \frac{1}{2}\rho u^2)$, where γ is a given constant. The computational domain is $0 < x < L$, with periodic boundary conditions at $x = 0$ and $x = L$. $\mathbf{u}$ is assumed to be given at the initial time. The grid points $x_j = (j-1)\Delta x$, $j = 1, \ldots, N$, where $\Delta x = L/(N-1)$, discretizes the computational domain. Undivided difference operators are denoted $\Delta_+ u_j = u_{j+1} - u_j$, $\Delta_0 u_j = (u_{j+1} - u_{j-1})/2$, and $\Delta_- u_j = (u_j - u_{j-1})$.

2.1 Skew-Symmetric Splitting

Splitting of the derivative of a product in conservative and non-conservative part is done by application of the formula

$$(ab)_x = \frac{1}{2}(ab)_x + \frac{1}{2}ab_x + \frac{1}{2}a_x b, \tag{3}$$

before discretization. An interesting property is that the split approximation can be written on conservative form,

$$\frac{1}{2}D_0(ab)_j + \frac{1}{2}a_j D_0 b_j + \frac{1}{2}b_j D_0 a_j = \frac{1}{4}D_+(a_j + a_{j-1})(b_j + b_{j-1}), \tag{4}$$

where $D_+u_j = (u_{j+1} - u_j)/\Delta x$. (4) can be generalized to arbitrary orders of accuracy if the second order operator D_0 is replaced by the $2p$th order accurate

$$D_{0p}u_j = \sum_{k=1}^{p} \alpha_k^{(p)} D_0(k)u_j. \tag{5}$$

The expanded operators are defined as

$$D_0(k)u_j = (u_{j+k} - u_{j-k})/(2k\Delta x)$$

and the coefficients satisfy

$$\sum_{k=1}^{p} \alpha_k^{(p)} = 1 \qquad \sum_{k=1}^{p} \alpha_k^{(p)} k^{2n} = 0, \;\; n = 1, \ldots, p-1. \tag{6}$$

For details see Ducros et al. [2]. Their key idea is to generalize a splitting that leads to kinetic energy conservation for the incompressible flow equations, to compressible flows.

There are many different ways that (3) can be used for the Euler equations. Different splittings are obtained from different ways to write the fluxes as products of two factors, and it is possible to apply splitting to only some of the equations. In the numerical investigations reported in [5], one of the best performing splittings for (2) was (here displayed with second order accuracy)

$$\begin{aligned}
\frac{d}{dt}\rho_j + \frac{1}{2}D_0\rho_j u_j + \frac{1}{2}\rho_j D_0 u_j + \frac{1}{2}u_j D_0\rho_j &= 0 \\
\frac{d}{dt}(\rho u)_j + \frac{1}{2}D_0\rho_j u_j^2 + \frac{1}{2}\rho_j u_j D_0 u_j + \frac{1}{2}u_j D_0\rho_j u_j + D_0 p_j &= 0 \\
\frac{d}{dt}e_j + \frac{1}{2}D_0 u_j(e_j + p_j) + \frac{1}{2}u_j D_0(e_j + p_j) + \frac{1}{2}(e_j + p_j)D_0 u_j &= 0.
\end{aligned} \tag{7}$$

In three space dimensions the recipe for (7) is to apply (4) to each of the two products in the general three dimensional flux $\mathbf{f} = u_n\mathbf{u} + p\mathbf{e}$, where u_n is the velocity normal to the cell interface, and $\mathbf{e} = (0,\ k_1,\ k_2,\ k_3,\ u_n)$, with $(k_1,\ k_2,\ k_3)$ being the cell interface normal.

See [4] for a comparison of splitting methods with different formulations of the energy equation. For a heuristic discussion on aliasing errors for split approximations, see [1].

The homogeneity of the Euler fluxes means that $\mathbf{f}(\mathbf{u}) = A(\mathbf{u})\mathbf{u}$, where $A(\mathbf{u})$ is the Jacobian of $\mathbf{f}(\mathbf{u})$. A natural splitting would therefore be

$$\frac{d}{dt}\mathbf{u}_j + \frac{1}{2}D_0\mathbf{f}_j + \frac{1}{2}A_j D_0\mathbf{u}_j + \frac{1}{2}D_0(A_j)\mathbf{u}_j = 0, \tag{8}$$

which is of a form that is more suitable for the norm estimate technique described in Sect. 1 for a scalar problem.

2.2 Entropy Conserving Schemes

Entropy conserving schemes were introduced in in the 1980s. See, e.g., [11]. These schemes are in conservation form, and admit a discrete conservation law for the entropy. An entropy, $E(\mathbf{u})$, and an entropy flux $F(\mathbf{u})$ are two functions satisfying

$$E_{\mathbf{u}}^T A(\mathbf{u}) = F_{\mathbf{u}}^T.$$

Here, $E_{\mathbf{u}}$ denotes the gradient of E with respect to $\mathbf{u}$. Furthermore, $E(\mathbf{u})$ is assumed to be a convex function. The entropy variables are defined by $\mathbf{v} = E_{\mathbf{u}}(\mathbf{u})$. Multiplying (2) by $\mathbf{v}^T$ gives the entropy equation

$$\mathbf{v}^T\mathbf{u}_t + \mathbf{v}^T A\mathbf{u}_x = E(u)_t + F_{\mathbf{u}}^T u_x = E(u)_t + F(u)_x = 0.$$

The entropy flux potential, defined by

$$\psi = \mathbf{v}^T\mathbf{f} - F$$

has the property that $\mathbf{f} = \psi_{\mathbf{v}}$.

The following construction defines a high order entropy conservation scheme.

Theorem 1. *The semi-discrete approximation of a system of conservation laws given by*

$$\Delta x \frac{d}{dt}\mathbf{u}_j + \sum_{k=1}^{p} \frac{\alpha_k^{(p)}}{k}(\mathbf{g}_{j+k/2}^{(k)} - \mathbf{g}_{j-k/2}^{(k)}) = \mathbf{0}, \tag{9}$$

where $\mathbf{g}_{j+k/2}^{(k)}$ *satisfies*

$$(\mathbf{v}_{j+k} - \mathbf{v}_j)^T \mathbf{g}_{j+k/2}^{(k)} = \psi_{j+k} - \psi_j \tag{10}$$

and where the kth flux differences approximate the flux derivative to second order with a truncation error of even powers of $k\Delta x$,

$$\mathbf{g}^{(k)}_{j+k/2} - \mathbf{g}^{(k)}_{j-k/2} = k\Delta x \mathbf{f}_x + k^3\Delta x^3\phi_1 + k^5\Delta x^5\phi_2 + \ldots, \tag{11}$$

is $2p$th order accurate, and admits a discrete entropy equation

$$\Delta x \frac{d}{dt} E_j + \sum_{k=1}^{p} \frac{\alpha_k^{(p)}}{k} (H^{(k)}_{j+k/2} - H^{(k)}_{j-k/2}) = 0, \tag{12}$$

where $H^{(k)}_{j+k/2} = \frac{1}{2}((\mathbf{v}_{j+k} + \mathbf{v}_j)^T \mathbf{g}^{(k)}_{j+k/2} - (\psi_{j+k} + \psi_j))$. Both (9) and (12) can be cast in conservation form, because

$$a_{j+k/2} - a_{j-k/2} = \Delta_+ (\sum_{m=0}^{k-1} a_{j-k/2+m})$$

for any arbitrary grid function $a_{j+k/2}$ that satisfies $a_{j+k/2-k} = a_{j-k/2}$.

Proof. Multiply (9) by $\mathbf{v}_j^T$ to obtain

$$\Delta x \frac{d}{dt} E(\mathbf{u}_j)_t + \sum_{k=1}^{p} \frac{\alpha_k^{(p)}}{k} (\mathbf{v}_j^T \mathbf{g}_{j+k/2} - \mathbf{v}_j^T \mathbf{g}_{j-k/2}) = 0.$$

Rewrite each flux difference as

$$\begin{aligned} \mathbf{v}_j^T \mathbf{g}^{(k)}_{j+k/2} - \mathbf{v}_j^T \mathbf{g}^{(k)}_{j-k/2} = {} & \frac{1}{2}(\mathbf{v}_{j+k} + \mathbf{v}_j)^T \mathbf{g}^{(k)}_{j+k/2} - \frac{1}{2}(\mathbf{v}_{j+k} - \mathbf{v}_j)^T \mathbf{g}^{(k)}_{j+k/2} \\ & - \frac{1}{2}(\mathbf{v}_j + \mathbf{v}_{j-k})^T \mathbf{g}^{(k)}_{j-k/2} - \frac{1}{2}(\mathbf{v}_j - \mathbf{v}_{j-k}) \mathbf{g}^{(k)}_{j-k/2} \end{aligned}$$

and use (10) to conclude that

$$\begin{aligned} \mathbf{v}_j^T \mathbf{g}^{(k)}_{j+k/2} - \mathbf{v}_j^T \mathbf{g}^{(k)}_{j-k/2} = {} & \frac{1}{2}((\mathbf{v}_{j+k} + \mathbf{v}_j)^T \mathbf{g}^{(k)}_{j+k/2} - (\mathbf{v}_j + \mathbf{v}_{j-k})^T \mathbf{g}^{(k)}_{j-k/2} \\ & - (\psi_{j+k} + \psi_j) + (\psi_j + \psi_{j-k})). \end{aligned} \tag{13}$$

It is clear from (13) that the entropy conservation (12) follows.

It remains to prove that the order of accuracy is $2p$. Assumption (11) gives

$$\sum_{k=1}^{p} \frac{\alpha_k^{(p)}}{k} (\mathbf{g}^{(k)}_{j+k/2} - \mathbf{g}^{(k)}_{j-k/2}) = \sum_{k=1}^{p} \alpha_k^{(p)} (\Delta x \phi_1 + \alpha_k k^2 \Delta x^3 \phi_3 + \alpha_k k^4 \Delta x^5 \phi_5 + \cdots).$$

(6) gives

$$\sum_{k=1}^{p} \frac{\alpha_k^{(p)}}{k} (\mathbf{g}_{j+k/2}^{(k)} - \mathbf{g}_{j-k/2}^{(k)}) = \Delta x \mathbf{f}_x + \mathscr{O}(\Delta x^{2p+1}),$$

showing that the order of accuracy is $2p$. □

This scheme was also described, although not implemented, in [10].

For a scalar conservation law the simple choice $g_{j+k/2}^{(k)} = (\psi_{j+k} - \psi_j)/(v_{j+k} - v_j)$ satisfies both (10) and (11). For the one dimensional Euler system [12, 13] defined entropy conserving fluxes based on integration in phase space. Here, we instead write ψ as a function of the entropy variables and determine functions φ_i consistent with the gradient of ψ and satisfying

$$(\psi_{j+k} - \psi_j) = \varphi_1((v_1)_{j+k} - (v_1)_j) + \ldots + \varphi_3((v_3)_{j+k} - (v_3)_j).$$

The definition $\mathbf{g}_{j+k/2}^{(k)} = (\varphi_1, \varphi_2, \varphi_3)$ determines an entropy conservative method. As an example, consider the entropy $E(u) = \frac{1+\gamma}{1-\gamma}(\rho p)^{\frac{1}{\gamma+1}}$, which has the entropy flux potential (for explicit expressions for the entropy variables, see [17] or [13])

$$\psi = -\frac{v_2}{v_3}((\gamma-1)(v_1 v_3 - v_2^2/2))^{\frac{1}{1-\gamma}}.$$

Denote $q = (\gamma-1)(v_1 v_3 - v_2^2/2)$, and perform the expansion by repeated use of the rule

$$\Delta ab = \overline{a}\Delta b + \overline{b}\Delta a$$

where Δa denotes $a_{j+k} - a_j$ and $\overline{a}$ denotes $(a_{j+k} + a_j)/2$. The expansion becomes,

$$\Delta\psi = (-\overline{1/v_3}\Delta v_2 - \overline{v_2}\Delta\frac{1}{v_3})\overline{q^{\frac{1}{1-\gamma}}} - \frac{\overline{v_2}}{v_3}\Delta q^{\frac{1}{1-\gamma}}$$

$$= \frac{1}{(v_3)_{j+k}(v_3)_j}(-\overline{v_3}\Delta v_2 + \overline{v_2}\Delta v_3)\overline{q^{\frac{1}{1-\gamma}}} - \frac{\overline{v_2}}{v_3}\Delta q^{\frac{1}{1-\gamma}}$$

with

$$\Delta q^{\frac{1}{1-\gamma}} = \frac{q_{j+k}^{\frac{1}{1-\gamma}} - q_j^{\frac{1}{1-\gamma}}}{q_{j+k} - q_j}\Delta q = \frac{q_{j+k}^{\frac{1}{1-\gamma}} - q_j^{\frac{1}{1-\gamma}}}{q_{j+k} - q_j}(\gamma-1)(\overline{v_3}\Delta v_1 + \overline{v_1}\Delta v_3 - \overline{v_2}\Delta v_2). \tag{14}$$

Denoting $Q = \frac{q_{j+k}^{\frac{1}{1-\gamma}} - q_j^{\frac{1}{1-\gamma}}}{q_{j+k} - q_j}(\gamma-1)$, the final expression becomes

$$\Delta\psi = -\overline{\frac{v_2}{v_3}}Q\overline{v_3}\Delta v_1 + (-\frac{\overline{v_3}}{(v_3)_{j+k}(v_3)_j}\overline{g^{\frac{1}{1-\gamma}}} + \overline{\frac{v_2}{v_3}}Q\overline{v_2})\Delta v_2$$
$$+(\frac{\overline{v_2}}{(v_3)_{j+k}(v_3)_j}\overline{g^{\frac{1}{1-\gamma}}} - \overline{\frac{v_2}{v_3}}Q\overline{v_1})\Delta v_3.$$

It is possible to obtain the numerical flux function in standard variables by transforming back from entropy variables. For example, the mass flux for the second order method ($k = 1$) becomes

$$-\overline{\frac{v_2}{v_3}}Q\overline{v_3} = \overline{u}\,\overline{\rho(p\rho^{-\frac{\gamma}{1+\gamma}})}\,Q.$$

The difference quotient Q tends to $p\rho^{\frac{\gamma}{\gamma+1}}$ when Δq becomes small. Therefore, this flux is consistent. For comparison, the mass flux in (7) is

$$\overline{u}\,\overline{\rho}.$$

Therefore, apart from the factor Q, the entropy conservative scheme can be interpreted as a splitting method. By redefining Q as $\overline{p\rho^{\frac{\gamma}{\gamma+1}}}$, the entropy conservative scheme would become a split scheme, but then perfect entropy conservation would no longer be certain.

3 Numerical Experiments

The isentropic vortex convection problem for the two dimensional Euler equations has initial data

$$\rho = (1 - \frac{(\gamma-1)\hat{\beta}^2}{8\gamma\pi^2}e^{1-r^2})^{\frac{1}{\gamma-1}}$$
$$u = 1 - \frac{\hat{\beta}(y-y_0)}{2\pi}e^{\frac{1-r^2}{2}}$$
$$v = \frac{\hat{\beta}(x-x_0)}{2\pi}e^{\frac{1-r^2}{2}}$$
$$p = \rho^\gamma$$

where $r^2 = (x-x_0)^2 + (y-y_0)^2$, (x_0, y_0) is the center of the vortex, and $\hat{\beta}$ is the strength of the vortex. The exact solution consists of the initial data translated with velocity one in the x-direction. We solve the isentropic vortex convection problem on the computational domain $0 \le x \le 18$, $0 \le y \le 18$ with periodic boundaries. The strength and center of the vortex are $\hat{\beta} = 5$ and $(x_0, y_0) = (9, 9)$, respectively. The grid spacing is $\Delta x = \Delta y = 0.25$. All computations use eighth order accurate

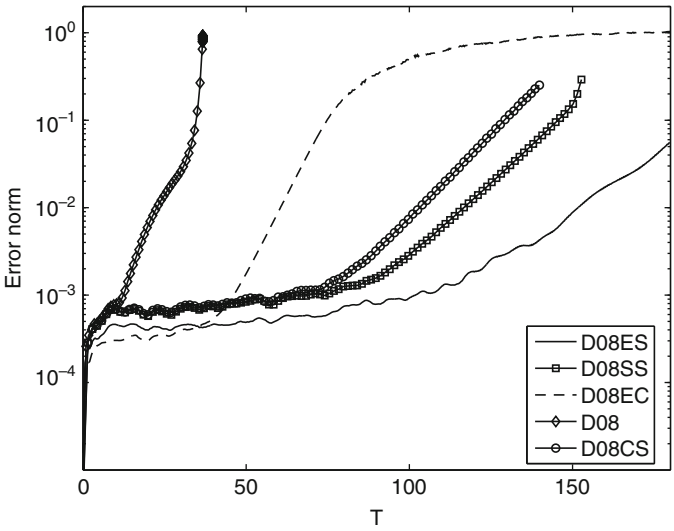


Fig. 1 Vortex convection. Norm of error vs. time for D08ES (*solid*), D08SS (*squares*), D08EC (*dash*), D08 (*diamonds*), and D08CS (*circles*). Inviscid computation

spatial discretizations with fourth order Runge–Kutta in time. Figure 1 displays a comparison of the norm of the solution error vs. time for five different methods. The final time of the computation is 180, which corresponds to 10 periods of vortex convection. D08ES (solid) denotes the non-conservative entropy splitting of Olsson and Oliger [6, 17] with splitting parameter $\beta = 2$, D08SS (squares) denotes the Ducros et al. split scheme (7), D08EC (dash) denotes the Tadmor entropy conservative scheme implemented as described in Sect. 2, D08 (diamonds) denotes the pure centered scheme, and D08CS (circles) denotes the natural split scheme (8). All schemes have small errors during the first period. The purely centered approximation, D08, breaks down due to the non-linear instability at a very early time. After two periods D08EC has the smallest error. The error grows to become large after three periods for D08EC, and after around five to six periods for the other schemes. This error is completely dispersive, and the solutions are highly oscillatory for all methods. The skew split schemes, D08SS and D08CS, break down with negative pressure at around time 140. This does not necessarily mean that they are unstable. They might be accurate for longer times on a finer grid. The entropy split scheme, D08ES, has the best performance, but it will eventually also reach a state where all accuracy has disappeared due to dispersive errors. It appears that the accuracy of D08EC is more sensitive to the small scale oscillations that develop. However, unlike D08SS and D08CS, the small oscillations do not make D08EC break down.

One reason for using entropy split and entropy conservative schemes for the Navier–Stokes equations is that all dissipation in the computed flow will be entirely due to physical viscosity of the Navier–Stokes operator. There is no numerical diffusion. Furthermore, the high frequency modes that cause instabilities in the inviscid

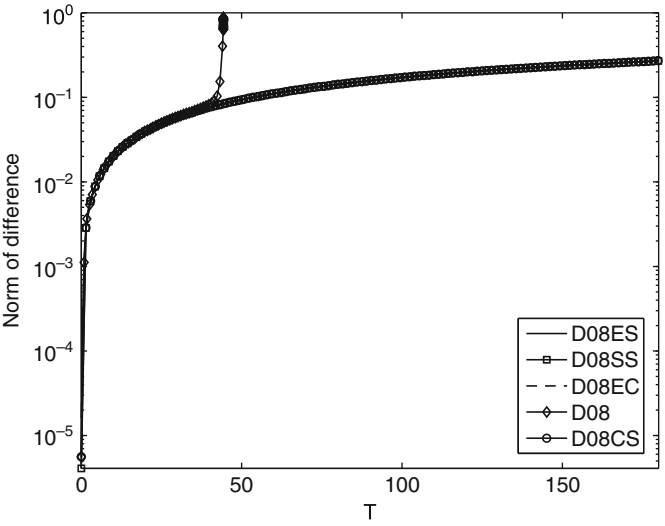


Fig. 2 Vortex convection. Norm of distance to inviscid solution vs. time for D08ES (*solid*), D08SS (*squares*), D08EC (*dash*), D08 (*diamonds*), and D08CS (*circles*). Navier–Stokes equations with $\mu = 0.001$. Results are indistinguishable for all schemes except D08

case will be limited by the physical viscosity. Figure 2 displays the norm of the difference between the inviscid solution and the computed solution vs. time for a solution of the Navier–Stokes equations. The same vortex convection problem as in Fig. 1 was solved, but with the added Navier–Stokes viscosity operator with a constant viscosity coefficient $\mu = 0.001$ and heat conduction corresponding to the Prandtl number 0.72. The viscosity was discretized by eighth order centered difference operators. The viscosity $\mu = 0.001$ is far from resolved on the grid, which has $\Delta x = 0.25$. The parabolic time step restriction is not activated. Even with this small dissipation, all methods, except D08, are well behaved. There is no accumulation of high frequency errors. The curves in Fig. 2 are indistinguishable. The viscosity is not large enough to prevent the blow-up of the pure centered scheme. However, increasing the viscosity to $\mu = 0.01$, which is also unresolved on the grid, gives more or less identical results with all methods (results not plotted), including the pure centered scheme.

In summary, the non-conservative entropy splitting and the Ducros et al. skew-symmetric split formulations perform the best for this particular smooth flow. However, Ducros et al.'s formulation is conservative and it is applicable to problems containing shock waves.

Acknowledgements The work of the first author performed under the auspices of the US Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. LLNL-CONF-417535. The financial support from the NASA Fundamental Aeronautics (Hypersonic) program for the second author is gratefully acknowledged.

References

1. G. A. Blaisdell, E. Spyropoulos, and J.H. Qin, *The Effect of the Formulation of Nonlinear Terms on Aliasing Errors in Spectral Methods*, Appl. Num. Math., **21** (1996) 207–219
2. F. Ducros, F. Laporte, T. Soulères, V. Guinot, P. Moinat, and B. Caruelle, *High-order Fluxes for Conservative Skew-Symmetric-like Schemes in Structured Meshes: Application to Compressible Flows*, J. Comput. Phys., **161** (2000) 114–139
3. W. J. Feiereisen, W. C. Reynolds, and J. H. Ferziger, *Numerical Simulation of a Compressible Homogeneous, Turbulent Shear Flow*, Report TF-13, Thermosciences Division, Department of Mechanical Engineering, Stanford University (1981)
4. A.E. Honein and P. Moin, *Higher Entropy Conservation and Numerical Stability of Compressible Turbulence Simulations*, J. Comput. Phys., **201** (2004) 531–545
5. A.E. Honein and P. Moin, *Numerical Aspects of Compressible Turbulence Simulations*, Report TF-92, Flow Physics and Computation Division, Department of Mechanical Engineering, Stanford University (2005)
6. P. Olsson and J. Oliger, *Energy and Maximum Norm Estimates for Nonlinear Conservation Laws*, RIACS Technical Report 94.01 (1994)
7. R.D. Richtmyer and K.W. Morton, *Difference Methods for Initial-Value Problems*, 2nd ed., John Wiley & Sons, New York (1967)
8. B. Sjögreen and H. C. Yee, *Grid Convergence of High Order Methods for Multiscale Complex Unsteady Viscous Compressible Flows*, J. Comput. Phys., **185** (2003) 1–26
9. B. Sjögreen and H. C. Yee, *Multiresolution Wavelet Based Adaptive Numerical Dissipation Control for Shock-Turbulence Computation*, RIACS Technical Report TR01.01, NASA Ames Research Center (Oct 2000); also, J. Sci. Comput., **20** (2004) 211–255
10. M. Svärd and S. Mishra, *Shock Capturing Artificial Dissipation for High-Order Finite Difference Schemes,* J. Sci. Comput., **39** (2009) 454–484
11. E. Tadmor, *Numerical Viscosity of Entropy Stable Schemes for Systems of Conservation Laws, I*, Math. Comput., **49** (1987) 91–103
12. E. Tadmor, *Entropy Stability Theory for Difference Approximations of Nonlinear Conservation Laws and Related Time-dependent Problems*, Acta Numer., **12** (2003) 451–512
13. E. Tadmor and W. Zhong, *Entropy Stable Approximations of Navier–Stokes Equations with no Artificial Numerical Viscosity*, J. Hyperbolic Diff. Eqn., **3** (2006) 529–559
14. H.C. Yee and B. Sjögreen, *Efficient Low Dissipative High Order Scheme for Multiscale MHD Flows, II: Minimization of Div(B) Numerical Error*, RIACS Technical Report TR03.10, NASA Ames Research Center (Jul 2003); also, J. Sci. Comput., **29** (2006) 115–164
15. H.C. Yee and B. Sjögreen, *Development of Low Dissipative High Order Filter Schemes for Multiscale Navier–Stokes/MHD Systems*, J. Comput. Phys., **225** (2007) 910–934
16. H.C. Yee, N.D. Sandham, and M.J. Djomehri, *Low Dissipative High Order Shock-Capturing Methods Using Characteristic-Based Filters*, J. Comput. Phys., **150** (1999) 199–238
17. H.C. Yee, M. Vinokur, and M.J. Djomehri, *Entropy Splitting and Numerical Dissipation*, J. Comput. Phys., **162** (2000) 33–81

Ideal Curved Elements and the Discontinuous Galerkin Method

Veronika Sobotíková

Abstract In this paper we prove a new result concerning Zlámal's ideal curved elements which allows us to employ these elements in a discontinuous Galerkin finite element method for a nonlinear convection-diffusion problem on a nonpolygonal domain, and to derive an H^1-optimal error estimate for this method.

1 Introduction

Discontinuous Galerkin finite element methods are often used for approximation of nonlinear convection–diffusion problems. Although in practice these problems are usually defined on nonpolygonal domains, theoretical papers consider almost exclusively problems on polygonal domains. In practical computations the curved parts of boundary are usually approximated by line segments. However, the obtained results are not always satisfactory in neighborhood of the boundary. Therefore we are interested in the possibility of using curved elements in the method instead of approximating the boundary.

We employ the ideal curved elements introduced by Zlámal in [3]. In this paper Zlámal showed that if the boundary of the domain is piecewise sufficiently smooth, then there exist regular mappings that map one-to-one the reference triangle on the ideal curved boundary triangles. Our aim is to prove boundedness of higher order derivatives of inverses of these mappings, which allows to use Zlámal's ideal curved elements in a DGFE approximation of a nonlinear non-stationary convection–diffusion problem and to derive an H^1-optimal error estimate for such a method.

V. Sobotíková
Czech Technical University Prague, Faculty of Electrical Engineering, Technická 2, 166 27 Praha 6, Czech Republic
e-mail: veronika@math.feld.cvut.cz

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_89,

2 Ideal Curved Triangulation

Let $\Omega \subset \mathbb{R}^2$ be a bounded domain with a Lipschitz-continuous boundary $\partial\Omega$ and let $\partial\Omega$ be piecewise of the class C^{k+1}, where $k \geq 2$ is an integer. Let $\{\Omega_h\}_{h\in(0,h_0)}$ be a system of polygonal approximations of Ω. On the domains Ω_h we consider triangulations $\mathcal{T}_h$ formed by a finite number of closed triangles K. We assume that the triangulations have the following properties

- All vertices of $\mathcal{T}_h$ lie in $\overline{\Omega}$,
- All vertices lying on the boundary of Ω_h lie on the boundary of Ω, too,
- At most two vertices of any triangle lie on the boundary of Ω_h,
- There are no hanging nodes on the boundary of Ω_h,
- All points from $\partial\Omega$ where the condition of C^{k+1}-smoothness of $\partial\Omega$ is not satisfied are vertices of $\mathcal{T}_h$,
- The system of triangulation $\{\mathcal{T}_h\}_{h\in(0,h_0)}$ is shape-regular.

Now, let $K \in \mathcal{T}_h$ be a *boundary triangle*. We denote its vertices by $P_1,\ P_2,\ P_3$ in such a way that $P_1,\ P_3 \in \partial\Omega_h$. Replacing in K the straight side $\overline{P_1P_3} \subset \partial\Omega_h$ by the arc $\overset{\frown}{P_1P_3} \subset \partial\Omega$, we get the *ideal curved triangle* $\widetilde{K}$ *associated with* K. We set $h_{\widetilde{K}} = h_K =$ diameter of K. If we add the set $\widetilde{\mathcal{T}}_h^B$ of all ideal curved boundary triangles to the set of all straight inner triangles of $\mathcal{T}_h$, we obtain the *ideal curved triangulation* $\widetilde{\mathcal{T}}_h$ of Ω *associated with* $\mathcal{T}_h$.

By $\hat{K}$ we denote a reference triangle with vertices $R_1 = (0,0),\ R_2 = (1,0)$ and $R_3 = (0,1)$.

Theorem 1. *Let h_0 be sufficiently small. Then for each $\widetilde{K} \in \widetilde{\mathcal{T}}_h^B$ there exists such a one-to-one mapping $\tilde{\mathbf{x}}_{\widetilde{K}} : \hat{K} \to \widetilde{K}$ that its Jacobian $J_{\tilde{\mathbf{x}}_{\widetilde{K}}}(\hat{\mathbf{x}})$ is different from zero on $\hat{K}$ and the mapping $\tilde{\mathbf{x}}_{\widetilde{K}}$ as well as its inverse are of class C^k. In addition, there exist positive constants $c_1,\ c_2,\ C_D$ independent of $\widetilde{K} \in \widetilde{\mathcal{T}}_h^B,\ h \in (0,h_0)$, such that*

$$c_1\, h_{\widetilde{K}}^2 \leq |J_{\tilde{\mathbf{x}}_{\widetilde{K}}}(\hat{\mathbf{x}})| \leq c_2\, h_{\widetilde{K}}^2, \tag{1}$$

$$|D^{\boldsymbol{\alpha}} \tilde{x}_{\widetilde{K}i}(\hat{\mathbf{x}})| \leq C_D\, h_{\widetilde{K}}^{|\boldsymbol{\alpha}|}, \qquad 1 \leq |\boldsymbol{\alpha}| \leq k,\ \ i = 1,2, \tag{2}$$

$$|D^{\boldsymbol{\alpha}} \hat{x}_{\widetilde{K}i}(\tilde{\mathbf{x}})| \leq C_D\, h_{\widetilde{K}}^{-1}, \qquad |\boldsymbol{\alpha}| = 1,\ \ i = 1,2, \tag{3}$$

where $(\tilde{x}_{\widetilde{K}1}, \tilde{x}_{\widetilde{K}2})(\hat{\mathbf{x}}) = \tilde{\mathbf{x}}_{\widetilde{K}}(\hat{\mathbf{x}})$, $(\hat{x}_{\widetilde{K}1}, \hat{x}_{\widetilde{K}2})(\tilde{\mathbf{x}}) = \hat{\mathbf{x}}_{\widetilde{K}}(\tilde{\mathbf{x}}) = \big(\tilde{\mathbf{x}}_{\widetilde{K}}\big)_{-1}(\tilde{\mathbf{x}})$ *and* $\alpha = (\alpha_1, \alpha_2)$, $|\alpha| = \alpha_1 + \alpha_2$.

For the proof of Theorem 1 and the definition of mappings $\tilde{\mathbf{x}}_{\widetilde{K}}$ see [3].

In what follows, we shall use this notation: Let $\widetilde{K} \in \widetilde{\mathcal{T}}_h^B$ and $\tilde{w}$ be a (scalar or vector) function defined on $\widetilde{K}$. Then we denote by $\hat{w}$ the function

$$\hat{w}(\hat{\mathbf{x}}) = \tilde{w}(\tilde{\mathbf{x}}(\hat{\mathbf{x}})), \qquad \hat{\mathbf{x}} \in \hat{K}.$$

Similarly, if $\hat{w}$ is a function defined on the reference triangle $\hat{K}$ and $\widetilde{K} \in \widetilde{\mathscr{T}}_h^B$, then we denote by $\tilde{w}$ the function

$$\tilde{w}(\tilde{\mathbf{x}}) = \hat{w}(\hat{\mathbf{x}}(\tilde{\mathbf{x}})), \qquad \tilde{\mathbf{x}} \in \widetilde{K}.$$

3 Properties of the Mappings $\hat{\mathbf{x}}_{\tilde{K}}$

To derive error estimates for a standard finite element method which employs ideal curved elements, knowledge of properties of first derivatives of the mappings $\hat{\mathbf{x}}_{\widetilde{K}} = \left(\tilde{\mathbf{x}}_{\widetilde{K}}\right)_{-1}$ is sufficient. However, if we want to use a discontinuous Galerkin method, we also need to know estimates of higher order derivatives of these mappings. For this reason we prove the following theorem.

Theorem 2. *Let* $\{\tilde{\mathbf{x}}_{\widetilde{K}}\}_{\widetilde{K} \in \widetilde{\mathscr{T}}_h^B,\ h \in (0,h_0)}$ *be any system of mappings with the properties from Theorem 1. Then there exists a positive constant* $\bar{C}_D$ *such that derivatives of the inverse mappings* $\hat{\mathbf{x}}_{\widetilde{K}} = \left(\tilde{\mathbf{x}}_{\widetilde{K}}\right)_{-1}$ *satisfy*

$$|D^{\boldsymbol{\alpha}} \hat{x}_{\widetilde{K}i}(\tilde{\mathbf{x}})| \le \bar{C}_D\, h_{\widetilde{K}}^{-1}, \qquad 1 \le |\boldsymbol{\alpha}| \le k, \quad i = 1, 2, \tag{4}$$

$$\tilde{\mathbf{x}} \in \widetilde{K} \in \widetilde{\mathscr{T}}_h^B, \quad h \in (0, h_0).$$

Remark 1. In fact, as boundedness (3) of the first derivatives of the inverse mappings is a consequence of properties (1) and (2), it is not necessary to require it explicitly.

We shall prove Theorem 2 with the use of the following lemmas:

Lemma 1. *Let* $m > 0$ *be an integer and* $\boldsymbol{\alpha}^* = (\alpha_1^*, \alpha_2^*)$ *be such a multiindex that* $0 \ne |\boldsymbol{\alpha}^*| = \alpha_1^* + \alpha_2^* \le m$. *Then it is possible to express the derivative* $D^{\boldsymbol{\alpha}^*}\tilde{w}$ *of any function* $\tilde{w} \in H^m(\widetilde{K})$, $\widetilde{K} \in \widetilde{\mathscr{T}}_h^B$, *in the form*

$$D^{\boldsymbol{\alpha}^*}\tilde{w}(\tilde{\mathbf{x}}) = \sum_{r=1}^{R^*} A_r^* \cdot B_r^*(\tilde{w}, \hat{\mathbf{x}})(\tilde{\mathbf{x}}) \cdot C_r^*(\hat{\mathbf{x}})(\tilde{\mathbf{x}}), \tag{5}$$

where

- $A_r^* \in \mathbb{N}$ *are constants,*
- $B_r^*(\tilde{w}, \hat{\mathbf{x}})(\tilde{\mathbf{x}}) = D^{\beta^{*r}}\hat{w}(\hat{\mathbf{x}}(\tilde{\mathbf{x}}))$ *for some nonzero multiindex* β^{*r}, $|\beta^{*r}| \le |\boldsymbol{\alpha}^*|$,
- $C_r^*(\hat{\mathbf{x}})(\tilde{\mathbf{x}}) = \prod_{s=1}^{S_r^*} E_s^*(\hat{\mathbf{x}})(\tilde{\mathbf{x}})$, *where* $S_r^* = |\beta^{*r}|$ *and* $E_s^*(\hat{\mathbf{x}})(\tilde{\mathbf{x}}) = (D^{\gamma^{*s}}\hat{x}_1)(\tilde{\mathbf{x}})$ *or* $E_s^*(\hat{\mathbf{x}})(\tilde{\mathbf{x}}) = (D^{\gamma^{*s}}\hat{x}_2)(\tilde{\mathbf{x}}))$ *for some nonzero multiindex* γ^{*s}, $|\gamma^{*s}| \le |\boldsymbol{\alpha}^*|$, $\sum_{s=1}^{S_r^*} |\gamma^{*s}| = |\boldsymbol{\alpha}^*|$.

The values of R^*, A_r^*, S_r^*, the multiindices β^{*r} and γ^{*s} and the forms of B_r^* and C_r^* depend on the multiindex $\boldsymbol{\alpha}^*$ only.

Proof of this lemma can be carried out with the use of mathematical induction on the order $|\boldsymbol{\alpha}^*|$ of the derivative. Using mathematical induction on the order of the derivative $|\bar{\alpha}|$, we can also prove the next lemma:

Lemma 2. *Let $\bar{\alpha} = (\bar{\alpha}_1, \bar{\alpha}_2)$ be a multiindex. Then for any pair of sufficiently smooth functions f, g it is possible to express the derivative $D^{\bar{\alpha}}\left(\frac{f}{g}\right)$ in the form*

$$D^{\bar{\alpha}}\left(\frac{f}{g}\right) = \sum_{r=1}^{\bar{R}} \bar{A}_r \cdot \frac{\bar{B}_r(f) \cdot \bar{C}_r(g)}{g^{\bar{k}_r}}, \tag{6}$$

where

- *$\bar{A}_r \in \mathbb{Z}$ are constants,*
- *$\bar{k}_r \in \{1, 2, \ldots, |\bar{\alpha}| + 1\}$,*
- *$\bar{B}_r(f) = D^{\bar{\beta}^r} f$ for a suitable multiindex $\bar{\beta}^r$, $0 \le |\bar{\beta}^r| \le |\bar{\alpha}|$,*
- *$\bar{C}_r(g) = \prod_{s=1}^{\bar{S}_r} \bar{E}_s(g)$, where $\bar{E}_s(g) = D^{\bar{\gamma}^s} g$ for a suitable multiindex $\bar{\gamma}^s$, $0 < |\bar{\gamma}^s| \le |\bar{\alpha}|$, and $\bar{S}_r = \bar{k}_r - 1$, $|\bar{\beta}^r| + \sum_{s=1}^{\bar{S}_r} |\bar{\gamma}^s| = |\bar{\alpha}|$, or $\bar{C}_r(g) = 1$ and $|\bar{\beta}^r| = |\bar{\alpha}|$ (in this case we set $\bar{S}_r = 0$).*

The values of $\bar{R}$, $\bar{A}_r$, $\bar{k}_r$, $\bar{S}_r$, the multiindices $\bar{\beta}^r$ and $\bar{\gamma}^s$ and the forms of $\bar{B}_r$ and $\bar{C}_r$ depend on the multiindex $\bar{\alpha}$ only.

Lemma 3. *Let $\{\tilde{\mathbf{x}}_{\widetilde{K}}\}_{\widetilde{K} \in \widetilde{\mathscr{T}}_h^B,\, h \in (0,h_0)}$ be any system of mappings with the properties from Theorem 1. Then there exists such a constant $C_J > 0$ that for all multiindices $\beta = (\beta_1, \beta_2)$, $|\beta| \le k - 1$, we have the estimate*

$$\left| \left(D^{\beta} J_{\tilde{\mathbf{x}}_{\widetilde{K}}} \right)(\hat{\mathbf{x}}) \right| \le C_J \cdot h_{\widetilde{K}}^{|\beta|+2} \qquad \hat{\mathbf{x}} \in \hat{K},\ \ \widetilde{K} \in \widetilde{\mathscr{T}}_h^B,\ \ h \in (0, h_0). \tag{7}$$

The constant C_J depends on k and C_D from Theorem 1 only.

Proof. By the definition of the Jacobian $J_{\tilde{\mathbf{x}}_{\widetilde{K}}}(\hat{\mathbf{x}})$ (we shall omit the subindex $\tilde{\mathbf{x}}_{\widetilde{K}}$), we have

$$J(\hat{\mathbf{x}}) = \frac{\partial \tilde{x}_1}{\partial \hat{x}_1}(\hat{\mathbf{x}}) \cdot \frac{\partial \tilde{x}_2}{\partial \hat{x}_2}(\hat{\mathbf{x}}) - \frac{\partial \tilde{x}_1}{\partial \hat{x}_2}(\hat{\mathbf{x}}) \cdot \frac{\partial \tilde{x}_2}{\partial \hat{x}_1}(\hat{\mathbf{x}}).$$

Using Leibniz rule for higher order derivatives of a product, we obtain

$$\begin{aligned}
\left(D^{\beta} J\right)(\hat{\mathbf{x}}) &= \frac{\partial^{\beta_2}}{\partial \hat{x}_2^{\beta_2}} \left(\frac{\partial^{\beta_1} J}{\partial \hat{x}_1^{\beta_1}} \right)(\hat{\mathbf{x}}) \\
&= \sum_{l_2=0}^{\beta_2} \sum_{l_1=0}^{\beta_1} \binom{\beta_2}{l_2} \binom{\beta_1}{l_1} \left\{ \frac{\partial^{l_2}}{\partial \hat{x}_2^{l_2}} \left(\frac{\partial^{l_1}}{\partial \hat{x}_1^{l_1}} \left(\frac{\partial \tilde{x}_1}{\partial \hat{x}_1} \right) \right) \cdot \frac{\partial^{\beta_2 - l_2}}{\partial \hat{x}_2^{\beta_2 - l_2}} \left(\frac{\partial^{\beta_1 - l_1}}{\partial \hat{x}_1^{\beta_1 - l_1}} \left(\frac{\partial \tilde{x}_2}{\partial \hat{x}_2} \right) \right) \right. \\
&\quad \left. - \frac{\partial^{l_2}}{\partial \hat{x}_2^{l_2}} \left(\frac{\partial^{l_1}}{\partial \hat{x}_1^{l_1}} \left(\frac{\partial \tilde{x}_1}{\partial \hat{x}_2} \right) \right) \cdot \frac{\partial^{\beta_2 - l_2}}{\partial \hat{x}_2^{\beta_2 - l_2}} \left(\frac{\partial^{\beta_1 - l_1}}{\partial \hat{x}_1^{\beta_1 - l_1}} \left(\frac{\partial \tilde{x}_2}{\partial \hat{x}_1} \right) \right) \right\} (\hat{\mathbf{x}}) \\
&= \sum_{l_2=0}^{\beta_2} \sum_{l_1=0}^{\beta_1} \binom{\beta_2}{l_2} \binom{\beta_1}{l_1} \left\{ \frac{\partial^{l_1+l_2+1} \tilde{x}_1}{\partial \hat{x}_2^{l_2} \partial \hat{x}_1^{l_1+1}} \cdot \frac{\partial^{\beta_1+\beta_2+1-l_1-l_2} \tilde{x}_2}{\partial \hat{x}_2^{\beta_2+1-l_2} \partial \hat{x}_1^{\beta_1-l_1}} \right. \\
&\quad \left. - \frac{\partial^{l_1+l_2+1} \tilde{x}_1}{\partial \hat{x}_2^{l_2+1} \partial \hat{x}_1^{l_1}} \cdot \frac{\partial^{\beta_1+\beta_2+1-l_1-l_2} \tilde{x}_2}{\partial \hat{x}_2^{\beta_2-l_2} \partial \hat{x}_1^{\beta_1+1-l_1}} \right\} (\hat{\mathbf{x}}).
\end{aligned}$$

From this, by the assumptions of Theorem 1, we deduce

$$\left| \left(D^{\beta} J\right)(\hat{\mathbf{x}}) \right| \leq \sum_{l_2=0}^{\beta_2} \sum_{l_1=0}^{\beta_1} 2\, C_D^2 \binom{\beta_2}{l_2} \binom{\beta_1}{l_1} h_{\widetilde{K}}^{\beta_1+\beta_2+2}.$$

To obtain (7), it is now sufficient to set $C_J = \max\limits_{|\beta| \leq k-1} \left\{ 2C_D^2 \sum\limits_{l_2=0}^{\beta_2} \sum\limits_{l_1=0}^{\beta_1} \binom{\beta_2}{l_2} \binom{\beta_1}{l_1} \right\}$. □

Proof of Theorem 2. Since $\alpha_1 + \alpha_2 = |\boldsymbol{\alpha}| > 0$, in what follows we can assume without the loss of generality that $\alpha_1 \neq 0$. We shall prove estimate (4) only for $i = 1$, in the case $i = 2$ one would proceed similarly.

1. If we differentiate the relation

$$(\tilde{x}_1, \tilde{x}_2) = \left(\tilde{x}_1 \left(\hat{x}_1(\tilde{\mathbf{x}}), \hat{x}_2(\tilde{\mathbf{x}}) \right),\ \tilde{x}_2 \left(\hat{x}_1(\tilde{\mathbf{x}}), \hat{x}_2(\tilde{\mathbf{x}}) \right) \right)$$

with respect to $\tilde{x}_1$, we obtain a system of two linear equations with two unknowns $\frac{\partial \hat{x}_1}{\partial \tilde{x}_1}$ and $\frac{\partial \hat{x}_2}{\partial \tilde{x}_1}$. Solving this system, we find that

$$\frac{\partial \hat{x}_1}{\partial \tilde{x}_1}(\tilde{\mathbf{x}}) = \left(\frac{\frac{\partial \tilde{x}_2}{\partial \hat{x}_2}}{\frac{\partial \tilde{x}_1}{\partial \hat{x}_1} \frac{\partial \tilde{x}_2}{\partial \hat{x}_2} - \frac{\partial \tilde{x}_1}{\partial \hat{x}_2} \frac{\partial \tilde{x}_2}{\partial \hat{x}_1}} \right) (\hat{\mathbf{x}}(\tilde{\mathbf{x}})) = \left(\frac{\frac{\partial \tilde{x}_2}{\partial \hat{x}_2}}{J} \right) (\hat{\mathbf{x}}(\tilde{\mathbf{x}})).$$

2. By Lemmas 2, 3, and the assumptions of the theorem, we have for $|\bar{\alpha}| \le k-1$

$$\left|\left(D^{\bar{\alpha}}\frac{\frac{\partial\tilde{x}_2}{\partial\hat{x}_2}}{J}\right)(\hat{\mathbf{x}})\right| \le \sum_{r=1}^{\bar{R}} |\bar{A}_r| \frac{\left|\bar{B}_r\left(\frac{\partial\tilde{x}_2}{\partial\hat{x}_2}\right)(\hat{\mathbf{x}})\right| \left|\bar{C}_r(J)(\hat{\mathbf{x}})\right|}{|J(\hat{\mathbf{x}})|^{k_r}}$$

$$\le \sum_{r=1}^{\bar{R}} |\bar{A}_r|\, C_D\, C_J^{\bar{S}_r} \frac{h_{\widetilde{K}}^{|\bar{\beta}^r|+1} h_{\widetilde{K}}^{\sum_{s=1}^{\bar{S}_r}(|\bar{\gamma}^s|+2)}}{c_1^{\bar{k}_r}\left(h_{\widetilde{K}}^2\right)^{\bar{k}_r}}.$$

From this, using the relations $|\bar{\beta}^r| + \sum_{s=1}^{\bar{S}_r} |\bar{\gamma}^s| = |\bar{\alpha}|$ and $\bar{S}_r = \bar{k}_r - 1$, we deduce

$$\left|\left(D^{\bar{\alpha}}\frac{\frac{\partial\tilde{x}_2}{\partial\hat{x}_2}}{J}\right)(\hat{\mathbf{x}})\right| \le \sum_{r=1}^{\bar{R}} |\bar{A}_r|\, C_D\, C_J^{\bar{S}_r} c_1^{-\bar{k}_r} h_{\widetilde{K}}^{|\bar{\alpha}|+1+2(\bar{k}_r-1)-2\bar{k}_r}$$

$$= \sum_{r=1}^{\bar{R}} C_D\, C_J^{\bar{S}_r} c_1^{-\bar{k}_r} |\bar{A}_r|\, h_{\widetilde{K}}^{|\bar{\alpha}|-1} \le C^* h_{\widetilde{K}}^{|\bar{\alpha}|-1},$$

where $C^* = \max\limits_{0<|\bar{\alpha}|\le k-1} \left\{ \sum\limits_{r=1}^{\bar{R}(\bar{\alpha})} C_D\, C_J^{\bar{S}_r} c_1^{-\bar{k}_r} |\bar{A}_r(\bar{\alpha})| \right\}$.

3. Now we can approach the proof of (4). As we assume $\alpha_1 \ne 0$, we can write

$$D^{\boldsymbol{\alpha}}\hat{x}_1(\tilde{\mathbf{x}}) = D^{(\alpha_1-1,\alpha_2)}\left(D^{(1,0)}\hat{x}_1(\tilde{\mathbf{x}})\right) = \frac{\partial^{\alpha_1+\alpha_2-1}}{\partial\tilde{x}_1^{\alpha_1-1}\partial\tilde{x}_2^{\alpha_2}}\left(\frac{\partial\hat{x}_1}{\partial\tilde{x}_1}(\tilde{\mathbf{x}})\right)$$
$$= \frac{\partial^{\alpha_1+\alpha_2-1}}{\partial\tilde{x}_1^{\alpha_1-1}\partial\tilde{x}_2^{\alpha_2}}\left(\frac{\frac{\partial\tilde{x}_2}{\partial\hat{x}_2}}{J}(\hat{\mathbf{x}}(\tilde{\mathbf{x}}))\right).$$

Using this expression, we can prove the estimate (4) by induction on $|\boldsymbol{\alpha}|$:

i. Let $|\boldsymbol{\alpha}| = 1$ (i.e., $\boldsymbol{\alpha} = (1,0)$). Then, by the assumptions of the theorem and the part 1) of this proof,

$$|D^{\boldsymbol{\alpha}}\hat{x}_1(\tilde{\mathbf{x}})| = \left|\frac{\partial\hat{x}_1}{\partial\tilde{x}_1}(\tilde{\mathbf{x}})\right| = \left|\left(\frac{\frac{\partial\tilde{x}_2}{\partial\hat{x}_2}}{J}\right)(\hat{\mathbf{x}}(\tilde{\mathbf{x}}))\right| \le \frac{C_D\, h_{\widetilde{K}}}{c_1\, h_{\widetilde{K}}^2} = C_D\, c_1^{-1} h_{\widetilde{K}}^{-1}.$$

Hence, the estimate (4) is valid whenever $\bar{C}_D \ge C_D \cdot c_1^{-1}$.

ii. Let $n \in \{2,\dots,k\}$ and let there exists such a constant $\bar{C}_{D,n-1} > 0$ that

$$\left|D^{\bar{\alpha}}\hat{x}_i(\tilde{\mathbf{x}})\right| \le \bar{C}_{D,n-1}\, h_{\widetilde{K}}^{-1} \qquad \forall\, \tilde{\mathbf{x}} \in \widetilde{K} \quad \forall\, \bar{\alpha},\ 0 < |\bar{\alpha}| \le n-1.$$

Let $|\boldsymbol{\alpha}| = n,\ \alpha_1 \neq 0$. Denoting

$$\tilde{w}(\tilde{\mathbf{x}}) = \left(\frac{\frac{\partial \tilde{x}_2}{\partial \hat{x}_2}}{J}\right)(\hat{\mathbf{x}}(\tilde{\mathbf{x}})), \qquad \boldsymbol{\alpha}^* = (\alpha_1 - 1, \alpha_2)$$

(obviously $|\boldsymbol{\alpha}^*| = n - 1$), we have by Lemma 1, part 2) of this proof and by the induction assumption (we use the notations introduced in Lemma 1)

$$\begin{aligned} |D^{\boldsymbol{\alpha}} \hat{x}_1(\tilde{\mathbf{x}})| &= \left| D^{\boldsymbol{\alpha}^*} \tilde{w}(\tilde{\mathbf{x}}) \right| \le \sum_{r=1}^{R^*} |A_r^*|\, |B_r^*(\tilde{w}, \hat{\mathbf{x}})(\tilde{\mathbf{x}})|\, |C_r^*(\hat{\mathbf{x}})(\tilde{\mathbf{x}})| \\ &\le \sum_{r=1}^{R^*} |A_r^*| C^* h_{\widetilde{K}}^{|\beta^{*r}|-1} \prod_{s=1}^{S_r^*} \bar{C}_{D,n-1} h_{\widetilde{K}}^{-1} \\ &= \sum_{r=1}^{R^*} \left(|A_r^*| C^* \bar{C}_{D,n-1}^{S_r^*} \right) h_{\widetilde{K}}^{|\beta^{*r}|-1-S_r^*} \le \widetilde{C}_{D,n} h_{\widetilde{K}}^{-1}, \end{aligned}$$

where $\widetilde{C}_{D,n}$ depends on n only. Here we used the relation $S_r^* = |\beta^{*r}| \le |\boldsymbol{\alpha}|$ and the fact that the values of $R^*,\ A_r^*$ and C^* depend on $|\boldsymbol{\alpha}^*| = n - 1$ only.

Thus, Theorem 2 is proved. □

4 Application to a Nonlinear Convection–Diffusion Problem

Let us consider the following nonlinear non-stationary convection–diffusion problem: Find $u:\ Q_T = \Omega \times (0,T) \to \mathbb{R}$ such that

$$\begin{aligned} \frac{\partial u}{\partial t} + \sum_{\ell=1}^{2} \frac{\partial f_\ell(u)}{\partial x_\ell} &= \varepsilon\, \Delta u + g \qquad \text{in} \quad Q_T, \\ u\big|_{\partial\Omega\times(0,T)} &= u_D\,, \\ u(x,0) &= u^0(x) \quad x \in \Omega. \end{aligned} \tag{8}$$

We seek an approximate solution of problem (8) on a time level t in the space of discontinuous piecewise "almost polynomial" functions $\widetilde{S}_{hp}$ defined by

$$\widetilde{S}_{hp} = \{\, \tilde{v}_h\,;\ \tilde{v}_h|_{\widetilde{K}} \in \widetilde{P}^p(\widetilde{K}) \ \text{ for all } \ \widetilde{K} \in \widetilde{\mathscr{T}}_h \,\},$$

where $\widetilde{P}^p(\widetilde{K}) = \{\tilde{w}\,;\ \hat{w} \text{ is a polynomial of degree} \le p \text{ on } \hat{K}\,\}$. Discretizing problem (8) using the standard DGFE techniques (we can consider both symmetric and non-symmetric variants of approximation), we get a form $\tilde{a}_h$ approximating the diffusion

term, an interior and boundary penalty $\widetilde{J}_h$ and a right-hand side form $\tilde{\ell}_h$; the convective terms are approximated by a form $\tilde{b}_h$ using a numerical flux, which we assume to be Lipschitz-continuous, consistent and conservative. (For more details see [2].)

We define an *approximate DGFE solution* of problem (8) as a function $\tilde{u}_h \in C^1([0,T];\widetilde{S}_{hp})$ satisfying the following conditions:

$$\begin{aligned} &\text{a)} \left(\frac{\partial \tilde{u}_h(t)}{\partial t}, \tilde{\varphi}_h\right) + \tilde{b}_h(\tilde{u}_h(t), \tilde{\varphi}_h) + \tilde{a}_h(\tilde{u}_h(t), \tilde{\varphi}_h) + \varepsilon \widetilde{J}_h(\tilde{u}_h(t), \tilde{\varphi}_h) = \tilde{\ell}_h(\tilde{\varphi}_h)\,(t) \\ &\qquad\qquad \forall\, \tilde{\varphi}_h \in \widetilde{S}_{hp} \quad \forall\, t \in (0,T), \\ &\text{b)} \quad \tilde{u}_h(0)|_{\widetilde{K}} = \tilde{\Pi}_{\widetilde{K}} u^0 \qquad \forall \widetilde{K} \in \widetilde{\mathscr{T}}_h, \end{aligned} \tag{9}$$

where $\tilde{\Pi}_{\widetilde{K}}$ ($\widetilde{K} \in \widetilde{\mathscr{T}}_h^B$) are projections defined by

$$\tilde{\Pi}_{\widetilde{K}} \tilde{w} = \tilde{v} \quad \Longleftrightarrow \quad \hat{v} \in P^p(\hat{K}) \quad \text{and} \quad (\hat{w}, \hat{\varphi}) = (\hat{v}, \hat{\varphi}) \quad \forall\, \hat{\varphi} \in P^p(\hat{K}). \tag{10}$$

Using Theorem 2, we can show that the projections $\tilde{\Pi}_{\widetilde{K}}$ have the following property

$$|\,\tilde{\Pi}_{\widetilde{K}} \tilde{v} - \tilde{v}\,|_{H^2(\widetilde{K})} \le C_\Pi\, h_K^{r-1} \left(||\,\tilde{v}\,||^2_{H^{r+1}(\widetilde{K})} - ||\,\tilde{v}\,||^2_{L^2(\widetilde{K})} \right)^{1/2} \tag{11}$$

$$\forall \tilde{v} \in H^{r+1}(\widetilde{K}), \qquad r \in \{1, \ldots, k-1\}.$$

This allows us, under the assumption that the boundary of the domain Ω is piecewise of the class C^{p+2} (i.e., $k = p+1$ in Sects. 2 and 3 and in (11)), to derive an error estimate of the type known for the polygonal case (cf. [1]):

Theorem 3. *Let u be the exact solution of problem (8) and let $\tilde{u}_h$ be its approximate DGFE solution. Let u meet the regularity conditions*

$$u \in L^2(0,T; H^{p+1}(\Omega))\,, \qquad \frac{\partial u}{\partial t} \in L^2(0,T; H^p(\Omega))\,.$$

Then there exists a constant $\widetilde{C} > 0$ such that the error of the method $\tilde{e}_h = u - \tilde{u}_h$ satisfies

$$\max_{t\in[0,T]} \|\tilde{e}_h(t)\|^2_{L^2(\Omega)} + \frac{\varepsilon}{2} \int_0^T \left(|\tilde{e}_h(\vartheta)|^2_{H^1(\Omega,\widetilde{\mathscr{T}}_h)} + \widetilde{J}_h^\sigma(\tilde{e}_h(\vartheta), \tilde{e}_h(\vartheta)) \right) \mathrm{d}\,\vartheta \;\le\; \widetilde{C} h^{2p}$$

for all $h \in (0, h_0)$.

For the proof see [2].

Acknowledgements The research was a part of the research project MSM 6840770010 financed by the Ministry of Education of the Czech Republic.

References

1. Dolejší, V. and Feistauer, M.: Error estimates of the discontinuous Galerkin method for nonlinear nonstationary convection–diffusion problems. Numer. Funct. Anal. Optim. **26**(3), 349–383 (2005)
2. Sobotíková, V.: Error analysis of a DG method employing curved elements applied to a nonlinear convection–diffusion problem (In preparation)
3. Zlámal, M.: Curved elements in the finite element method, I. SIAM J. Numer. Anal. **10**, 229–240 (1973)

Analysis of the Parallel Finite Volume Solver for the Anisotropic Allen–Cahn Equation in 3D

Pavel Strachota, Michal Beneš, Marco Grottadaurea, and Jaroslav Tintěra

Abstract In this contribution, a parallel implementation of the finite volume solver is introduced, designated to numerically solve the initial boundary value problem for the Allen–Cahn equation with anisotropy on large 3D grids. The choice of a suitable numerical scheme is discussed and its convergence properties are investigated by means of evaluation of the experimental order of convergence. Afterwards, the consequent limitations for the theoretical error estimate are pointed out. Furthermore, the results of parallel algorithm efficiency measurements are shown, based on extensive tests performed on high performance computing systems. The final part gives a brief overview of a magnetic resonance tractography (neural tract tracking and visualization) method consisting in the solution of the above problem.

1 Introduction

The Allen–Cahn equation having its origin in phase modeling in physics [1] has since found its application in other fields, including image processing and mathematical visualization [2, 8]. In particular, in order to visualize the streamlines of a given tensor field in 3D, an initial boundary value problem for the modified Allen–Cahn equation with incorporated anisotropy can be used [8, 10]. We introduce its parallel numerical solver using several flux approximation schemes on a rectangular

P. Strachota (✉) and M. Beneš
Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering,
Czech Technical University in Prague
e-mail: pavel.strachota@fjfi.cvut.cz, michal.benes@fjfi.cvut.cz

M. Grottadaurea
Department of Engineering, University of Leicester
e-mail: mg165@le.ac.uk

J. Tintěra
Institute for Clinical and Experimental Medicine, Prague
e-mail: jati@medicon.cz

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_90,

grid, justify the choice of the suitable scheme with respect to the undesired artificial dissipation effect and focus on its convergence properties.

2 Problem for the Allen–Cahn Equation with Anisotropy

2.1 Formulation

Assume there is a symmetric positive definite tensor field $\mathbf{D} : \bar{\Omega} \mapsto \mathbb{R}^{3\times3}$ where $\Omega \subset \mathbb{R}^3$ is a block shaped domain. On the time interval $\mathscr{J} = (0, T)$, the initial boundary value problem for the anisotropic Allen–Cahn equation reads

$$\xi \frac{\partial p}{\partial t} = \xi \nabla \cdot \mathbf{D} \nabla p + \frac{1}{\xi} f_0(p) \qquad \text{in } \mathscr{J} \times \Omega, \tag{1}$$

$$\left.\frac{\partial p}{\partial n}\right|_{\partial\Omega} = 0 \qquad \text{on } \bar{\mathscr{J}} \times \partial\Omega, \tag{2}$$

$$p|_{t=0} = I \qquad \text{in } \Omega \tag{3}$$

where $f_0(p) = p(1-p)\left(p - \frac{1}{2}\right)$. Let $x \in \Omega$. Thanks to $\mathbf{D}(x)$ in the diffusion term on the right hand side of (1), the diffusion of p at x is focused into the direction of the principal eigenvector of $\mathbf{D}(x)$, or more precisely, with the directional distribution described by the ellipsoid $\left\{ \boldsymbol{\eta} \in \mathbb{R}^3 \,\middle|\, \boldsymbol{\eta}^{\mathrm{T}} \mathbf{D}(x)^{-1} \boldsymbol{\eta} = 1 \right\}$. In terms of tensor field visualization, we choose the initial condition I in (3) as a noisy texture, preferably an impulse noise. Due to the anisotropic diffusion process carried out by solving (1)–(3), the solution p changes in time from noise to an organized structure. Streamlines of the field of principal eigenvectors of $\mathbf{D}$ can be recognized there as parts with locally similar value of p. The term f_0 efficiently increases contrast of the resulting 3D image provided that the parameter ξ and the final time T are chosen appropriately (in our case by experiment). In order to actually view the resulting 3D image $p(\cdot, T)$, 2D slices through Ω can be helpful.

2.2 Numerical Solution

For numerical solution, the *method of lines* [9] is utilized. Applying a finite volume discretization scheme in space, the problem (1)–(3) is converted to a system of ODE in the general form

$$\frac{\mathrm{d}\mathbf{p}}{\mathrm{d}t} = \mathbf{f}(t, \mathbf{p}). \tag{4}$$

Thereafter, we employ the 4th order Runge–Kutta–Merson solver with adaptive time stepping to solve (4).

Describing the finite volume scheme, (4) can also be referred to as the semidiscrete scheme and written in the form

$$\xi \frac{\mathrm{d}}{\mathrm{d}t} p_K(t) = \xi \sum_{\sigma \in \mathscr{E}_K} F_{K,\sigma}(t) + \frac{1}{\xi} f_{0,K}(t) \qquad \forall K \in \mathscr{T} \tag{5}$$

where $\mathscr{T}$ is an admissible finite volume mesh [4], $K \in \mathscr{T}$ is one particular control volume (cell) and $\mathscr{E}_K$ is the set of all faces of the cell K. $F_{K,\sigma}(t)$ represent the respective numerical fluxes at the time t, which contain difference quotients approximating the derivatives $\partial_x p$, $\partial_y p$, $\partial_z p$ at the center of the face σ.

2.3 Artificial Dissipation and Finite Volume Scheme Design

One can assess the behavior of the numerical solution with respect to *artificial (numerical) dissipation* depending on the exact form of $F_{K,\sigma}$. In the case of tensor field visualization, this phenomenon demonstrating itself as an additional *isotropic diffusion* may significantly deteriorate the visual quality of the result. This is because the streamlines emerging in the solution are thin high frequency structures. To be treated correctly, they require the difference operators used in $F_{K,\sigma}$ to be of an appropriate order [6].

Having the results obtained using different schemes available, one can decide on the best of them by mere visual comparison. We have compared finite volume schemes based on three different discretizations of $F_{K,\sigma}$ together with a standard 1st order forward–backward finite difference scheme. The comparison performed in two different settings was restricted to $\mathbb{R}^2$ and is shown in Fig. 1. In both cases, the initial condition depicted on the very left underwent a process of anisotropic diffusion directed along the axis $y = x$. Least artificial dissipation was produced by the *multipoint flux approximation* (MPFA) scheme where the numerical flux $F_{K,\sigma}$ was obtained using the rules below:

- The difference quotient approximating the derivative in the direction perpendicular to the face σ uses a non-equidistant point distribution in order to avoid redundant interpolation (Fig. 2a). Its 1-dimensional analog for a function $u \in \mathrm{C}^1(\mathbb{R})$ can be represented by the formula

$$u'\left(x_{i+\frac{1}{2}}\right) \approx \frac{1}{24h}\left(u_{i-1} - 27u_i + 27u_{i+1} - u_{i+2}\right)$$

  where $x_j = j \cdot h$, $u_j = u\left(x_j\right)$ for $j \in \mathbb{Z}, h > 0$.
- The remaining derivatives are approximated using a uniform 5-point stencil. Again, its 1D analog can be written as

$$u'(x_i) \approx \frac{1}{12h}\left(u_{i-2} - 8u_{i-1} + 8u_{i+1} - u_{i+2}\right).$$

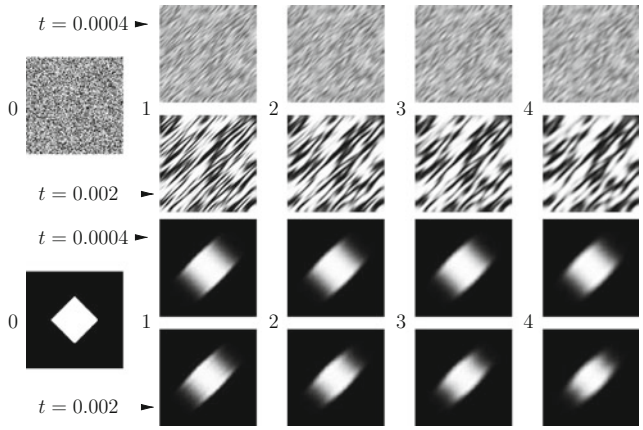


Fig. 1 Artificial diffusion in different numerical schemes. Two time levels for two different initial conditions

Moreover, the stencil points (the crosses along the dashed line in Fig. 2b) are interpolated from the neighboring grid nodes using 1-dimensional cubic interpolation.

3 Convergence Properties

We have been dealing with the derivation of the error estimate for a general finite volume scheme with first order flux approximation on a general mesh. The goal is to prove a first order error bound in the sense of the following result, so far available for the isotropic case and a special centered difference scheme only (see also [4]).

Let $\Omega \subset \mathbb{R}^d$ be a polygonal domain and $T > 0$. Denote by $\mathscr{T}$ an *admissible* [4] mesh defined on Ω, let $k \in (0, T)$ and $N_k = \{n \in \mathbb{N} \,|\, nk \leq T\}$.

Furthermore, for all $K \in \mathscr{T}$, denote by p_K^n the value obtained by numerical solution of 5 approximating $p\,(x_K, nk)$ where $x_K \in K$. The pointwise error is then given by

$$e_K^n = p\,(x_K, t_n) - p_K^n.$$

for all $K \in \mathscr{T}$ and $n \in N_k$. Assuming $p \in \mathrm{C}^2\left(\bar{\Omega} \times \mathscr{J}\right)$, $I \in \mathrm{C}^2\left(\bar{\Omega}, \mathbb{R}\right)$, $\mathbf{D} = \mathbf{I}$, and using the Dirichlet boundary condition $p|_{\partial\Omega} = g$, $g \in \mathrm{C}\,(\partial\Omega \times (0, T))$ instead of (2), there exist positive constants C and k_0 depending only on u, Ω, T , and ξ such that

$$\sqrt{\sum_{K \in \mathscr{T}} \left(e_K^n\right)^2 m(K)} \leq C\,(h + k) \qquad \forall n \in N_k$$

provided that $k \leq k_0$.

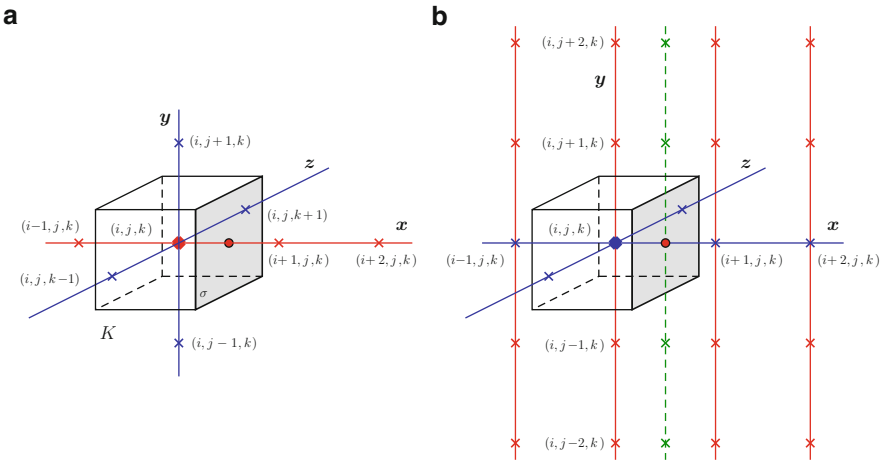


Fig. 2 Point stencils of difference quotients for derivative approximations in the MPFA finite volume scheme

3.1 Experimental Convergence Measurement

Below we provide the experimentally measured convergence rates for some of the schemes. Although not being a general proof of convergence, these results give rise to the following conjectures:

- Despite the difficulties arising from treating the anisotropic diffusion operator on a general mesh, one may expect to prove first order convergence at least for structured meshes thanks to the evidence of experimental convergence analysis.
- On the other hand, proving higher order convergence rate is out of the question even for MPFA schemes as long as cell centered finite volume approach is used.

The experimental order of convergence (EOC) is obtained by computing the solution on a sequence of gradually refining grids and is defined as

$$\mathrm{EOC}_i = \log\left(\frac{\mathrm{Error}_i}{\mathrm{Error}_{i-1}}\right) \Big/ \log\left(\frac{h_i}{h_{i-1}}\right),$$

where $h = \max_K \mathrm{diam}(K)$ is the mesh size and Error_i is the difference of the ith solution from the precise (analytical) solution measured in an appropriate norm. To be able to calculate the analytical solution, we modify the right hand side of (1) to obtain an alternate problem with any prescribed solution of class $\mathrm{C}^2\left(\bar{\Omega} \times \mathscr{J}\right)$. Of course, the prescribed solution must satisfy the initial and boundary condition. The results of the experimental convergence analysis for **D** constant are summarized in Tables 1 and 2.

Table 1 EOC results for the standard central difference scheme

h	$L_\infty(\mathscr{J};L_2(\Omega))$ error $\times 10^{-4}$	EOC in $L_\infty(\mathscr{J};L_2(\Omega))$	$L_\infty(\mathscr{J};L_\infty(\Omega))$ error $\times 10^{-3}$	EOC in $L_\infty(\mathscr{J};L_\infty(\Omega))$
0.00990	2.5560	–	5.5110	–
0.00497	0.6389	2.015	1.3560	2.038
0.00332	0.2844	2.005	0.6097	1.979
0.00249	0.1601	2.002	0.3431	2.004

Table 2 EOC results for the MPFA scheme

h	$L_\infty(\mathscr{J};L_2(\Omega))$ error $\times 10^{-3}$	EOC in $L_\infty(\mathscr{J};L_2(\Omega))$	$L_\infty(\mathscr{J};L_\infty(\Omega))$ error $\times 10^{-2}$	EOC in $L_\infty(\mathscr{J};L_\infty(\Omega))$
0.00971	3.2350	–	2.2190	–
0.00493	1.6190	1.021	1.1140	1.016
0.00330	1.0790	1.012	0.7440	1.008
0.00248	0.8095	1.008	0.5585	1.005

Table 3 Parallel computation efficiency results on IBM BladeCenter LS21 at CINECA, Bologna, Italy. Grid size $400 \times 400 \times 100$ nodes, slice cross section 400×100 nodes

Number of cores n	Time $t_n[s]$	Speedup S_n	Efficiency $E_n[\%]$
1	54044.3	1.000	100.0
2	26377.2	2.049	102.5
4	13238.0	4.083	102.1
8	6752.1	8.004	100.1
20	2688.8	20.100	100.5
40	1366.4	39.552	98.9
80	706.9	76.458	95.6
100	575.7	93.874	93.9

4 Parallel Computation Performance

In our application (see Sect. 5), large meshes (hundreds of nodes in each dimension) need to be dealt with. The numerical solution algorithm has therefore been developed as a parallel code from the beginning, using a simple domain decomposition technique by means of the MPI library. The computational domain is divided into an arbitrary number of successive parallel slices. Each slice is assigned to one MPI process. As each process needs to synchronize solution data with its (at most) two neighbors, this spatial configuration requires the smallest possible number of interprocess communication operations for a given number of processes. On the other hand, it is the least effective solution in terms of the amount of data transferred. In order to investigate how serious the overhead is resulting from heavy communication, series of tests have been performed on the IBM BladeCenter LS21 at CINECA, Bologna, Italy (see Acknowledgements). OpenMPI together with the Intel C/C++ compiler were used. The results in Table 3 confirm that excellent scalability can be achieved even with the decomposition model described above. Slightly superlinear speedup (efficiency over 100%) can be observed thanks to the cache effect [5].

5 Application in MR-DTI

Medical examination of human brain by means of Magnetic Resonance Diffusion Tensor Imaging (MR-DTI) [3,7] generates a tensor field. Applying the visualization procedure described in Sect. 2, we arrive at a 3D texture where the streamlines can be interpreted as neural fiber tracts. Sample DTI visualization is shown in Fig. 3. For details on how tensor data correspond to neural fiber location and orientation, see [7, 11]. Here we focus on the effects of this particular application on the design of the parallel code:

- The choice of the slice orientation in domain decomposition is intended to minimize the amount of data for synchronization, i.e., to minimize the cross section at slice boundaries. Fortunately, the design of the MR scanner determines the shape of the computational grid which has a significantly lower resolution in one of the dimensions. The favorable slice orientation therefore stays fixed and a straightforward algorithm is used to organize data in memory at the start of the computation, so that no `MPI_Pack` operations are necessary for synchronization.
- Tests indicated the appropriate grid size for a whole brain visualization (see the image dimensions in Fig. 3) and also implied that memory is of greater concern than computing power.

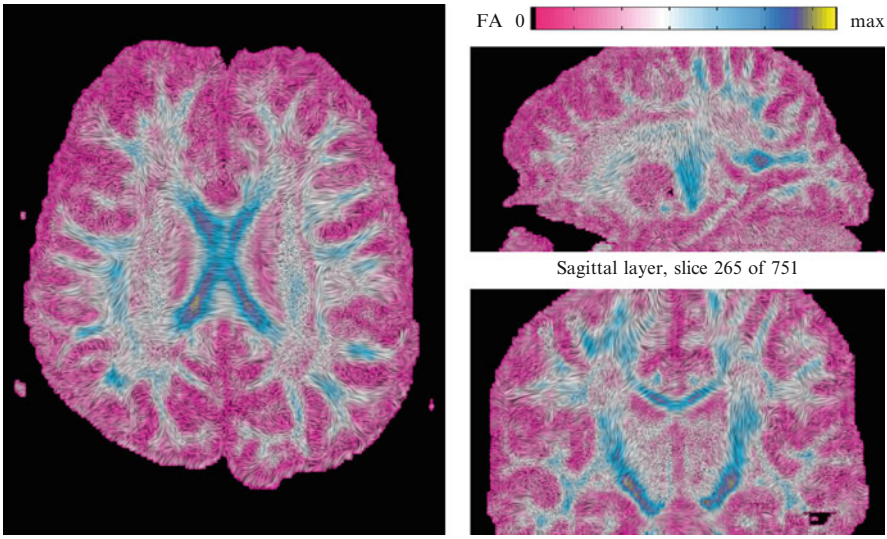


Fig. 3 Sample results of DTI brain visualization. Cuts through the 3D volume in three principal perpendicular planes, colorization by fractional anisotropy [3]. Input data provided by the Institute for Clinical and Experimental Medicine (IKEM), Prague

6 Conclusion

We have developed a tensor field visualization procedure based on numerical solution of the problem for the anisotropic Allen–Cahn equation. We have designed and implemented a parallel numerical solver, gradually employing several numerical schemes. Concerning artificial dissipation, we have observed that the MPFA scheme with cubic interpolation exhibits satisfactory properties for visualization purposes. We have experimentally verified the convergence of all numerical schemes as well as very good scalability of the parallel code. As of the future work, it remains to complete the theoretical error estimation procedure.

Acknowledgements This work was partially supported by the following projects: The HPC-EUROPA project (RII3-CT-2003-506079), with the support of the European Community – Research Infrastructure Action under the FP6 Structuring the European Research Area Program. The HPC-EUROPA++ project (project number: 211437), with the support of the European Community – Research Infrastructure Action of the FP7 "Coordination and support action" Program. The project "Jindřich Nečas Center for Mathematical Modeling," No. LC06052.

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. Beneš, M., Chalupecký, V., Mikula, K.: Geometrical image segmentation by the Allen–Cahn equation. Appl. Numer. Math. **51**(2), 187–205 (2004)
3. Bihan, D.L., et al.: Diffusion tensor imaging: Concepts and applications. J. Magn. Reson. Imaging **13**, 534–546 (2001)
4. 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, Amsterdam (2000)
5. Gustafson, J.L.: Fixed time, tiered memory, and superlinear speedup. In: Proc. 5th Distributed Memory Computing Conference, pp. 1255–1260 (1990)
6. Lomax, H., Pulliam, T.H., Zingg, D.W.: Fundamentals of Computational Fluid Dynamics. Springer, Berlin (2001)
7. Mori, S., Zhang, J.: Principles of diffusion tensor imaging and its applications to basic neuroscience research. Neuron **51**, 527–539 (2006)
8. Preußer, T., Rumpf, M.: Anisotropic nonlinear diffusion in flow visualization. In: Proc. IEEE Visualization 1999, pp. 325–332 (1999)
9. Schiesser, W.E.: The Numerical Method of Lines: Integration of Partial Differential Equations. Academic Press, San Diego (1991)
10. Strachota, P.: Application of anisotropic diffusion in MR tractography. In: Science and Supercomputing in Europe – Report 2008, pp. 279–284. CINECA, Bologna (2009)
11. Tschumperlé, D., Deriche, R.: Tensor field visualization with PDE's and application to DT-MRI fiber visualization. INRIA Sophia-Antipolis, Odyssée Lab, France (2004)

Stabilized Finite Element Approximations of Flow Over a Self-Oscillating Airfoil

Petr Sváček and Jaromír Horáček

Abstract The paper presents the comparison of numerical solution of a 2D aeroelastic problem and experimental results. For the numerical approximation the coupled formulation of a turbulent flow over an oscillating solid airfoil is considered. The flow is modelled by the incompressible Reynolds averaged Navier–Stokes (RANS) equations rewritten in Arbitrary Lagrangian–Eulerian (ALE) form and discretized by the stabilized finite element method (FEM). The numerical results are compared with the results of optical measurements of flow field around an elastically supported vibrating double circular arc (DCA) 18% profile. The measurements were performed above the critical airflow velocity for loss of the system stability by flutter. The numerical results for the time dependent pressure distribution on the fluttering airfoil are presented.

1 Introduction

The interaction of fluid flow and a vibrating structure is important in many technical disciplines, see e.g., [4]. Number of advanced numerical and computational methods for simulation of the fluid–structure interaction were developed during last decades, see [1]. In this paper the main attention is paid to the comparison of numerical simulations and experimental measurement of self-sustained vibrations of a profile in turbulent incompressible flow. The used numerical method was previously developed and applied onto several benchmark problems, see [5, 11].

P. Sváček (✉)
Faculty of Mechanical Engineering, Czech Technical University in Prague,
Department of Technical Mathematics, Karlovo nám. 13, Praha 2, Czech Republic
e-mail: Petr.Svacek@fs.cvut.cz

J. Horáček
Institute of Thermomechanics, Academy of Sciences of the Czech Republic, Dolejškova 5,
Praha 8, Czech Republic
e-mail: jaromirh@it.cas.cz

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_91,

The mathematical model consists of the 2D flow model in interaction with a flexibly supported profile in a channel. The numerical solution of RANS equations is carried out using the finite element method for the spatial discretization of the problem. The finite elements for velocity and pressure were selected to satisfy the Babuška–Brezzi condition in order to guarantee the stability of the scheme, see [7]. The stabilization based on GLS (Galerkin Least-Squares) method together with div-div stabilization was employed in order to suppress the appearance of spurious oscillations due to high Reynolds numbers $Re \approx 10^6$, cf. [6, 8]. The choice of the stabilization parameters is based on the numerical analysis of the problem as well as the numerical experience – see [8, 11]. The Spalart–Allmaras one equation turbulence model is approximated by the FEM stabilized by the streamline upwind/Petrov–Galerkin (SUPG) method.

2 Mathematical Model

The turbulent incompressible flow can be modelled by the RANS equations written in ALE form

$$D^{\mathscr{A}}\mathbf{u}/Dt - \nabla \cdot \Big((\nu + \nu_T)\big(\nabla \mathbf{u} + (\nabla \mathbf{u})^T\big)\Big) + ((\mathbf{u} - \mathbf{w}_g) \cdot \nabla)\mathbf{u} + \nabla p = 0, \quad \text{on } \Omega_t$$
$$\nabla \cdot \mathbf{u} = 0 \qquad (1)$$

where $D^{\mathscr{A}}/Dt$ denotes the ALE derivative, $\mathbf{w}_g$ is the domain velocity, $\mathbf{u} = (u_1, u_2)$ is the mean value of the fluid velocity, ν is the kinematic fluid viscosity, p is the mean value of the kinematic pressure (i.e., pressure divided by the fluid density), and ν_T is a turbulent viscosity that can be obtained by the solution of one or more partial differential equations for additional quantities, see [14]. In order to numerically simulate aeroelastic problems for large vibration amplitudes the ALE formulation of Reynolds equations is used following the notation in [5]. System (1) is considered in a time-dependent domain Ω_t (see Fig. 1). The symbol $\mathscr{A}_t$ denotes a regular one-to-one ALE mapping of the reference configuration Ω_0 onto the current configuration Ω_t for any time instant – see [9]. The system of (1) is equipped with suitable boundary and initial conditions. On the moving part of boundary (airfoil surface Γ_{Wt}) the kinematic boundary condition is prescribed, i.e., $\mathbf{u} = \mathbf{w}_g$ on Γ_{Wt}. At the inlet and on the fixed impermeable channel walls Γ_D the Dirichlet conditions $\mathbf{u} = \mathbf{u}_D$ are considered and at the outlet Γ_O the modification of "do-nothing" boundary condition is used (cf. [12]):

$$-(\nu + \nu_T)\sum_{j=1}^{2}\left(\frac{\partial u_i}{x_j} + \frac{\partial u_j}{x_i}\right) + (p - p_{ref})n_i, \qquad i = 1, 2, \qquad (2)$$

where $\mathbf{n} = (n_1, n_2)^T$ is the unit normal to the boundary of the domain , Γ_O is the outlet and p_{ref} denotes a prescribed reference outlet pressure. The turbulent

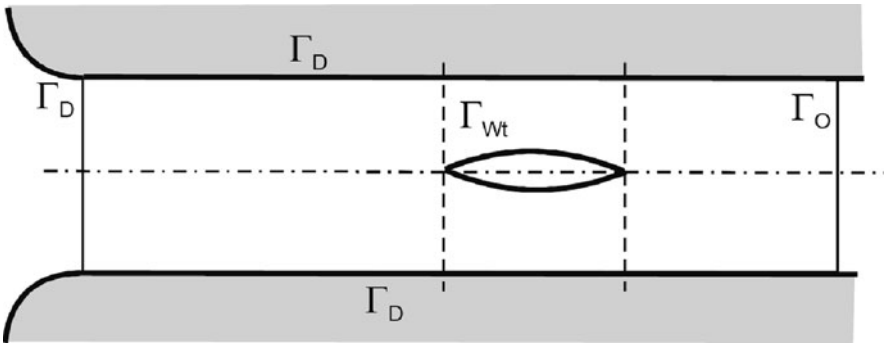


Fig. 1 Scheme of the computational region around the vibrating airfoil in the channel

viscosity ν_T is determined with the aid of the Spalart–Allmaras turbulence model written in the ALE form [14]:

$$\frac{D^{\mathscr{A}}\tilde{\nu}}{Dt} + (\mathbf{u}-\mathbf{w}_g)\cdot\nabla\tilde{\nu} = \left[\sum_{i=1}^{2}\frac{\partial}{\partial x_i}\left(\frac{(\nu+\tilde{\nu})}{\beta}\frac{\partial\tilde{\nu}}{\partial x_i}\right) + \frac{c_{b_2}}{\beta}(\nabla\tilde{\nu})^2\right] + G(\tilde{\nu}) - Y(\tilde{\nu}), \tag{3}$$

for an additional quantity and equipped with the boundary condition $\tilde{\nu} = 0$ on $\Gamma_{Wt} \cup \Gamma_D$ and $\partial\tilde{\nu}/\partial\mathbf{n} = 0$ on Γ_O.

The turbulent viscosity ν_T is defined as

$$\nu_T = \tilde{\nu} f_{v_1}, \qquad f_{v_1} = \frac{\xi^3}{\xi^3 + c_v^3}, \qquad \xi = \frac{\tilde{\nu}}{\nu}.$$

and $G(\tilde{\nu})$ and $Y(\tilde{\nu})$ are functions of the tensor of rotation of the mean velocity ($\omega_{ij} = \frac{1}{2}(\frac{\partial u_i}{\partial x_j} - \frac{\partial u_j}{\partial x_i})$) depending on the wall distance y:

$$G(\tilde{\nu}) = c_{b_1}\widetilde{S}\tilde{\nu}, \quad Y(\tilde{\nu}) = c_{w_1}\frac{\tilde{\nu}^2}{y^2}\left(\frac{1+c_{w_3}^6}{1+c_{w_3}^6/g^6}\right)^{\frac{1}{6}}, \quad \widetilde{S} = \left(S + \frac{\tilde{\nu}}{\kappa^2 y^2} f_{v_2}\right),$$

$$f_{v_2} = 1 - \frac{\chi}{1+\chi f_{v_1}}, \quad g = r + c_{w_2}(r^6 - r), \quad r = \frac{\tilde{\nu}}{\widetilde{S}\kappa^2 y^2}, \quad S = \sqrt{2\sum_{i,j}\omega_{ij}^2},$$

The following constants are used $c_{b_1} = 0.1355$, $c_{b_2} = 0.622$, $\beta = \frac{2}{3}$, $c_v = 7.1$, $c_{w_2} = 0.3$, $c_{w_3} = 2.0$, $\kappa = 0.41$, $c_{w_1} = c_{b_1}/\kappa^2 + (1 + c_{b_2})/\beta$.

3 Structural Model

The profile can vertically vibrate with the displacement $h(t)$ and rotate around the elastic axis EA with the rotation angle α. The elastic support of the profile on translational and rotational springs with a bending stiffness k_h and torsion stiffness k_α is shown in Fig. 2. The airfoil motion is described by nonlinear equations for large vibrating amplitudes

$$\begin{aligned} m\ddot{h} + S_\alpha\, \ddot{\alpha}\, \cos\alpha - S_\alpha \dot{\alpha}^2 \sin\alpha + k_h h &= -L(t), \\ S_\alpha \ddot{h} \cos\alpha + I_\alpha \ddot{\alpha} + k_\alpha \alpha &= M(t). \end{aligned} \tag{4}$$

where m is the mass of the airfoil, S_α is the static moment and I_α is the inertia moment around the elastic axis. The pressure and viscous forces acting on the vibrating airfoil in fluid result in the lift force $L(t)$ and the torsional moment $M(t)$ defined by

$$L = -l \int_{\Gamma_{Wt}} \sum_{j=1}^{2} \tau_{2j} n_j \, dS, \qquad M = l \int_{\Gamma_{Wt}} \sum_{i,j=1}^{2} \tau_{ij} n_j r_i^{\mathrm{ort}} \, dS, \tag{5}$$

where

$$\begin{aligned} \tau_{ij} &= \rho \left[-p\delta_{ij} + (\nu + \nu_T) \left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} \right) \right], \\ r_1^{ort} &= -(x_2 - x_{EA2}), r_2^{ort} = (x_1 - x_{EA1}), \end{aligned}$$

τ_{ij} we denotes the components of the stress tensor, δ_{ij} the Kronecker symbol, $\mathbf{n}$ is the unit outer normal to $\partial\Omega_t$ on Γ_{Wt} (pointing into the profile) and $x_{EA} = (x_{EA1}, x_{EA2})$ is the position of the elastic axis EA.

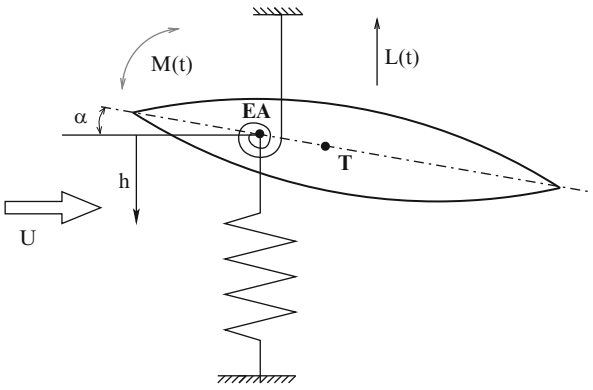


Fig. 2 Scheme of the elastically supported DCA profile

4 Numerical Approximation of the Flow Model

In order to solve the problem numerically, we start from the time discretization of the flow model. The ALE derivative is approximated by a two step backward difference formula. The problem discretized in time is solved by the FEM. The construction of the finite element space is based on a triangulation $\mathcal{T}_\Delta$ of a polygonal approximation of the computational domain Ω_t at time t.

In the finite element solution of incompressible Navier–Stokes equations several important obstacles need to be overcome. First, it is necessary to take into account that the finite element velocity/pressure pair has to be suitably chosen in order to satisfy the Babusjka-Brezzi condition, which guarantees the stability of the scheme see, e.g., [7]. In practical computations, the finite element spaces are defined over a triangulation $K \in \mathcal{T}_\Delta$, formed by a finite number of closed triangles $K \in \mathcal{T}_\Delta$. In our computations, the well-known Taylor-Hood P2/P1 conforming elements are used for the velocity/pressure approximation. This means that the finite element approximation of the pressure p_Δ is a piecewise linear function and the approximation of the velocity $\mathbf{u}_\Delta$ is a piecewise quadratic vector-valued function.

The standard Galerkin discretization may produce approximate solutions suffering from spurious oscillations for high Reynolds numbers. In order to avoid this drawback, the stabilization via streamline-diffusion/Petrov-Galerkin technique is applied - see, e.g., [6, 11]. Moreover, it is necessary to design carefully the computational mesh, using adaptive grid refinement in order to allow an accurate resolution of time oscillating thin boundary layers, wakes and vortices. We use the anisotropic mesh adaptation technique [3] for the construction and adaptive refinement of the mesh.

The nonlinear Spalart–Allmaras equation (5) is discretized by piecewise linear elements. In order to guarantee the positivity of the function $\tilde{\nu}$ needs to preserve, the SUPG/GLS stabilization applied as in [3]. However, the use of SUPG/GLS stabilization still does not avoid local oscillations near sharp layers, which can lead to negative viscosity. In order to solve this problem, the additional artificial viscosity stabilizing procedure based on crosswind diffusion is introduced, cf. [2].

5 Numerical Results

In this section the approximation of flow around elastically supported vibrating profile in wind tunnel is considered. The double circular arc (DCA) 18% profile with the chord length $b = 120$ mm and thickness 21.6 mm was installed in the test section of the wind tunnel. The test section was 80 mm wide and 210 mm high. Centre of rotation (EA) of the profile was at 1/3 of the chord behind the leading edge, for more details on the measurement see [13]. Experimentally established eigenfrequencies f_1 and f_2 and damping of the system for zero airflow velocity are presented in Table 1. By increasing the oncoming flow velocity both frequencies converged to the flutter frequency 20.4 Hz at the critical velocity at about Mach

Table 1 Dynamic characteristics of the model (M = 0)

Mode no	Natural frequency	Damping ratio	Mode shape
1	18.38 [Hz]	3.22 [%]	Translation
2	38.13 [Hz]	0.96 [%]	Torsion
3	146.9 [Hz]	0.72 [%]	Parasitic

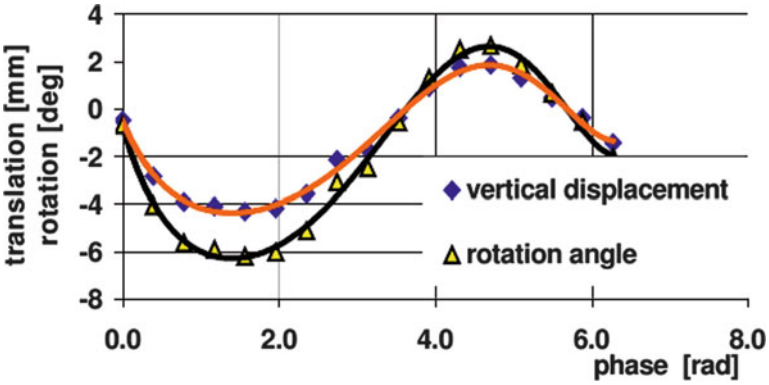


Fig. 3 Measured and prescribed (full line) angle of attack α and vertical translation h during one period of the self-excited airfoil motion

number M = 0.38 (Reynolds number $Re = U_\infty b/\nu = 1.0 \times 10^6$). Above this velocity, the system become unstable with rapid increase of vibration amplitudes resulting in the self-excited motion of the profile in a limit cycle oscillations (LCO).

The numerical simulation of the aeroelastic system behavior bellow and after the loss of the stability was considered in, e.g., [10]. Here, the flow around the airfoil is numerically simulated for the prescribed airfoil motion given by the attack angle α and vertical translation h of the elastic axis corresponding to the measured self-vibration regime, see Fig. 3. The oscillation frequency was 20.4 Hz, the angle varied from $-6°$ to $3°$ and the translation h from mm to $+2$ mm. The following input parameters were considered in the numerical computations: oncoming airflow velocity $U_0 = 130\,\mathrm{m\,s^{-1}}$, air density $\rho = 1.225\,\mathrm{kg\,m^{-3}}$, kinematic viscosity $\nu = 1.5 \times 10^{-5}\,\mathrm{m\,s^{-2}}$ and total pressure $p_0 = 9761.8\,\mathrm{Pa}$ in the oncoming flow (approximately equal to the atmospheric pressure).

The numerical approximation of flow velocity and pressure during one LCO is shown in Fig. 4. The pressure measurement and the numerical approximation of pressure computed along the profile on the upper and lower surfaces during one vibration period are presented in Fig. 5. Figure 6 shows the comparison of the measured lift coefficient and its numerical approximation.

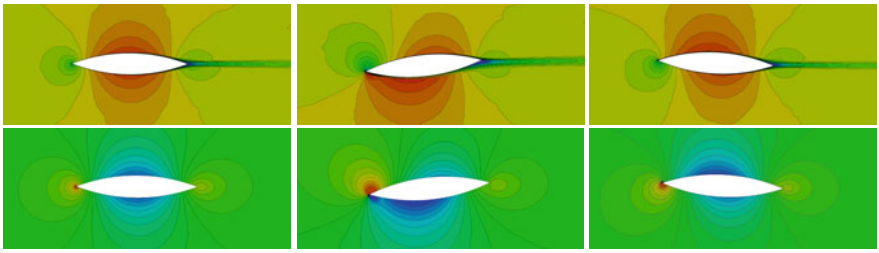


Fig. 4 Numerical approximation of velocity magnitude (*up*) and pressure distribution (*down*) during LCO

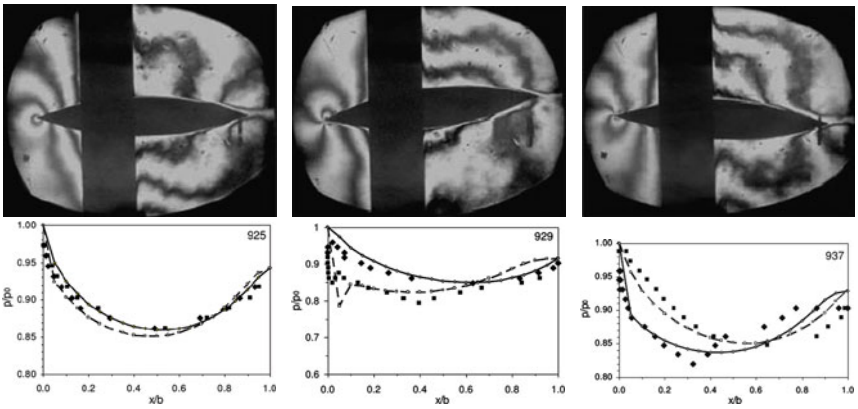


Fig. 5 Interferograms measured during one LCO (*up*) and comparison of experimental pressure distribution with numerical approximation (*down*)

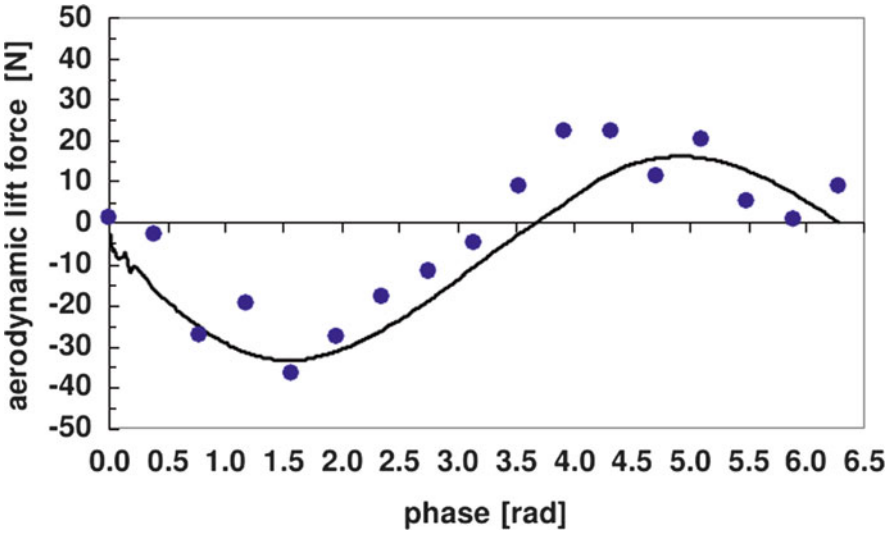


Fig. 6 Resulting measured and computed (full line) aerodynamic lift force during one period of the self excited airfoil motion.

6 Conclusion

Aeroelastic model of a double arc circle profile was investigated in wind tunnel in the regime of flutter instability at the Mach number $M = 0.38$. The method developed for the numerical simulation of airfoil aeroelastic behavior in turbulent flow was successfully validated by unique experimental data in case of the prescribed periodic airfoil vibrations with large amplitudes for flow velocities above the instability threshold for flutter.

Acknowledgements The authors acknowledge the financial support of the Grant Agency of Academy of Sciences of the Czech Republic by the project No. IAA200760613 Computer Modelling of Aeroelastic Phenomena for Real Fluid Flowing past Vibrating Airfoils Particularly after the Loss of System Stability. The research was also supported under the Research Plan MSM 6840770003 of the Ministry of Education of the Czech Republic.

References

1. Bathe, K.J.: Computational Fluid and Solid Mechanics. Elsevier, Amsterdam (2007)
2. Codina, R.: A discontinuity capturing crosswind-dissipation for the finite element solution of the convection diffusion equation. Comp. Meth. Appl. Mech. Eng., 110, 325–342 (1993)
3. Dolejší, V.: Anisotropic mesh adaptation technique for viscous flow simulation. East W. J. Numer. Math., 9, 1 (2001)
4. Dowell, E.H.: A Modern Course in Aeroelasticity. Kluwer, Dordecht (1995)
5. Dubcová, L., Feistauer, M., Horáček, J., Sváček, P.: Numerical simulation of airfoil vibrations induced by turbulent flow. J. Comput. Appl. Math., 218(1), 34–42 (2008)
6. Gelhard, T., Lube, G., Olshanskii, M.A, Starcke, J.H.: Stabilized finite element schemes with LBB-stable elements for incompressible flows. J. Comput. Appl. Math., 177, 243–267 (2005)
7. Girault, V., Raviart, P.A.: Finite Element Methods for the Navier–Stokes Equations, Springer, Berlin (1986)
8. Lube, G.: Stabilized Galerkin finite element methods for convection dominated and incompressible flow problems. Num. Anal. Math. Model., 29, 85–104 (1994)
9. Nomura, T., Hughes, T.J.R.: An arbitrary Lagrangian–Eulerian finite element for interaction of fluid and a rigid body. Comput. Meth. Appl. Mech. Eng., 95, 115–138 (1992)
10. Růžička, M., Feistauer, M., Horáček, J. and Sváček, P.: Interaction of incompressible flow and a moving airfoil. Electron. Trans. Numer. Anal., 32, 123–133 (2008)
11. Sváček, P., Feistauer, M., Horáček, J.: Numerical simulation of flow induced airfoil vibrations with large amplitudes. J. Fluids Struct., 23, 391–411 (2007)
12. Turek, S.: Efficient Solvers for Incompressible Flow Problems: An Algorithmic and Computational Approach. Springer, Berlin (1999)
13. Vlček, V., Horáček, J., Luxa, M., Veselý J.: Visualization of unsteady flow around a vibrating profile. 9th International Conference on Flow Induced Vibration FIV 2008, Praha, 30.6.–3.7.2008, Flow-Induced Vibration (Zolotarev, I., Horzek, J. eds.). Praha: Institute of Thermomechanics, Academy of Sciences of the Czech Republic, v.v.i., pp. 531–536 (2008)
14. Wilcox, D.C.: Turbulence modeling for CFD. DCW Industries, La Canada, California (1993)

Multigrid Methods for Elliptic Optimal Control Problems with Neumann Boundary Control

Stefan Takacs and Walter Zulehner

Abstract In this article we discuss multigrid methods for solving discretized optimality systems for elliptic optimal control problems. We concentrate on a model problem of tracking type with Neumann boundary control, whose optimality system is a linear system for the state y, the control u and the adjoined state p. An Uzawa-type smoother is used for the multigrid method. Moreover, we will compare this approach with standard smoothers, like damped Jacobi iteration applied to the normal equation of the Karush–Kuhn–Tucker system. A rigorous multigrid convergence analysis is presented for both smoothers.

1 Formulation of the Model Problem

We discuss the solution of optimal control problems of tracking type. Let Ω be a bounded convex and polygonal domain in $\mathbb{R}^2$ with boundary $\partial\Omega$. We want to minimize the functional

$$J(y,u) := \frac{1}{2}\|y - y_D\|^2_{L^2(\Omega)} + \frac{\gamma}{2}\|u\|_{L^2(\partial\Omega)}, \tag{1}$$

where y is the state variable and u is the control variable. Here, y_D is given and $\gamma > 0$ is some fixed regularization or cost parameter.

The minimization is done subject to the following constraint: the state variable fulfills some elliptic boundary value problem (BVP) with Neumann boundary data u. For this paper we restrict ourselves to the simple case of a Laplace-type equation:

S. Takacs (✉)
Doctoral Program Computational Mathematics, Johannes Kepler University Linz, Austria
e-mail: stefan.takacs@dk-compmath.jku.at

W. Zulehner
Institute of Computational Mathematics, Johannes Kepler University Linz, Austria
e-mail: zulehner@numa.uni-linz.ac.at

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_92,

$$-\Delta y + y = 0 \text{ in } \Omega \qquad \text{and} \qquad \frac{\partial y}{\partial n} = u \text{ on } \partial\Omega. \tag{2}$$

The functions y and u live in standard Lebesgue and Sobolev spaces:

$$y \in H^1(\Omega) \qquad \text{and} \qquad u \in L^2(\partial\Omega). \tag{3}$$

Observe that for this setting the BVP is uniquely solvable in y for every given control u. The BVP (2) can be written in variational form:

$$(y, p)_{H^1(\Omega)} - (u, p)_{L^2(\partial\Omega)} = 0 \qquad \text{for all } p \in H^1(\Omega).$$

Based on the variational formulation, we introduce the Lagrange functional

$$\mathscr{L}(y, u, p) := \frac{1}{2}\|y - y_D\|^2_{L^2(\Omega)} + \frac{\gamma}{2}\|u\|^2_{L^2(\partial\Omega)} + (y, p)_{H^1(\Omega)} - (u, p)_{L^2(\partial\Omega)}.$$

Solving the original optimal control problem is equivalent to finding a saddle point of the Lagrange functional which leads to the first order optimality conditions (the Karush–Kuhn–Tucker system), given by: Find $(y, u, p) \in X := H^1(\Omega) \times L^2(\partial\Omega) \times H^1(\Omega)$ such that

$$\begin{aligned} (y, \tilde{y})_{L^2(\Omega)} \qquad\qquad + (p, \tilde{y})_{H^1(\Omega)} &= (y_D, \tilde{y})_{L^2(\Omega)} \\ \gamma\,(u, \tilde{u})_{L^2(\partial\Omega)} - (p, \tilde{u})_{L^2(\partial\Omega)} &= 0 \\ (y, \tilde{p})_{H^1(\Omega)} - (u, \tilde{p})_{L^2(\partial\Omega)} &= 0 \end{aligned} \tag{4}$$

holds for all $(\tilde{y}, \tilde{u}, \tilde{p}) \in X$.

The optimality system has a natural 2-by-2 block-structure:

$$\begin{aligned} a((y, u), (\tilde{y}, \tilde{u})) + b((\tilde{y}, \tilde{u}), p) &= (y_D, \tilde{y})_{L^2(\Omega)} \\ b((y, u), \tilde{p}) &= 0, \end{aligned}$$

where

$$\begin{aligned} a((y, u), (\tilde{y}, \tilde{u})) &:= (y, \tilde{y})_{L^2(\Omega)} + \gamma\,(u, \tilde{u})_{L^2(\partial\Omega)} \\ b((y, u), \tilde{p}) &:= (y, \tilde{p})_{H^1(\Omega)} - (u, \tilde{p})_{L^2(\partial\Omega)}. \end{aligned}$$

Observe that a is a symmetric and bounded bilinear form which is coercive on the kernel of b and b is a bounded bilinear form. Moreover b fulfills the inf-sup-condition

$$\inf_{0 \neq \tilde{p} \in H^1(\Omega)} \sup_{0 \neq (y,u) \in H^1(\Omega) \times L^2(\partial\Omega)} \frac{b((y, u), \tilde{p})}{\|(y, u)\|_{H^1(\Omega) \times L^2(\partial\Omega)}\, \|\tilde{p}\|_{H^1(\Omega)}} \geq C > 0,$$

which can be seen by plugging in $(y, u) := (\tilde{p}, 0)$.

By adding the three equations in (4) the optimality system can be rewritten as one single variational equation:

$$\text{Find } x \in X \text{ such that} \qquad \mathsf{a}(x, \tilde{x}) = \langle \mathscr{F}, \tilde{x} \rangle \qquad \text{for all } \tilde{x} \in X. \tag{5}$$

Using Brezzi's theorem we obtain:

Lemma 1. *Let $\gamma > 0$ be fixed. The problem* (4) *is well posed in the space X, i.e., there are constants $\underline{C} > 0$ and $\overline{C}$ such that*

$$\underline{C}\, \|x\|_X \leq \sup_{0 \neq \tilde{x} \in X} \frac{\mathsf{a}(x, \tilde{x})}{\|\tilde{x}\|_X} \leq \overline{C}\, \|x\|_X$$

for all $x \in X$. For every right-hand-side $\mathscr{F} \in X^*$ the problem (5) has a unique solution $x \in X$.

The discretization is done by standard techniques. For the model problem we use a family of meshes which is obtained based on some coarsest triangular mesh (grid level $k = 0$) and uniform refinement. For $k = 0, 1, \ldots$ we denote the size of the largest edge of the triangulation by h_k. Due to the fact that we have uniform refinement $h_k = 2^{-k} h_0$ holds.

The space of discretized functions $X_k = Y_k \times U_k \times P_k$ is constructed by the Courant element: $Y_k = P_k$ is the set of continuous and piecewise linear functions. U_k is the set of continuous and piecewise linear functions on the boundary. This setting allows us to show the statement of lemma 1 also if X is replaced by X_k.

Using the standard nodal basis, we can rewrite the optimality system (4) in matrix-vector notation as follows:

$$\underbrace{\left(\begin{array}{cc|c} M_k & 0 & K_k \\ 0 & \gamma M_{\Gamma\Gamma k} & -M_{\Gamma\Omega k}^T \\ \hline K_k & -M_{\Gamma\Omega k} & 0 \end{array}\right)}_{\mathscr{A}_k := \left(\begin{smallmatrix} A_k & B_k^T \\ B_k & 0 \end{smallmatrix}\right) :=} \underbrace{\begin{pmatrix} \underline{y}_k \\ \underline{u}_k \\ \underline{p}_k \end{pmatrix}}_{\underline{x}_k :=} = \underbrace{\begin{pmatrix} \underline{g}_k \\ 0 \\ 0 \end{pmatrix}}_{\underline{f}_k :=} \tag{6}$$

with mass matrices M_k, $M_{\Gamma\Omega k}$, $M_{\Gamma\Gamma k}$ and the stiffness matrix K_k. The symbols $\underline{y}_k, \underline{u}_k, \ldots$ denote the coordinate vectors of the corresponding functions $y_k, u_k, \ldots$ with respect to the nodal basis.

Possible multigrid approaches for such a 3-by-3 block formulation are all-at-once methods, where the multigrid idea is directly applied to the optimality system (see, e.g., [8] for distributed control), or block-preconditioned methods, where multigrid techniques are used for constructing components of the block preconditioner (see, e.g., [6] and [5]).

Moreover, multigrid methods based on a reduction to a 2-by-2-formulation (see, e.g., [1] for an all-at-once approach) or to a 1-by-1-formulation (see, e.g., [3]) have also been proposed.

In this paper we will concentrate on the all-at-once approach for (6).

2 Multigrid Solvers for Saddle Point Problems

Starting from an initial approximation $\underline{x}_k^{(0)}$ one step of the multigrid method for solving the discretized equation (6) on grid level k is given by:

- Apply ν smoothing steps

$$\underline{x}_k^{(0,m)} := \underline{x}_k^{(0,m-1)} + \hat{\mathscr{A}}_k^{-1}(\underline{f}_k - \mathscr{A}_k\, \underline{x}_k^{(0,m-1)}) \qquad \text{for } m = 1, \ldots, \nu \tag{7}$$

  with $\underline{x}_k^{(0,0)} := \underline{x}_k^{(0)}$.

- Apply the coarse-grid correction
  - Compute the defect and restrict it to the coarser grid
  - Solve the problem on the coarser grid
  - Prolongate and add the result

  If the problem on the coarser grid is solved exactly, then we obtain

$$\underline{x}_k^{(1)} := \underline{x}_k^{(0,\nu)} + I_{k-1}^k \mathscr{A}_{k-1}^{-1} I_k^{k-1}(\underline{f}_k - \mathscr{A}_k\, \underline{x}_k^{(0,\nu)})$$

  for the next iterate (two-grid method).

In practice the problem on grid level $k-1$ is done by applying one (V-cycle) or two (W-cycle) steps of the multigrid method, recursively. On grid level $k = 0$ the problem is solved exactly. The convergence of the two-grid method implies the convergence of the W-cycle multigrid method under weak assumptions.

The intergrid-transfer operators I_{k-1}^k and I_k^{k-1} are chosen in a canonical way: we use the canonical embedding for I_{k-1}^k and its adjoined as restriction operator I_k^{k-1}.

The smoother will be specified in Sect. 3.

The classical convergence theory of multigrid methods is based on two properties:

- Smoothing property:

$$|||\underline{x}_k^{(0,\nu)} - \underline{x}_k|||_{2,k} \leq \eta(\nu)|||\underline{x}_k^{(0)} - \underline{x}_k|||_{0,k} \tag{8}$$

  should hold for some function $\eta(\nu)$ independent of k with $\lim_{\nu\to\infty} \eta(\nu) = 0$.
- Approximation property:

$$|||\underline{x}_k^{(1)} - \underline{x}_k|||_{0,k} \leq C_A|||\underline{x}_k^{(0,\nu)} - \underline{x}_k|||_{2,k} \tag{9}$$

  should hold for some constant $C_A > 0$ independent of k.

We have the freedom to choose two norms in (8) and (9). This is done in the following way:

We first introduce the norm $\|\cdot\|_{X_k^-}$ by replacing in $\|\cdot\|_X$ all H^1-norms by L^2-norms scaled by the factor h_k^{-1}:

$$\|(y_k, u_k, p_k)\|_{X_k^-}^2 := h_k^{-2}\|y_k\|_{L^2(\Omega)}^2 + \|u_k\|_{L^2(\partial\Omega)}^2 + h_k^{-2}\|p_k\|_{L^2(\Omega)}^2.$$

This corresponds to a norm for $(\underline{y}_k, \underline{u}_k, \underline{p}_k)$ involving mass matrices. If the mass matrices are replaced by properly scaled identity matrices, we obtain the desired norm $|||\cdot|||_{0,k}$, given by:

$$|||(\underline{y}_k, \underline{u}_k, \underline{p}_k)|||_{0,k}^2 := \Bigg(\underbrace{\begin{pmatrix} I & & \\ & h_k I & \\ & & I \end{pmatrix}}_{\mathscr{L}_k :=} \begin{pmatrix} \underline{y}_k \\ \underline{u}_k \\ \underline{p}_k \end{pmatrix}, \begin{pmatrix} \underline{y}_k \\ \underline{u}_k \\ \underline{p}_k \end{pmatrix} \Bigg)_{\ell^2}.$$

According to standard techniques, we choose $|||\cdot|||_{2,k}$ as residual norm corresponding to $|||\cdot|||_{0,k}$, i.e.,

$$|||\underline{x}_k|||_{2,k} := \sup_{\tilde{\underline{x}}_k \in \mathbb{R}^n} \frac{(\mathscr{A}_k\, \underline{x}_k, \tilde{\underline{x}}_k)}{|||\tilde{\underline{x}}_k|||_{0,k}}.$$

3 Construction of Smoothers

Next we construct two simple iterative methods fulfilling the smoothing property (8).

The first kind of smoothers, we want to discuss, are *Uzawa-type smoothers* which have already been successfully applied to distributed control problems (e.g., [8]). These methods can also be extended to Neumann boundary control problems.

We construct the preconditioner $\hat{\mathscr{A}}_k$ in (7) based on the block-LU-factorization of $\mathscr{A}_k$: We have

$$\mathscr{A}_k = \begin{pmatrix} A_k & B_k^T \\ B_k & 0 \end{pmatrix} = \begin{pmatrix} A_k & 0 \\ B_k & -S_k \end{pmatrix} \begin{pmatrix} I & A_k^{-1} B_k^T \\ 0 & I \end{pmatrix},$$

where $S_k := B_k A_k^{-1} B_k^T$ is the Schur-complement. Based on this decomposition we define the preconditioner $\hat{\mathscr{A}}_k$ by replacing A_k and S_k by diagonal matrices $\hat{A}_k$ and $\hat{S}_k$:

$$\hat{\mathscr{A}}_k := \begin{pmatrix} \hat{A}_k & 0 \\ B_k & -\hat{S}_k \end{pmatrix} \begin{pmatrix} I & \hat{A}_k^{-1} B_k^T \\ 0 & I \end{pmatrix}.$$

Then $\hat{\mathscr{A}}_k$ can be inverted easily.

The main issue is how to choose the matrices $\hat{A}_k$ and $\hat{S}_k$. Normally, one would expect to choose $\hat{A}_k$ as the diagonal part of A_k. Instead we propose to choose for $\hat{A}_k$ the $(1,1)$-block of $\mathscr{L}_k$. For $\hat{S}_k$ we take the corresponding inexact Schur-complement. This leads to

$$\hat{A}_k := \frac{1}{\omega} \begin{pmatrix} I & \\ & h_k I \end{pmatrix} \qquad \text{and} \qquad \hat{S}_k := \frac{1}{\sigma} \operatorname{diag}(B_k \hat{A}_k^{-1} B_k^T)$$

with additional damping parameters ω and σ which are chosen independent of k and such that

$$\hat{A}_k \geq A_k \qquad \text{and} \qquad \hat{S}_k \geq S_k \tag{10}$$

holds. This is possible, as we can choose ω and σ equal to the reciprocal of the number of non-zero entries of A_k or S_k, respectively, which are bounded.

An *alternative approach* is to construct smoothers that are based on the *normal equation* $\mathscr{A}_k^* \mathscr{A}_k \underline{x}_k = \mathscr{A}_k^* \underline{f}_k$, where $\mathscr{A}_k^*$ denotes the adjoined of $\mathscr{A}_k$ with respect to the inner product corresponding to the norm $|||\cdot|||_{0,k}$. Using $\mathscr{L}_k$ we can formulate this in standard matrix-vector notation:

$$\mathscr{A}_k^T \mathscr{L}_k^{-1} \mathscr{A}_k \underline{x}_k = \mathscr{A}_k^T \mathscr{L}_k^{-1} \underline{f}_k.$$

We can apply some standard smoother which is applicable to symmetric positive definite problems, like the damped Jacobi iteration:

$$\underline{x}_k^{(0,m)} := \underline{x}_k^{(0,m-1)} + \tau \operatorname{diag}(\mathscr{A}_k^T \mathscr{L}_k^{-1} \mathscr{A}_k)^{-1} \mathscr{A}_k^T \mathscr{L}_k^{-1} (\underline{f}_k - \mathscr{A}_k \underline{x}_k^{(0,m-1)}),$$

where the parameter τ is chosen such that the smallest eigenvalue of the iteration matrix is non-negative.

4 Convergence Analysis

A convergence analysis for *distributed control problems* was already done based on approximation and smoothing property, e.g., [8]. The approximation property was shown following the ideas from [2].

Alternative approaches to obtain convergence results are local mode analysis for distributed control (e.g., [1]) and compactness arguments (e.g., [3] and [1]).

Here we follow the lines of the analysis in [8]. We can show in our framework:

Lemma 2 (Smoothing property). *The smoothing property holds for both alternatives of smoothers we discussed in this work with smoothing rate* $\eta(\nu) := C_S/\sqrt{\nu}$ *where* $C_S > 0$ *is a constant independent of* k*, i.e.,*

$$|||\underline{x}_k^{(0,\nu)} - \underline{x}_k|||_{2,k} \leq C_S \frac{1}{\sqrt{\nu}} |||\underline{x}_k^{(0)} - \underline{x}_k|||_{0,k}.$$

The proof for the Uzawa-type smoother follows the proof in [8] which is based on [7]. The analysis for the smoothers based on the normal equation uses the fact that the normal equation is symmetric and positive definite. Therefore the analysis can be done by standard techniques.

We can show the *approximation property* similar to the proof given in [2] using the following regularity result: For $f := (f_1, f_2) \in L^2(\Omega) \times L^2(\Omega)$ let $x \in X$ solve

$$\mathsf{a}(x, \tilde{x}) = (f_1, \tilde{y})_{L^2(\Omega)} + (f_2, \tilde{p})_{L^2(\Omega)} \quad \text{for all } \tilde{x} = (\tilde{y}, \tilde{u}, \tilde{p}) \in X$$

Since Ω is convex, it follows from standard H^2-regularity results for y and p that

$$\|x\|_{H^2(\Omega)\times H^1(\partial\Omega)\times H^2(\Omega)} \leq C \ \|f\|_{L^2(\Omega)\times L^2(\Omega)}.$$

Using this result we can show the following lemma.

Lemma 3 (Approximation property). *There is a constant $C_A > 0$ such that*

$$|||\underline{x}_k^{(1)} - \underline{x}_k|||_{0,k} \leq C_A |||\underline{x}_k^{(0,\nu)} - \underline{x}_k|||_{2,k}.$$

Lemma 2 and lemma 3 lead to:

Theorem 1 (Convergence of the two-grid method). *The two-grid method converges for sufficiently large values of ν:*

$$|||\underline{x}_k^{(1)} - \underline{x}_k|||_{0,k} \leq q \ |||\underline{x}_k^{(0)} - \underline{x}_k|||_{0,k}$$

with convergence rate $q := C_A C_S / \sqrt{\nu} < 1$ independent of the grid level k.

This implies the convergence of the W-cycle multigrid method, see e.g., [4].

5 Numerical Results

The numerical tests were done for the unit square. The coarsest mesh (level $k = 0$) was constructed by separating the domain into two congruent triangles. The refinement was done by splitting each triangle into four congruent sub-triangles. The parameter γ was set to 1. The parameters ω and σ for the Uzawa-type smoother were chosen such that (10) holds (on a coarse level). For the normal equation method it turned out that it suffices to choose $\tau := 1/4$. Starting with randomly chosen initial approximations the multigrid iteration was performed until the $|||.|||_{0,k}$-norm of the error was reduced by a factor of 10^{-5}.

Table 1 shows the number of iterations and the computing time for the W-cycle.

According to the theory the number of iterations is independent of the grid level. Moreover, the number of iterations decreases as the number of smoothing steps is

Table 1 Number of iterations and computing time

Level	Number of unknowns	Uzawa-type				Jacobi-type			
		Smoothing steps ν							
		$2+2$		$6+6$		$2+2$		$6+6$	
5	2,306	30	0.7 s	9	0.4 s	10	0.3 s	5	0.3 s
6	8,706	31	3.0 s	10	2.3 s	10	1.2 s	5	1.5 s
7	33,794	31	12.9 s	10	10.6 s	10	5.0 s	5	6.6 s
8	133,122	31	57.8 s	11	52.3 s	11	24.2 s	5	28.9 s
9	528,386	31	237.2 s	11	222.6 s	10	91.7 s	5	121.0 s

increased. The computing time increases linearly with the number of unknowns (optimal complexity). The computing time shows that the performance of the multigrid method with the Jacobi smoother applied to the normal equation is better.

Roughly the same number of iterations were observed for the V-cycle, for which the computing time is about 30% less than for the W-cycle.

We compared this multigrid methods with a Bramble–Pasciak CG-method implemented along the guidelines of [5]. In terms of computing times the performance of the Bramble–Pasciak CG-method lies between the performance of the V-cycle and the W-cycle proposed in this paper. While the proposed multigrid methods can also be applied to problems with singular $(1,1)$-block (and show reasonable efficiency in selected experiments), this is not possible for the Bramble-Pasciak CG-method, which requires a non-singular $(1,1)$-block.

6 Conclusion and Further Work

This work shows that the results for the Uzawa-type smoother and the strategy proposed in [8] for the distributed control problem carry over to the boundary control problem. It was possible to generate comparable results also for the Jacobi-type smoother applied to the normal equations. The more general approach of the method based on the normal equation will hopefully allow an extension of the method for a larger class of optimal control problems. Further work has to be done to find smoothers that are robust in the parameter γ.

Acknowledgements The work is supported by the Austrian Science Fund (FWF) under grant W1214/DK12.

References

1. Borzi, A., Kunisch, K., Kwak, D.Y. Accuracy and convergence properties of the finite difference multigrid solution of an optimal control optimality system. *SIAM J. Control. Optim.*, 41(5):1477–1497, 2003

2. Brenner, S.C. Multigrid methods for parameter dependent problems. *RAIRO, Modélisation Math. Anal. Numér.*, 30:265–297, 1996
3. Hackbusch, W. Fast solution of elliptic control problems. *J. Optim. Theor. Appl.*, 31:565–581, 1980
4. Hackbusch, W. *Multi-Grid Methods and Applications*. Springer, Berlin, 1985
5. Rees, T., Stoll, M. Block-triangular preconditioners for PDE-constrained optimization. *Numer. Linear Algebra Appl.*, 2010 (To appear) (http://onlinelibrary.wiley.com/doi/10.1002/nla.693/abstract)
6. Rees, T., Dollar, S., Wathen, A. Optimal solvers for PDE-Constrained Optimization. *SIAM J. Sci. Comput.*, 32(1):271–298, 2010
7. Schöberl, J., Zulehner, W. On Schwarz-type smoothers for saddle point problems. *Numer. Math.*, 95:377–399, 2003
8. Simon, R., Zulehner, W. On Schwarz-type smoothers for saddle point problems with applications to PDE-constrained optimization problems. *Numer. Math.*, 111:445–468, 2009

Extension of the Complete Flux Scheme to Time-Dependent Conservation Laws

J.H.M. ten Thije Boonkkamp and M.J.H. Anthonissen

Abstract We present the stationary and transient complete flux schemes for the advection-diffusion-reaction equation. In the first scheme, the numerical flux is derived from a local BVP for the stationary equation. The transient scheme is an extension, since it includes the time derivative in the flux computation. The resulting semidiscretization is an implicit ODE system, which has much smaller dissipation and dispersion errors than the semidiscretization based on the stationary flux, at least for smooth problems. Both schemes are validated for a test problem.

1 Introduction

Conservation laws are ubiquitous in continuum physics, they occur in disciplines like fluid mechanics, combustion theory, plasma physics, semiconductor theory etc. These conservation laws are often of advection-diffusion-reaction type, describing the interplay between different processes such as advection or drift, diffusion or conduction and (chemical) reaction or recombination/generation.

The numerical solution of these equations requires accurate and robust space discretization and time integration methods and efficient (iterative) solution methods for the resulting algebraic system. In this paper we address the first two topics for the model equation

$$\frac{\partial \varphi}{\partial t} + \frac{\partial}{\partial x}\left(u\varphi - \varepsilon \frac{\partial \varphi}{\partial x}\right) = s, \tag{1}$$

where u is the advection velocity, $\varepsilon \geq \varepsilon_{\min} > 0$ a diffusion/conduction coefficient and s a source term. Associated with (1) we introduce the flux f defined by

$$f := u\varphi - \varepsilon \frac{\partial \varphi}{\partial x}. \tag{2}$$

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

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_93,

For space discretization we use the finite volume method (FVM) [1] in combination with the complete flux (CF) scheme for the numerical fluxes [4, 5]. For stationary equations, the CF approximation is based on the solution of a local boundary value problem for the entire equation and is therefore an extension of exponentially fitted schemes, which are based on the corresponding constant coefficient, homogeneous equation; see e.g., [3]. The CF approximation is second order accurate, even for strongly advection dominated flow, and gives rise to a tridiagonal system.

In this paper we consider the extension to time-dependent equations. A first obvious choice would be to combine the stationary CF approximation with a suitable time integration method. We refer to this flux approximation as the stationary complete flux (SCF) scheme. However, for strong advection, the space discretization error reduces to first order. Therefore, we propose to include the time derivative $\partial\varphi/\partial t$ already in the numerical approximation of the flux. More precisely, we put the time derivative in the source term and solve the corresponding quasi-stationary BVP. The resulting scheme, referred to as the transient complete flux (TCF) scheme, does not have this drawback, and moreover, has usually much smaller dissipation and dispersion errors than the SCF scheme.

We have organized our paper as follows. The SCF scheme is briefly summarized in Sect. 2 and its extension to time-dependent equations is presented in Sect. 3. The SCF and TCF semidiscretizations are analysed in terms of dissipation and dispersion in Sect. 4. The performance of both schemes is shown in Sect. 5.

2 Numerical Approximation of the Stationary Flux

In this section we present the complete flux scheme for the stationary flux, which is based on the integral representation of the flux. The derivation is a summary of the theory in [4, 5].

The stationary conservation law can be written as $\mathrm{d}f/\mathrm{d}x = s$ with the flux f defined in (2). In the FVM we cover the domain with a finite number of control volumes (cells) I_j of size Δx. We choose the grid points x_j, where the variable φ has to be approximated, in the cell centres, the so-called cell centred approach; see e.g., [6]. Consequently, we have $I_j := [x_{j-1/2}, x_{j+1/2}]$ with $x_{j+1/2} := \frac{1}{2}(x_j + x_{j+1})$. Integrating the equation over I_j and applying the midpoint rule for the integral of s, we obtain the discrete conservation law

$$F_{j+1/2} - F_{j-1/2} = s_j\,\Delta x, \tag{3}$$

where $F_{j+1/2}$ is the numerical approximation of the flux f at the interface at $x = x_{j+1/2}$ and where $s_j := s(x_j)$.

The integral representation of the flux $f_{j+1/2} := f(x_{j+1/2})$ at the cell edge $x_{j+1/2}$ is based on the following model boundary value problem (BVP) for the variable φ

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(u\varphi - \varepsilon\frac{\mathrm{d}\varphi}{\mathrm{d}x}\right) = s, \quad x_j < x < x_{j+1}, \tag{4a}$$

$$\varphi(x_j) = \varphi_j, \quad \varphi(x_{j+1}) = \varphi_{j+1}. \tag{4b}$$

We like to emphasize that $f_{j+1/2}$ corresponds to the solution of the *inhomogeneous* BVP (4), implying that $f_{j+1/2}$ not only depends on u and ε, but also on the source term s. It is convenient to introduce the variables λ, P, Λ and S for $x \in (x_j, x_{j+1})$ by

$$\lambda := \frac{u}{\varepsilon}, \quad P := \lambda \Delta x, \quad \Lambda(x) = \int_{x_{j+1/2}}^{x} \lambda(\xi)\,\mathrm{d}\xi, \quad S(x) := \int_{x_{j+1/2}}^{x} s(\xi)\,\mathrm{d}\xi. \tag{5}$$

Here, P and Λ are the Peclet function and Peclet integral, respectively, generalizing the well-known (numerical) Peclet number. Integrating the differential equation (4a) from $x_{j+1/2}$ to x we get the integral balance $f(x) - f_{j+1/2} = S(x)$. Using the definition of Λ in (5), it is clear that the flux can be rewritten as $f = -\varepsilon \mathrm{e}^{\Lambda}\,\mathrm{d}\big(\varphi\,\mathrm{e}^{-\Lambda}\big)/\mathrm{d}x$. Substituting this into the integral balance and integrating from x_j to x_{j+1} we obtain the following expression for the flux

$$f_{j+1/2} = f^{(\mathrm{h})}_{j+1/2} + f^{(\mathrm{i})}_{j+1/2}, \tag{6a}$$

$$f^{(\mathrm{h})}_{j+1/2} = -\big(\mathrm{e}^{-\Lambda_{j+1}}\varphi_{j+1} - \mathrm{e}^{-\Lambda_j}\varphi_j\big) \Big/ \int_{x_j}^{x_{j+1}} \varepsilon^{-1}\mathrm{e}^{-\Lambda}\,\mathrm{d}x, \tag{6b}$$

$$f^{(\mathrm{i})}_{j+1/2} = -\int_{x_j}^{x_{j+1}} \varepsilon^{-1}\mathrm{e}^{-\Lambda} S\,\mathrm{d}x \Big/ \int_{x_j}^{x_{j+1}} \varepsilon^{-1}\mathrm{e}^{-\Lambda}\,\mathrm{d}x, \tag{6c}$$

where $f^{(\mathrm{h})}_{j+1/2}$ and $f^{(\mathrm{i})}_{j+1/2}$ are the homogeneous and inhomogeneous part, corresponding to the homogeneous and particular solution of (4), respectively.

In the following we assume that u and ε are constant; extension to variable coefficients is discussed in [4, 5]. In this case we can determine all integrals in (6b). Moreover, substituting the expression for $S(x)$ in (6c) and changing the order of integration, we can derive an alternative expression for the inhomogeneous flux. This way we obtain

$$f^{(\mathrm{h})}_{j+1/2} = -\frac{\varepsilon}{\Delta x}\big(B(P)\varphi_{j+1} - B(-P)\varphi_j\big), \tag{7a}$$

$$f^{(\mathrm{i})}(x_{j+1/2}) = \Delta x \int_0^1 G(\sigma; P) s(x_j + \sigma\Delta x)\,\mathrm{d}\sigma, \quad \sigma(x) := \frac{x - x_j}{\Delta x}. \tag{7b}$$

Here $B(z) := z/\big(\mathrm{e}^z - 1\big)$ is the Bernoulli function and $G(\sigma; P)$ the Green's function for the flux, given by

$$G(\sigma; P) = \begin{cases} \dfrac{1 - \mathrm{e}^{-P\sigma}}{1 - \mathrm{e}^{-P}} & \text{for} \quad 0 \le \sigma \le \frac{1}{2}, \\[2ex] -\dfrac{1 - \mathrm{e}^{P(1-\sigma)}}{1 - \mathrm{e}^{P}} & \text{for} \quad \frac{1}{2} < \sigma \le 1; \end{cases} \tag{8}$$

see Fig. 1.

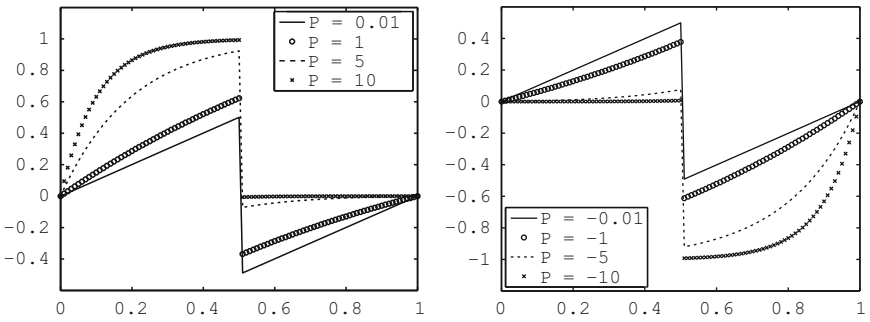


Fig. 1 Green's function for the flux for $P > 0$ (*left*) and $P < 0$ (*right*)

Next, we give the numerical flux $F_{j+1/2}$. For the homogeneous component $F^{(\mathrm{h})}_{j+1/2}$ we simply take (7a), i.e., $F^{(\mathrm{h})}_{j+1/2} = f^{(\mathrm{h})}_{j+1/2}$. Note that for dominant diffusion ($|P| \ll 1$) the integral (average) of $G(\sigma; P)$ is small, whereas for dominant advection ($|P| \gg 1$) $G(\sigma; P)$ has a clear bias towards the upwind side of the interval. For this reason we replace $s(x)$ in (7b) by its upwind value $s_{\mathrm{u},j+1/2}$, i.e., $s_{\mathrm{u},j+1/2} = s_j$ if $u \geq 0$ and $s_{\mathrm{u},j+1/2} = s_{j+1}$ if $u < 0$, and evaluate the resulting integral exactly. This way we obtain

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

where $W(z) := \big(\mathrm{e}^z - 1 - z\big)/\big(z\big(\mathrm{e}^z - 1\big)\big)$. From this expression it is clear that the inhomogeneous component is only of importance for dominant advection. We refer to (9) as the complete flux (CF) scheme, as opposed to the homogeneous flux (HF) scheme for which we only take into account $F^{(\mathrm{h})}_{j+1/2}$. Finally, substituting (9) in (3) we obtain the discretization

$$\frac{1}{\Delta x}\Big(F^{(\mathrm{h})}_{j+1/2} - F^{(\mathrm{h})}_{j-1/2}\Big) = \big(\tfrac{1}{2} + W(|P|)\big)s_j + \big(\tfrac{1}{2} - W(|P|)\big)s_{j(\mathrm{u})}, \tag{10}$$

where $j(\mathrm{u})$ is the index of the grid point upwind of j, i.e., $j(\mathrm{u}) = j - 1$ if $u \geq 0$ and $j(\mathrm{u}) = j + 1$ if $u < 0$.

3 Extension to Time-Dependent Conservation Laws

In this section we present the extension of the complete flux scheme to time-dependent conservation laws.

Equation (1) can be written as $\partial\varphi/\partial t + \partial f/\partial x = s$. Integrating this equation over the control volume I_j and applying the midpoint rule for the integrals of $\partial\varphi/\partial t$ and s, we obtain the semidiscrete conservation law

$$\dot{\varphi}_j \, \Delta x + F_{j+1/2} - F_{j-1/2} = s_j \, \Delta x, \tag{11}$$

where $\dot{\varphi}_j = \mathrm{d}\varphi_j/\mathrm{d}t$. Note that the numerical flux $F_{j+1/2}$ still depends on t.

For the numerical flux $F_{j+1/2}$ in (11) we have two options. First, we can simply take the stationary flux (9), henceforth referred to as the SCF scheme. Alternatively, we can take into account $\partial\varphi/\partial t$ if we determine the numerical flux from the following quasi-stationary BVP

$$\frac{\partial}{\partial x}\Big(u\varphi - \varepsilon\frac{\partial\varphi}{\partial x}\Big) = s - \frac{\partial\varphi}{\partial t}, \quad x_j < x < x_{j+1}, \tag{12a}$$

$$\varphi(x_j, t) = \varphi_j(t), \quad \varphi(x_{j+1}, t) = \varphi_{j+1}(t). \tag{12b}$$

Thus, we have a modified source term $\tilde{s} := s - \partial\varphi/\partial t$. Repeating the derivation in the previous section, we obtain

$$F_{j+1/2} = F^{(\mathrm{h})}_{j+1/2} + \big(\tfrac{1}{2} - W(P)\big)\big(s_{\mathrm{u},j+1/2} - \dot{\varphi}_{\mathrm{u},j+1/2}\big)\,\Delta x. \tag{13}$$

This flux contains the upwind value $\dot{\varphi}_{\mathrm{u},j+1/2}$ of the time derivative and is referred to as the transient complete flux (TCF) scheme. Analogous to the stationary case, we conclude that inclusion of the time derivative is only of importance for dominant advection.

Combining the expression in (13) with the semi-discrete conservation law (11) we find

$$\begin{aligned}\big(\tfrac{1}{2} + W(|P|)\big)\dot{\varphi}_j + \big(\tfrac{1}{2} - W(|P|)\big)\dot{\varphi}_{j(\mathrm{u})} + \frac{1}{\Delta x}\Big(F^{(\mathrm{h})}_{j+1/2} - F^{(\mathrm{h})}_{j-1/2}\Big) = \\ \big(\tfrac{1}{2} + W(|P|)\big)s_j + \big(\tfrac{1}{2} - W(|P|)\big)s_{j(\mathrm{u})}.\end{aligned} \tag{14}$$

Finally, we have to apply a suitable time integration method to (14), for which we will take the trapezoidal rule.

4 Dissipation and Dispersion of the Semidiscrete System

It is interesting to compare the SCF and TCF semidiscretizations in terms of dissipation (damping) and dispersion. Therefore, consider (1) with u and ε constant and $s = 0$. In the following we assume that $u \geq 0$; the analysis for $u \leq 0$ is similar. Following [2], we look for a planar wave solution

$$\varphi(x, t) = \mathrm{e}^{\mathrm{i}(\kappa x - \omega t)}, \tag{15}$$

where κ is the wave number and ω is the frequency. Substituting (15) in (1) we obtain the dispersion relation

$$\mathrm{i}\omega(\kappa) = \mathrm{i}u\kappa + \varepsilon\kappa^2. \tag{16}$$

The frequency ω determines the time evolution of the solution (15). Comparing the (exact) solution of (1) at two consecutive time levels, we can define the amplification factor $g = g(\kappa)$ as follows

$$g(\kappa) := \varphi(x, t_{n+1})/\varphi(x, t_n) = \mathrm{e}^{-\mathrm{i}\omega\Delta t}, \tag{17}$$

where $t_n := n\Delta t$ ($n = 0, 1, 2, \ldots$) and $\Delta t > 0$ is the time step. Note that g is independent of x. Combining the relations (16) and (17) we find the following amplification factor for (1), i.e.,

$$g(\psi) = \mathrm{e}^{-d\psi^2}\mathrm{e}^{-\mathrm{i}c\psi}, \tag{18}$$

with $d := \varepsilon\Delta t/\Delta x^2$ the diffusion number, $c := u\Delta t/\Delta x$ the Courant number and $\psi := \kappa\Delta x$ the phase angle ($0 \le \psi < \pi$).

We will now compute the amplification factors of the SCF and TCF semidiscretizations and compare these to (18). First, consider the SCF semidiscretization of (1), which coincides with the HF semidiscretization since $s = 0$, given by

$$\dot{\varphi}_j\,\Delta x + \frac{\varepsilon}{\Delta x}B^-\big(\varphi_j - \varphi_{j-1}\big) - \frac{\varepsilon}{\Delta x}B^+\big(\varphi_{j+1} - \varphi_j\big) = 0, \tag{19}$$

with $B^\pm := B(\pm P)$. Substituting $\varphi(x_j, t)$, with φ defined in (15), we obtain the discrete dispersion relation

$$\mathrm{i}\omega(\kappa) = \mathrm{i}u\kappa\frac{\sin\psi}{\psi} + \varepsilon\kappa^2\tfrac{1}{2}\big(B^+ + B^-\big)\Big(\frac{\sin\psi/2}{\psi/2}\Big)^2 =: \mathrm{i}u\kappa\xi + \varepsilon\kappa^2\eta. \tag{20}$$

The variables ξ and η in the right hand side define the deviation of ω from the expression in (16). From (17) and (20) we can derive the following expression for the amplification factor, i.e.,

$$g(\psi) = \mathrm{e}^{-d\psi^2\eta}\mathrm{e}^{-\mathrm{i}c\psi\xi}. \tag{21}$$

To quantify dissipation and dispersion, we define the (relative) amplitude error ϵ_a and the (relative) phase error ϵ_f as follows:

$$\epsilon_\mathrm{a}(\psi) := 1 - \mathrm{e}^{d\psi^2(1-\eta)}, \quad \epsilon_\mathrm{f}(\psi) := 1 - \xi. \tag{22}$$

Plots of ϵ_a and ϵ_f are given in Fig. 2 (solid lines).

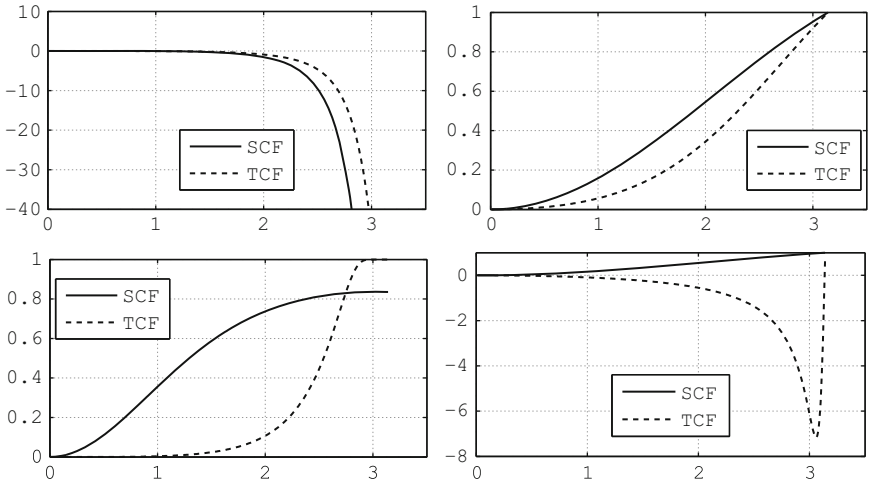


Fig. 2 The amplitude error (*left*) and the phase error (*right*). Parameter values are $P = 1$ (*top*), $P = 50$ (*bottom*) and $c = 1$

Next, consider the TCF semidiscretization of (1), which reads

$$\left(\frac{1}{2} - W^+\right)\dot{\varphi}_{j-1}\,\Delta x + \left(\frac{1}{2} + W^+\right)\dot{\varphi}_j\,\Delta x + \frac{\varepsilon}{\Delta x}B^-\left(\varphi_j - \varphi_{j-1}\right) - \frac{\varepsilon}{\Delta x}B^+\left(\varphi_{j+1} - \varphi_j\right) = 0, \tag{23}$$

with $W^+ := W(P)$. Note that substituting $W^+ = \frac{1}{2}$ in (23) we recover the SCF semidiscretization (19). Once more substituting $\varphi(x_j, t)$ we find the discrete dispersion relation

$$\mathrm{i}\omega(\kappa) = \frac{\varepsilon}{\Delta x^2}\frac{-B^+\left(\mathrm{e}^{\mathrm{i}\psi} - 1\right) + B^-\left(1 - \mathrm{e}^{-\mathrm{i}\psi}\right)}{\left(\frac{1}{2} - W^+\right)\mathrm{e}^{-\mathrm{i}\psi} + \left(\frac{1}{2} + W^+\right)} =: \mathrm{i}u\kappa\xi + \varepsilon\kappa^2\eta. \tag{24}$$

This relation implicitly defines the factors ξ and η, given by

$$\xi = \frac{\sin\psi}{\psi}\frac{\cos^2\psi/2 + F_1(P)\sin^2\psi/2}{\cos^2\psi/2 + 4W^2(P)\sin^2\psi/2}, \tag{25a}$$

$$\eta = \left(\frac{\sin\psi/2}{\psi/2}\right)^2\frac{\cos^2\psi/2 + F_2(P)\sin^2\psi/2}{\cos^2\psi/2 + 4W^2(P)\sin^2\psi/2}, \tag{25b}$$

where $F_1(z) := 2W(z) - \big(B(z) + B(-z)\big)\big(2W(z) - 1\big)/z$ and $F_2(z) := W(z)\big(B(z) + B(-z)\big)$. The corresponding amplification factor is given in (21). Combining (22)

with (25) we can determine the amplitude and phase errors, which are shown in Fig. 2 (dashed lines).

From these figures we conclude the following. For dominant diffusion, i.e., small P, the TCF scheme has slightly smaller amplitude and phase errors than the SCF scheme. On the other hand, for dominant advection, i.e., large P, the amplitude and phase errors of the TCF scheme are significantly smaller than those of the SCF scheme, at least for low wave number modes with typically $0 \leq \psi \leq \frac{1}{2}$. For smooth solutions this can always be attained if we choose Δx small enough. However, for high wave number modes, say $2 \leq \psi \leq \pi$, the dispersion error of the TCF scheme is large. Therefore, spurious oscillations cannot be excluded for nonsmooth solutions with steep interior/boundary layers. These have to be controlled by a (dissipative) time integration method.

5 Numerical Example

In this section we apply the SCF and TCF schemes to a model problem to asses their (order of) accuracy. We consider both diffusion-dominated and advection-dominated flow.

Consider the test equation [6]

$$\frac{\partial \varphi}{\partial t} + u\frac{\partial \varphi}{\partial x} - \varepsilon \frac{\partial^2 \varphi}{\partial x^2} = s, \quad s(x,t) = \beta^2 \varepsilon \cos\big(\beta(x-ut)\big), \quad 0 < x < 1, \quad t > 0. \tag{26}$$

Initial and boundary condition are chosen such that the exact solution is given by $\varphi(x,t) = \cos\big(\beta(x-ut)\big) + \mathrm{e}^{-\alpha^2 \varepsilon t} \cos\big(\alpha(x-ut)\big)$. We take the following parameter values: $\alpha = 4\pi$, $\beta = 2\pi$, $u = 1.1$ and $\varepsilon = 2 \times 10^{-2}$ (dominant diffusion) or $\varepsilon = 10^{-8}$ (dominant advection). Furthermore, we choose $\Delta x = \Delta t =: h$. To determine the accuracy of a numerical solution we compute the average error $e_h := h\,||\boldsymbol{\varphi} - \boldsymbol{\varphi}^*||_1$ at $t = 1$, where $\boldsymbol{\varphi}^*$ denotes the exact solution restricted to the grid, as a function of the reciprocal grid size h^{-1}. Table 1 shows e_h and the reduction factors $e_h/e_{h/2}$. Clearly, for $\varepsilon = 2 \times 10^{-2}$, $e_h/e_{h/2} \to 4$ for $h \to 0$ for both the SCF and

Table 1 Average errors and error quotients

	$\varepsilon = 2 \times 10^{-2}$				$\varepsilon = 10^{-8}$			
	SCF		TCF		SCF		TCF	
h^{-1}	e_h	$e_h/e_{h/2}$	e_h	$e_h/e_{h/2}$	e_h	$e_h/e_{h/2}$	e_h	$e_h/e_{h/2}$
20	$7.479 \cdot 10^{-2}$	3.36	$1.415 \cdot 10^{-2}$	2.72	$3.879 \cdot 10^{-1}$	1.26	$2.430 \cdot 10^{-2}$	3.69
40	$2.224 \cdot 10^{-2}$	3.81	$5.197 \cdot 10^{-3}$	3.33	$3.070 \cdot 10^{-1}$	1.50	$6.586 \cdot 10^{-3}$	3.87
80	$5.843 \cdot 10^{-3}$	3.94	$1.563 \cdot 10^{-3}$	3.66	$2.046 \cdot 10^{-1}$	1.71	$1.703 \cdot 10^{-3}$	3.93
160	$1.482 \cdot 10^{-3}$	3.98	$4.268 \cdot 10^{-4}$	3.83	$1.200 \cdot 10^{-1}$	1.84	$4.333 \cdot 10^{-4}$	3.97
320	$3.723 \cdot 10^{-4}$	3.99	$1.114 \cdot 10^{-4}$	3.92	$6.532 \cdot 10^{-2}$	1.92	$1.092 \cdot 10^{-4}$	3.98
640	$9.324 \cdot 10^{-5}$	4.00	$2.844 \cdot 10^{-5}$	3.96	$3.411 \cdot 10^{-2}$	1.96	$2.742 \cdot 10^{-5}$	3.99
1280	$2.333 \cdot 10^{-5}$		$7.186 \cdot 10^{-6}$		$1.743 \cdot 10^{-2}$		$6.868 \cdot 10^{-6}$	

TCF scheme, and consequently, both schemes display second order convergence behaviour for $h \to 0$. The numerical errors are approximately the same for both schemes. However, the situation is quite different for the case $\varepsilon = 10^{-8}$. In this case $e_h/e_{h/2} \to 2$ for $h \to 0$ for the SCF scheme, which means that the method is only first order convergent. The TCF-scheme still displays second convergence behaviour. Obviously, the TCF-solution is in this case much more accurate than the SCF-solution.

References

1. Eymard, R., Gallouët, T., Herbin, R.: Finite Volume Methods. In: Ciarlet, P.G., Lions, J.L. (eds.) Handbook of Numerical Analysis, Volume VII, pp. 713–1020. North-Holland, Amsterdam (2000)
2. Mattheij, R.M.M., Rienstra, S.W., Ten Thije Boonkkamp, J.H.M.: Partial Differential Equations, Modeling, Analysis, Computing. SIAM, Philadelphia (2005)
3. Morton, K.W.: Numerical Solution of Convection-Diffusion Problems, Applied Mathematics and Mathematical Computation 12. Chapman & Hall, London (1996)
4. Ten Thije Boonkkamp, J.H.M., Anthonissen, M.J.H.: The Finite Volume-Complete Flux Scheme for One-Dimensional Advection-Diffusion-Reaction Equations. CASA report 08–28, Eindhoven University of Technology
5. Van 't Hof, B., Ten Thije Boonkkamp, J.H.M., Mattheij, R.M.M.: Discretisation of the Stationary Convection-Diffusion-Reaction Equation. Numer. Meth. for Part. Diff. Eq. **14**, 607–625 (1998)
6. Wesseling, P.: Principles of Computational Fluid Dynamics, Springer Series in Computational Mathematics 29. Springer, Berlin (2000)

Solution of Navier–Stokes Equations Using FEM with Stabilizing Subgrid

M. Tezer-Sezgin, S. Han Aydın, and A.I. Neslitürk

Abstract The Galerkin finite element method (FEM) is used for solving the incompressible Navier–Stokes equations in 2D. Regular triangular elements are used to discretize the domain and the finite-dimensional spaces employed consist of piecewise continuous linear interpolants enriched with the residual-free bubble (RFB) functions. To find the bubble part of the solution, a two-level FEM with a stabilizing subgrid of a single node is described in our previous paper [Int. J. Numer. Methods Fluids 58, 551–572 (2007)]. The results for backward facing step flow and flow through 2D channel with an obstruction on the lower wall show that the proper choice of the subgrid node is crucial to get stable and accurate solutions consistent with the physical configuration of the problems at a cheap computational cost.

1 Introduction

Applications of the Galerkin finite element method to incompressible flow equations in velocity-pressure form were carried out in the early 1970s. But, the use of equal-order interpolations for both velocity and pressure produces some spurious oscillations. To overcome this, either Babuška–Brezzi condition [2, 3] must be satisfied, or some stabilized methods such as SUPG (streamline upwind Petrov–Galerkin) should be considered [6]. The finite element methods of the SUPG type reduces

M. Tezer-Sezgin (✉)
Department of Mathematics & Institute of Applied Mathematics, Middle East Technical University 06531 Ankara, Turkey
e-mail: munt@metu.edu.tr

S. Han Aydın
Computer Center, Middle East Technical University 06531 Ankara, Turkey
e-mail: saydin@metu.edu.tr

A.I. Neslitürk
Izmir Institute of Technology, Department of Mathematics, 35430 Izmir, Turkey
e-mail: alinesliturk@iyte.edu.tr

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_94, © Springer-Verlag Berlin Heidelberg 2010

the oscillations in the standard Galerkin method of piecewise linears and achieves stability by adding mesh-dependent perturbation terms to the formulation [9].

Later it has been shown that the SUPG type stabilized methods for the equations modeling the flow problems can be derived by adding the bubble functions to the velocity space in the standard Galerkin finite element formulation, and then eliminating the bubbles by using the static condensation approach [4,5]. In this approach the optimal choice of the stabilization parameter in the SUPG method was simply translated into the problem of the optimal choice of the bubble space. Therefore, the bubble functions should be chosen appropriately such as residual free bubble functions [4] by using two-level FEM [8]. Since the two-level FEM is a computationally expensive procedure, a cheap efficient algorithm which generates qualitatively the same bubble functions is sought.

In this work, the stabilizing subgrid method (SSM) for solving Navier–Stokes equations, which is given in [11] is applied for obtaining solutions of step flows. SSM approximates the solution well and proves good stability features. It is further computationally cheap and able to adapt itself between different flow regimes.

2 The RFB Method Through Two-Level FEM with a Stabilizing Subgrid

The steady incompressible Navier–Stokes equations in an open bounded domain $\Omega \subset^2 I\!R$ with the boundary $\partial\Omega$ are given by

$$\begin{cases} \mathbf{u}\cdot\nabla\mathbf{u} - \frac{1}{Re}\triangle^2\mathbf{u} + \nabla p = \mathbf{f} \ \text{ in } \Omega\,, \\ \nabla\cdot\mathbf{u} = 0 \ \text{ in } \Omega\,, \end{cases} \tag{1}$$

where $\mathbf{u}$ is the velocity field, p is the scalar pressure function, $\mathbf{f}$ is a given source function and Re is the Reynolds number.

We use standard notation for function spaces: $C^0(\bar{\Omega})$ is the space of continuous functions on the closure of Ω, $L^2(\Omega)$ is the space of square integrable functions over the domain Ω, $H^1(\Omega)$ is the Sobolev space of $L^2(\Omega)$ functions whose derivatives are square integrable functions in Ω, and $H_0^1(\Omega)$ is the Sobolev subspace of $H^1(\Omega)$ functions in Ω with zero value on the boundary $\partial\Omega$.

The weak formulation of the problem (1) is obtained by employing the pair of function spaces $V = (H_0^1(\Omega))^2$ and $P = C^0(\bar{\Omega}) \cap L^2(\Omega)$, and it reads: Find $\mathbf{u} \in V$, $p \in P$ such that

$$B(\mathbf{u};\mathbf{u},p;\mathbf{v},q) = (\mathbf{f},\mathbf{v}) \qquad \text{for all} \quad \mathbf{v} \in V,\ q \in P, \tag{2}$$

where

$$B(\mathbf{u};\mathbf{u},p;\mathbf{v},q) = (\mathbf{u}\cdot\nabla\mathbf{u},\mathbf{v}) + \frac{1}{Re}(\nabla\mathbf{u},\nabla\mathbf{v}) - (\nabla\mathbf{v},p) + (\nabla\cdot\mathbf{u},q)$$

and (u,v) is the inner product of u and v.

Let Ω_h be the discretization of Ω by triangles. Define finite dimensional subspaces on Ω_h

$$V_h = \{\mathbf{v} \in (H_0^1(\Omega))^2 \mid \mathbf{v}\mid_K \in P_1(K)^2\,,\ K \in \Omega_h\},$$
$$P_h = \{p \in C^0(\bar{\Omega}) \cap L^2(\Omega) \mid p\mid_K \in P_1(K)\,,\ K \in \Omega_h\}$$

where $P_1(K)$ is the space of piecewise linear functions of a typical element K. The standard Galerkin finite element method is based on employing the same function space for both test and trial spaces and it is equivalent to finding the pair $\{\mathbf{u}_h, p_h\}$ from $V_h \times P_h$ such that

$$B(\mathbf{u}_h; \mathbf{u}_h, p_h; \mathbf{v}_h, q_h) = (\mathbf{f}, \mathbf{v}_h) \quad \forall\{\mathbf{v}_h, q_h\} \in V_h \times P_h\,, \tag{3}$$

where

$$B(\mathbf{u}_h; \mathbf{u}_h, p_h; \mathbf{v}_h, q_h) = (\mathbf{u}_h \cdot \boldsymbol{\nabla}\mathbf{u}_h, \mathbf{v}_h) + \frac{1}{Re}(\boldsymbol{\nabla}\mathbf{u}_h, \boldsymbol{\nabla}\mathbf{v}_h) - (\boldsymbol{\nabla}\mathbf{v}_h, p_h) + (\boldsymbol{\nabla}\cdot\mathbf{u}_h, q_h).$$

The nonlinearity in (3) due to the advection term is resolved with an iteration on the approximate solution $\mathbf{u}_h$ and p_h as

$$\mathbf{u}_h^{n+1} = \mathbf{u}_h^n + \hat{\mathbf{u}}_h \tag{4}$$
$$p_h^{n+1} = p_h^n + \hat{p}_h \tag{5}$$

where n denotes the iteration step and $\hat{\mathbf{u}}_h$ and $\hat{p}_h$ are the corrections to the approximations at the previous iteration step. We linearize the problem (3) by taking

$$\mathbf{u}_h{}^{n+1} \cdot \nabla\mathbf{u}_h{}^{n+1} \approx \mathbf{u}_h{}^n \cdot \nabla\mathbf{u}_h{}^n + \hat{\mathbf{u}}_h \cdot \nabla\mathbf{u}_h{}^n + \mathbf{u}_h{}^n \cdot \nabla\hat{\mathbf{u}}_h$$

for the solution of $\hat{\mathbf{u}}_h$ and $\hat{p}_h$.

SUPG formulation of the Navier–Stokes equations for linear elements is given in [6] as: Find $\{\mathbf{u}_h, p_h\}$ from $V_h \times P_h$

$$\begin{aligned}&(\mathbf{u}_h \cdot \nabla\mathbf{u}_h, \mathbf{v}_h) + \tfrac{1}{Re}(\nabla\mathbf{u}_h, \nabla\mathbf{v}_h) - (\nabla\mathbf{v}_h, p_h) + (\nabla\cdot\mathbf{u}_h, q_h)\\ &+ \sum \tau_K \int_{\Omega K} ((\mathbf{u}_h \cdot \nabla\mathbf{u}_h + \nabla p_h - \mathbf{f}) \cdot (\mathbf{u}_h \cdot \nabla\mathbf{v}_h + \nabla q_h))\, d\Omega_K = (\mathbf{f}, \mathbf{v}_h)\end{aligned} \tag{6}$$

$\forall\{\mathbf{v}_h, q_h\} \in V_h \times P_h$ with the stabilization parameter τ_K such that [6]

$$\tau_K = \frac{h_K}{2|\mathbf{u}_h{}^n|_K}\varepsilon(Pe_K) \tag{7}$$

where h_k is the diameter of the element and ε is a function given as

$$\varepsilon(Pe_K) = \begin{cases} Pe_k & \text{if } Pe_k < 1 \\ 1 & \text{if } Pe_k \geq 1 \end{cases}$$

and

$$Pe_K = \frac{|\mathbf{u}_h{}^n|_K h_K}{6\frac{1}{Re}}.$$

SSM is based on the selection of a single subgrid point whose location has the role in the stabilization of the convection dominated flows [10]. It uses triangular elements. We note that the SSM and the SUPG formulation of the Navier–Stokes equations have the identical structure except for the value of the stabilization parameter τ_K [11]. The value of the stabilization parameter τ_K is given in terms of the bubble function b_K as,

$$\tau_K = \frac{1}{|K|}\int_K b_K \, dK \tag{8}$$

and b_K is the unique bubble function defined by the following boundary value problem in K:

$$\begin{cases} Lb_K = -\frac{1}{Re}\triangle^2 b_K + \mathbf{u}_1^n \cdot \nabla b_K = 1 & \text{in } K \\ b_K = 0 & \text{on } \partial K. \end{cases} \tag{9}$$

Since (9) can be viewed as a linear advection-diffusion equation, finding the exact solution of the problem may not be an easy task in an arbitrary triangular domain. Therefore, a cheap efficient approximation by b_K that generates qualitatively the same behavior with the exact bubble function b_K is required [10].

The subgrid point N is joined to the three vertices denoted by V_i splitting the triangle K into three sub-triangles called K_i. We will choose the point N along one of the three medians of K. We denote the area of i^{th} sub-triangle by $|K_i|$, the edge of K opposite to V_i by e_i and the length of e_i by $|e_i|$. The location of the subgrid point is determined using the procedure given in [11]. Then, the stabilization parameter τ_K can be obtained approximately as

$$\tilde{\tau}_K = \frac{1}{|K|}\int_K b_N^* = \frac{1}{|K|}\frac{(\int_K b_N)^2}{\frac{1}{Re}\int_K |\nabla b_N|^2} = \frac{4|K|}{9\frac{1}{Re}\sum_i |e_i|^2/|K_i|}. \tag{10}$$

The values of $\tilde{\tau}_K$s are then used in the global formulation (6) in place of τ_K.

3 Numerical Results

Navier–Stokes equations are solved in rectangular channels containing steps or obstructions using FEM with stabilizing subgrid method.

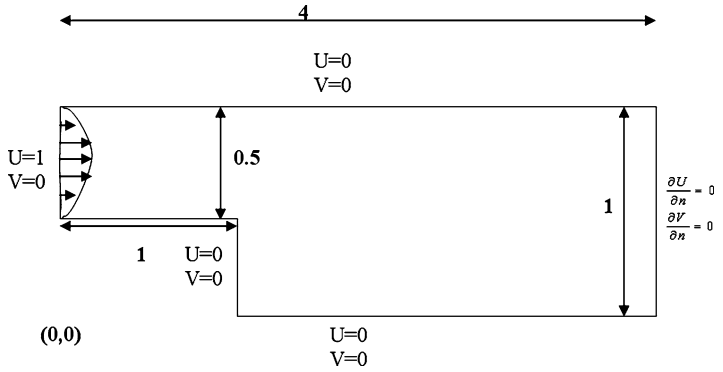


Fig. 1 The statement of the backward facing step flow

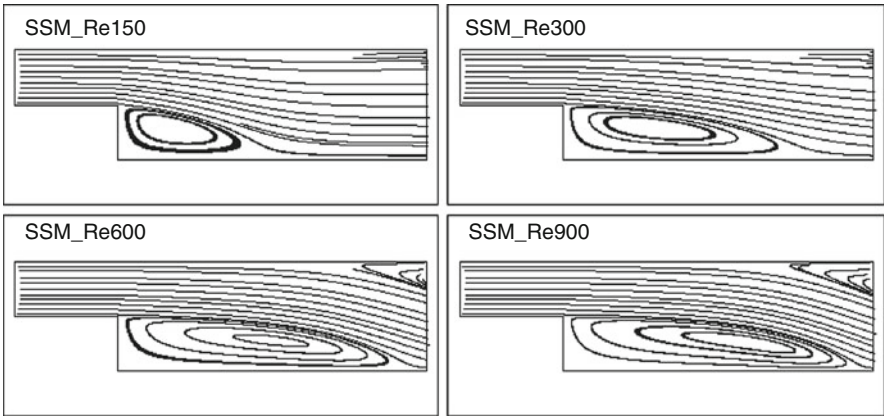


Fig. 2 Changes in the flow as Reynolds number increases for the backward facing step problem

3.1 Backward Facing Step Flow

This is a standard benchmark problem [1]. The problem specifications are given in Fig. 1. It is known that, the results obtained with standard Galerkin FEM show oscillations in pressure values even for small Reynolds numbers [1].

We present the streamlines for $Re = 150, 300, 600$ and 900 in Fig. 2. As Reynolds number increases, the vortex in front of the step enlarges and existence of new vortices are observed. Figure 3 shows pressure contours for the same Reynolds numbers. The stabilizing subgrid method is effective especially for the advection dominated flows on the rough mesh. Figure 4 shows the configuration of subgrid points in the same mesh for $Re = 150$ and $Re = 300$, respectively. As the Reynolds number increases the problem becomes advection dominated and therefore the adaptation of the position of the subgrid point is strongly pronounced. The stabilization is now effective on a larger portion of the entire domain.

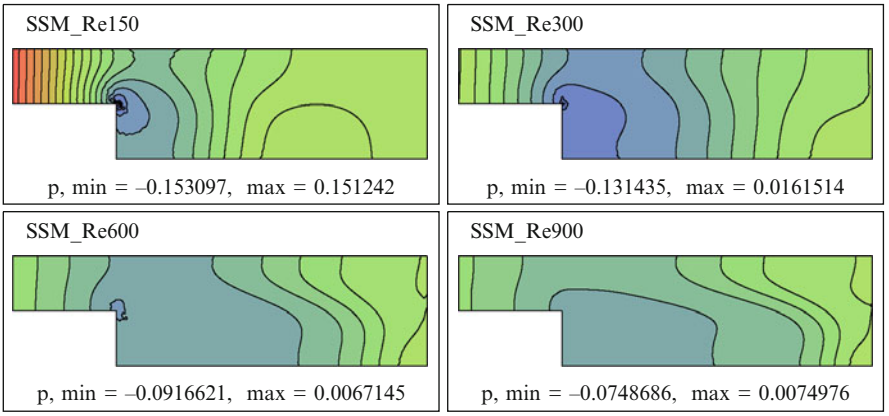


Fig. 3 Pressure contours for the backward facing step for different Reynolds numbers

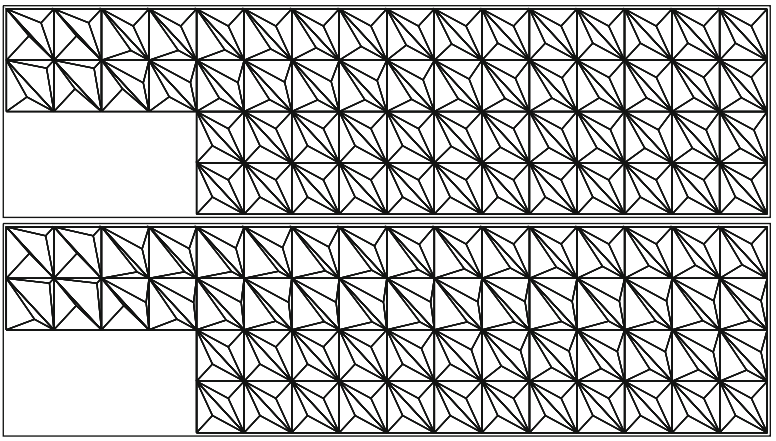


Fig. 4 Adaptation of the subgrid points in SSM as the problem becomes convection dominated for $Re = 150$ and $Re = 300$

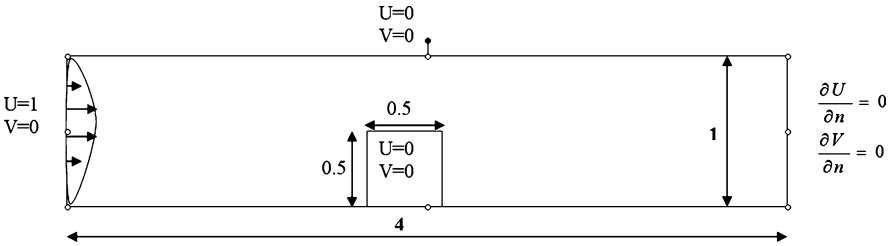


Fig. 5 The statement of the flow in the channel with an obstruction

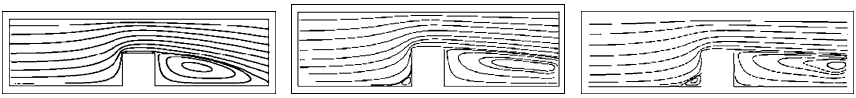


Fig. 6 Streamlines for the flow in the channel with an obstruction for Reynolds numbers $Re = 100, 200, 400$ and $L = 4$

Fig. 7 Pressure contours for the flow in the channel with an obstruction for Reynolds numbers $Re = 100, 200, 400$ and $L = 4$

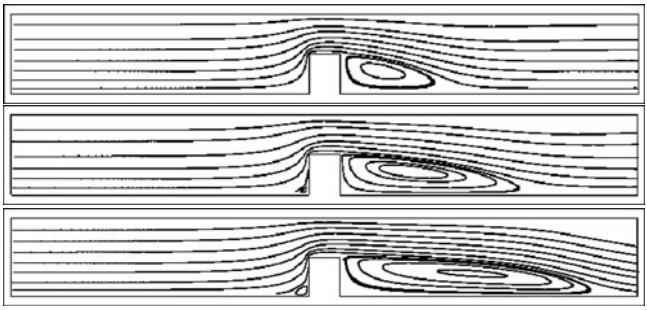


Fig. 8 Streamlines for the flow in the channel with an obstruction for Reynolds numbers $Re = 100, 200, 400$ and $L = 10$

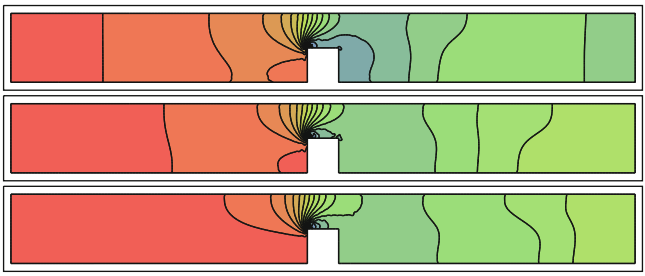


Fig. 9 Pressure contours for the flow in the channel with an obstruction for Reynolds numbers $Re = 100, 200, 400$ and $L = 10$

3.2 Flow Through 2D Channel with an Obstruction on the Lower Wall

This is another test problem [7]. The statement of the problem is given in Fig. 5. The problem is solved for short and long channel cases. In the first case, the channel length is taken as $L = 4$. Changes in the flow in terms of streamlines (in Fig. 6) and

pressure contours (in Fig. 7) are presented for different Reynolds numbers. For high Reynolds number values, existence of a new vortex is captured on the left side of the obstruction.

The complete structure of the vortices and pressure contours are seen more clearly in Figs. 8 and 9 as the channel length is taken longer ($L = 10$).

References

1. Aydın, S.H.: The finite element method over a simple stabilizing grid applied to fluid flow problems. *Ph.D. Thesis.* Middle East Technical University, Ankara, Turkey (2008)
2. Babuska, I.: The finite element method with Lagrangian multipliers. Numer.Math. **20**, 179–192 (1973)
3. Brezzi, F.: On the existence, uniqueness and approximation of saddle-point problems arising from Lagrange multipliers. RAIRO Ser. Rouge. **8**, 129–151 (1974)
4. Brezzi, F., Russo, A.: Choosing bubbles for advection-diffusion problems. M3AS. **4**, 571–587 (1994)
5. Brezzi, F., Bristeau, M.O., Franca, L.P., Mallet, M., Rogé, G.: A relationship between stabilized finite element methods and the Galerkin method with bubble functions. Comput. Methods Appl. Mech. Engrg. **96**, 117–129 (1992)
6. Brooks, A.N., Hughes, T.J.R.: Streamline upwind/Petrov–Galerkin formulations for convection dominated flows with particular emphasis on the incompressible Navier–Stokes equations. Comput. Methods Appl. Mech. Engrg. **32**, 199–259 (1982)
7. Demirkaya, G., Wafo Soh, C., Ilegbusi, O.J.: Direc solution of Navier–Stokes equations by radial basis functions. Appl. Math. Modell. **32**, 1848–1858 (2008)
8. Franca, L.P., Macedo, A.P.: A two-level finite element method and its application to the Helmholtz equation. Int. J. Numer. Methods Eng. **43**, 23–32 (1998)
9. Johnson, C., Nävert, U., Pitkäranta, J.: Finite element methods for linear hyperbolic problem. Comput. Methods Appl. Mech. Engrg. **45**, 285–312 (1984)
10. Nesliturk, A.I.: A Stabilizing subgrid for convection-diffusion problem. M3AS. **16** (2), 211–231 (2006)
11. Nesliturk, A.I., Aydin, S.H., Tezer-Sezgin, M.: Two-level finite element method with a stabilizing subgrid for the incompressible Navier–Stokes equations. Int. J. Numer. Methods Fluids **58**, 551–572 (2007)

Multigrid Methods for Control-Constrained Elliptic Optimal Control Problems

Michelle Vallejos and Alfio Borzì

Abstract Multigrid schemes that solve control-constrained elliptic optimal control problems discretized by finite differences are presented. A gradient projection method is used to treat the constraints on the control variable. A comparison is made between two multigrid methods, the multigrid for optimization (MGOPT) method and the collective smoothing multigrid (CSMG) method. To illustrate both techniques, we focus on minimization problems governed by elliptic differential equations with constraints on the control variable.

1 Introduction

Multigrid methods solve elliptic optimal control problems with optimal computational complexity. Due to recent theoretical and experimental results, multigrid is now considered as one of the most promising approaches for the development of efficient optimization schemes. Some recent developments include the application of multigrid to unconstrained optimization problems [7, 8], to optimal control problems [1, 2, 6] and to inverse problems [10, 11].

The purpose of this paper is to investigate two representative multigrid methods for optimization: the collective smoothing multigrid method (CSMG) and the multigrid for optimization method (MGOPT). We consider the application of these methods for solving control-constrained elliptic optimal control problems. While both schemes are based on the well known full approximation storage (FAS) scheme [4], they represent different approaches to the solution of optimization problems. The CSMG scheme solves optimal control problems by solving the corresponding

M. Vallejos (✉)
Institute of Mathematics, University of the Philippines, Diliman, Quezon City, Philippines
e-mail: michelle.vallejos@uni-graz.at, michelle.vallejos@up.edu.ph

A. Borzì
Dipartimento e Facoltà di Ingegneria, Università degli Studi del Sannio, 82100 Benevento, Italia
e-mail: alfio.borzi@unisannio.it

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_95,

PDE optimality system and treating all optimization variables collectively. As typical in multigrid development, this approach needs to customize the collective smoothing strategy for each individual problem. On the other hand, an appropriate design of the CSMG multigrid components results in a robust algorithm with typical multigrid efficiency [3]. The MGOPT method was first introduced in [7, 8]. The motivation for investigating the MGOPT scheme is that it can be formulated in a way that is not problem specific and therefore it appears to have much larger applicability. In the MGOPT scheme the multigrid solution process represents the outer loop where the control function is considered as the unique dependent variable. The inner loop in this scheme consists of a classical one-grid optimization scheme. In this paper, we discuss an extension of techniques developed in [12] for the case of control-unconstrained elliptic optimal control problems.

In the next sections, constrained optimal control problems are presented together with the discretization scheme and a detailed description of appropriate smoothing algorithms. In Sect. 4, the multigrid scheme is formulated. Numerical experiments follow to demonstrate the ability of multigrid in solving control-constrained optimal control problems. A section of conclusion completes this paper.

2 Constrained Optimal Control Problems

In this section, we discuss constrained optimal control problems. The corresponding optimality system is presented and the multigrid solution process will be detailed in the next section.

We consider a constrained optimal control problem governed by a partial differential equation given by

$$\begin{aligned} &\min_{u\in U_{ad}} J(y,u),\\ &c(y,u)=0, \end{aligned}$$

where $c(y,u)$ is an elliptic partial differential equation (PDE) that represents the equality constraint and the control space is a closed convex subset of $L^2(\Omega)$,

$$U_{ad}=\{u\in L^2(\Omega)|\underline{u}\le u\le \overline{u} \text{ a.e. in } \Omega\}, \tag{1}$$

where $\underline{u}$ and $\overline{u}$ are elements of $L^\infty(\Omega)$. Let c be defined in an open domain $\Omega\subset \Re^d$ together with homogeneous Dirichlet boundary conditions. By assumption, given a control variable $u\in U$, the equality constraint admits a unique solution y, called the state variable, such that the mapping $u\to y$ is affine and continuous.

We focus on a control-constrained linear elliptic optimal control problem

$$\begin{aligned} \min_{u\in U_{ad}} J(y,u) :&= \tfrac{1}{2}\|y-z\|^2_{L^2(\Omega)}+\tfrac{\nu}{2}\|u\|^2_{L^2(\Omega)},\\ -\Delta y-u &= f \ \text{ in } \Omega,\\ y &= 0 \ \text{ on } \partial\Omega, \end{aligned} \tag{2}$$

where $\nu > 0$ is the weight of the cost of the control, $z \in L^2(\Omega)$ is the target function, and $f \in L^2(\Omega)$.

We define the Lagrange functional $L(y, u, p) = J(y, u) + \langle < -\Delta y - u - f, p \rangle_{H_1^{-1}, H_1}$, where p is the Lagrange multiplier. Equating to zero the Fréchet derivatives of L with respect to the triple (y, u, p) results to the first-order necessary optimality conditions for a minimum. The existence of a unique solution to (2) and its characterization are well known. For completeness, a short derivation is as follows:

Let the solution $y(u)$ of the equality constraint, also called the state equation, be a function of u such that $u \to y(u)$ is an affine and continuous mapping from $L^2(\Omega)$ to $H^2(\Omega) \cap H_0^1(\Omega)$. Let us denote its first derivative at u in the direction δu by $y'(u, \delta u)$. It is characterized as the solution to

$$-\Delta y'(u, \delta u) - \delta u = 0 \text{ in } \Omega, \quad y'(u, \delta u) = 0 \text{ on } \partial\Omega. \tag{3}$$

The second derivative of $u \to y(u)$ is zero. We now introduce the reduced cost functional $\hat{J}(u) = J(y(u), u)$, together with its gradient with respect to u given by $\hat{J}'(u) = \nu u - p$. The mapping $u \to \hat{J}(u)$ is twice Fréchet differentiable and its derivatives are given by

$$\begin{aligned} \hat{J}'(u, \delta u) &= (y(u) - z, y'(u, \delta u))_{L^2(\Omega)} + \nu(u, \delta u)_{L^2(\Omega)}, \\ \hat{J}''(u)(\delta u, \delta u) &= \|y'(u, \delta u)\|^2_{L^2(\Omega)} + \nu\|\delta u\|^2_{L^2(\Omega)}. \end{aligned}$$

If $\nu > 0$, $u \to \hat{J}(u)$ is uniformly convex and this implies the existence of a unique solution u^* to (2). The solution u^* is characterized by the following optimality condition:

$$\hat{J}'(u^*, v - u^*) = (y(u^*) - z, y'(u^*, v - u^*))_{L^2(\Omega)} + \nu(u^*, v - u^*)_{L^2(\Omega)} \geq 0,$$

for all $v \in U_{ad}$. Let $p^* = p(u^*) \in H^2(\Omega) \cap H_0^1(\Omega)$ be a solution to

$$-\Delta p^* + y(u^*) = z \text{ in } \Omega, \quad p^* = 0 \text{ on } \partial\Omega, \tag{4}$$

where p^* is the Lagrange multiplier. Then by (3) and (4), we have

$$\begin{aligned} \hat{J}'(u^*, v - u^*) &= (y(u^*) - z, y'(u^*, v - u^*))_{L^2(\Omega)} + \nu(u^*, v - u^*)_{L^2(\Omega)} \\ &= (\Delta p^*, -\Delta^{-1}(v - u^*))_{L^2(\Omega)} + \nu(u^*, v - u^*)_{L^2(\Omega)} \\ &= (\nu u^* - p^*, v - u^*)_{L^2(\Omega)} \geq 0 \text{ for all } v \in U_{ad}, \end{aligned}$$

which constitutes the necessary and sufficient optimality condition for the given optimal control problem (2). Hence we have

$$\begin{aligned} -\Delta y - u &= f \quad \text{in } \Omega, & y &= 0 \quad \text{on } \partial\Omega, \\ -\Delta p + y &= z \quad \text{in } \Omega, & p &= 0 \quad \text{on } \partial\Omega, \\ (\nu u - p, v - u) &\geq 0 \quad \text{for all } v \in U_{ad}. \end{aligned} \tag{5}$$

The first equation is called the state equation, the second is the adjoint equation and the inequality condition is the optimality condition. Equation (5) is called the optimality system which is a characterization of the solution to the given optimization problem (2).

Next we discuss the finite difference discretization scheme together with the smoothing algorithms associated to CSMG and MGOPT methods.

3 Discretization Scheme and Smoothing Algorithms

Our discussion on multigrid methods requires to define a hierarchy of problems $A_k u_k = f_k$ in Ω_k, indexed by $k = 1, 2, \ldots, L$. Here Ω_k denotes the set of grid points with uniform grid spacing h_k for the finite difference discretization in Ω taken as a square domain. For simplicity, we assume that $h_{k-1} = 2h_k$ such that $h_1 > h_2 > \cdots > h_L > 0$. The number of interior grid points is n_k and any function in Ω_k is a vector of size n_k. We denote this vector space with V_k and we introduce the inner product$(\cdot,\cdot)_k$ with the corresponding norm $\|u\|_k = \sqrt{(u,u)_k}$. For multigrid purpose we define a restriction operator $I_k^{k-1} : V_k \to V_{k-1}$ and a prolongation operator $I_{k-1}^k : V_{k-1} \to V_k$ such that $(I_k^{k-1}u, v)_{k-1} = (u, I_{k-1}^k v)_k$ for all $u \in V_k$ and $v \in V_{k-1}$.

Now we consider the discrete version of the optimality system (5). We have

$$\begin{aligned} -\Delta_k y_k - u_k &= f_k, \\ -\Delta_k p_k + y_k &= z_k, \\ (\nu u_k - p_k, v_k - u_k) &\geq 0. \end{aligned}$$

Let $x \in \Omega_k$ where $x = (ih_k, jh_k)$ and i, j are the indices of the grid points arranged lexicographically. We use the standard five point stencil for the Laplacian. We first set $A = -(y_{i-1,j} + y_{i+1,j} + y_{i,j-1} + y_{i,j+1}) - h^2 f_{i,j}$, and $B = -(p_{i-1,j} + p_{i+1,j} + p_{i,j-1} + p_{i,j+1}) - h^2 z_{i,j}$. The values A and B are considered constant during the update of the variables at ij. Hence, we have

$$\begin{aligned} A_{i,j} + 4y_{i,j} - h^2 u_{i,j} &= 0, \\ B_{i,j} + 4p_{i,j} + h^2 y_{i,j} &= 0, \\ (\nu u_{i,j} - p_{i,j}, v_{i,j} - u_{i,j}) &\geq 0. \end{aligned}$$

Let $w_k = (y_k, u_k, p_k)$. A collective smoothing step on w updates the values $y_{i,j}$, $u_{i,j}$, and $p_{i,j}$ such that the resulting residuals of the state and adjoint equations at that point are zero. We can compute the updates for the variables $y_{i,j}$ and $p_{i,j}$ in

the following way

$$\begin{aligned} y_{i,j}(u_{i,j}) &= \tfrac{1}{4}(h^2 u_{i,j} - A_{i,j}), \\ p_{i,j}(u_{i,j}) &= \tfrac{1}{16}(-h^4 u_{i,j} + h^2 A_{i,j} - 4B_{i,j}). \end{aligned} \tag{6}$$

To obtain an update $u_{i,j}$, we replace the expression for $p_{i,j}$ in the inequality constraint and define the auxiliary variable as

$$\widetilde{u}_{i,j} = \frac{1}{16\nu + h^4}(h^2 A - 4B). \tag{7}$$

Then the new value for $u_{i,j}$ resulting from the smoothing step is given by

$$u_{i,j} = \begin{cases} \overline{u}_{i,j} & \text{if } \widetilde{u}_{i,j} \geq \overline{u}_{i,j} \\ \widetilde{u}_{i,j} & \text{if } \underline{u}_{i,j} < \widetilde{u}_{i,j} < \overline{u}_{i,j} \\ \underline{u}_{i,j} & \text{if } \widetilde{u}_{i,j} \leq \underline{u}_{i,j} \end{cases}. \tag{8}$$

With this new value of $u_{i,j}$, new values for $y_{i,j}$ and $p_{i,j}$ are obtained. This completes the description of the collective smoothing step for the constrained case.

The collective smoothing step defined by (6)–(8) satisfies the inequality constraint in the optimality system (5). Consider any grid point wherein $\widetilde{u} \leq \underline{u}$, then from (8) $u = \underline{u}$. Thus, $(v - u) \geq 0$ for any $v \in U_{ad}$. On the other hand, we have

$$\nu u - p = \nu u - \frac{1}{16}(-h^4 u_{i,j} + h^2 A - 4B) \geq \frac{1}{16}\left[\left(16\nu + h^4\right)\widetilde{u} - \left(h^2 A - 4B\right)\right] = 0.$$

Therefore, $(\nu u - p, v - u) \geq 0$ for all $v \in U_{ad}$. Similarly, one proves that if $\widetilde{u} \geq \overline{u}$, then the choice $u = \overline{u}$ satisfies the inequality constraint. The case $\underline{u} < \widetilde{u} < \overline{u}$ is obvious.

For the MGOPT case, the gradient projection method is utilized as the smoothing algorithm. In this setting, we want to find a solution u of $\min_u \left(\hat{J}(u) - (f, u)\right)$ such that $u \in U_{ad}$. Define the projection P onto U_{ad} by

$$P_{U_{ad}}(u) = \begin{cases} \overline{u} & \text{if } u \geq \overline{u} \\ u & \text{if } \underline{u} < u < \overline{u}. \\ \underline{u} & \text{if } u \leq \underline{u} \end{cases}$$

Given the current iterate u^ℓ, define the new iterate by $u^\ell(\alpha) = P_{U_{ad}}(u^\ell + \alpha d^\ell)$, where d^ℓ is the search direction given by $d^\ell = -\hat{J}'(u^\ell) - f$ and α satisfies the sufficient decrease condition for bound constrained problems [5, 9].

A report on the applicability and efficiency of these smoothing algorithms are shown in Sect. 5.

4 The Multigrid Method

In this section we present the two multigrid schemes for solving control-constrained elliptic optimal control problems, the CSMG and the MGOPT methods.

The CSMG scheme is based on the nonlinear multigrid full approximation storage (FAS) scheme applied to the optimality system with a collective smoothing. Some recent applications of the CSMG method to optimal control problems with control constraints are presented in [1, 6].

The multigrid for optimization (MGOPT) method was first introduced by Lewis and Nash [7, 8] as an extension of the multigrid scheme to optimization problems. This method is very similar to the CSMG scheme and some recent developments include the application of MGOPT to optimal control problems [12].

To illustrate both methods we consider a discrete problem

$$A_k w_k = f_k,$$

where A_k represents a discrete linear operator on Ω_k. The MGOPT method is applied to solve $\min_{u_k} \left(\hat{J}_k(u_k) - (f_k, u_k)_k\right)$. Hence in this case, $w := u$ and $A_k u_k = \hat{J}_k'(u_k)$. In the CSMG case, we solve (5) and define $w := (y, u, p)$.

Let the smoothing iteration at level k be given by S_k such that we get an update $w_k^\ell = S_k(w_k^{\ell-1}, f_k)$, $\ell = 1, 2, \ldots, \gamma_1$. Starting with an initial approximation w_k^0, we apply γ_1 times the smoothing scheme S_k and obtain $w_k^{\gamma_1}$. On a coarse grid V_{k-1}, the problem is given by

$$A_{k-1} w_{k-1} = f_{k-1},$$

where $f_{k-1} = I_k^{k-1} f_k + \tau_{k-1}$ and $\tau_{k-1} = A_{k-1}(I_k^{k-1} w_k^{\gamma_1}) - I_k^{k-1} A_k w_k^{\gamma_1}$ is called the fine-to-coarse residual/gradient correction. Once this problem is solved, the coarse grid correction step follows

$$w_k^{\gamma_1+1} = w_k^{\gamma_1} + \alpha I_{k-1}^k (w_{k-1} - I_k^{k-1} w_k^{\gamma_1}).$$

For CSMG $\alpha = 1$ and for MGOPT α is the step length obtained after a line search procedure in the direction $I_{k-1}^k (w_{k-1} - I_k^{k-1} w_k^{\gamma_1})$. Finally, we apply γ_2 iterations of the smoothing algorithm to damp possible high frequency errors that may arise from the coarse grid correction process. The following algorithm presents the method described above.

Algorithm (Multigrid method)
Choose w_k^0 to be an initial approximation at resolution k. If $k = 1$, solve $A_k w_k = f_k$ and return. Else if $k > 1$,

1. Apply γ_1 iterations of a smoothing algorithm. $w_k^\ell = S_k(w_k^{\ell-1}, f_k)$, $\ell = 1, 2, \ldots, \gamma_1$
2. Compute the right hand side $f_{k-1} = I_k^{k-1} f_k + \tau_{k-1}$, where $\tau_{k-1} = A_{k-1} w_{k-1}^{\gamma_1} - I_k^{k-1} A_k w_k^{\gamma_1}$.
3. Apply γ cycles of MG (γ_1, γ_2) to the coarse grid problem $A_{k-1} w_{k-1} = f_{k-1}$.

4. For a given step length α, $w_k^{\gamma_1+1} = w_k^{\gamma_1} + \alpha I_{k-1}^k (w_{k-1} - w_{k-1}^{\gamma_1})$.
5. Apply γ_2 iterations of a smoothing algorithm. $w_k^\ell = S_k(w_k^{\ell-1}, f_k)$, $\ell = \gamma_1 + 2, \ldots, \gamma_1 + \gamma_2 + 1$.

5 Numerical Results

In this section, we present the results of the experiments on the computational performance of the proposed multigrid schemes as solvers for control-constrained elliptic optimal control problems. Using different values of the cost of the control, we gathered the number of iterations and the CPU time (in seconds) until a stopping tolerance of $\|u^\ell - u^\ell(1)\|_{L^2} < 10^{-6}$ is satisfied for both the CSMG and the MGOPT methods. For all computations, we use $\gamma_1 = \gamma_2 = 2$ pre and post smoothing steps. This means that one iteration of the CSMG and the MGOPT method uses $\gamma_1 + \gamma_2 = 4$ iterations of the smoothing algorithm on the finest level. We consider problem (2) with the zero function as an initial guess, $\Omega = (0, 1) \times (0, 1)$ and f, $z \in L^2(\Omega)$ are

$$f(x_1, x_2) = 0, \quad z(x_1, x_2) = \sin(\pi\, x_1) \sin(2\pi\, x_2).$$

The numerical results are shown in Table 1. We can see that using different parameters ν for CSMG, the method converges within nine iterations and the number of iterations is independent on the mesh size. On the other hand, the MGOPT method converges within two iterations and the results show that the number of iterations is independent both on the weighting parameter ν and the mesh size. Moreover, the CPU time (in seconds) approximately increase by a factor of four by halving the mesh size. This shows an almost optimal computational complexity of the MGOPT approach. The numerical solutions y and u for $\nu = 10^{-4}$ are shown in Fig. 1.

Table 1 Numerical results using CSMG and MGOPT schemes for different weighting parameters

ν	Mesh	Iter	CSMG	Iter	MGOPT
10^{-2}	129×129	6	26.0	2	10.5
	257×257	6	396.1	2	47.6
	513×513	6	3360.3	2	221.7
10^{-3}	129×129	7	29.9	2	12.4
	257×257	7	461.9	2	50.2
	513×513	7	3970.5	2	236.0
10^{-4}	129×129	9	38.9	2	12.8
	257×257	9	610.9	2	56.0
	513×513	9	5104.8	2	238.7

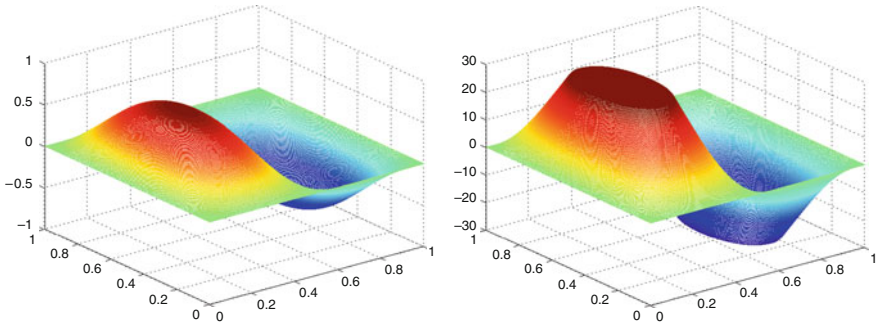


Fig. 1 Numerical solutions for the state (*left*) and control (*right*) variables of the control-constrained problem using $\nu = 10^{-4}$

6 Conclusion

In this paper the solution of control-constrained elliptic optimal control problems discretized by finite differences and solved by means of CSMG and MGOPT methods is presented. The numerical results show that the MGOPT method exhibits a faster convergence rate compared to the CSMG method. Since any optimization algorithm can be used as a smoothing iteration for MGOPT, it is easier to implement than the CSMG method where the smoothing iteration is problem specific. Because of the modularity of MGOPT, it can be easily applied to a large class of PDE-based optimization problems. A topic which can be considered for future research is the appropriate use of different optimization algorithms as smoothing iteration for MGOPT.

Acknowledgements Supported by the Office of the Chancellor, in collaboration with the Office of the Vice-Chancellor for Research and Development, of the University of the Philippines Diliman through the Ph.D. Incentive Award.

References

1. Borzì, A. and Kunisch, K.: A multigrid scheme for elliptic constrained optimal control problems. Comput. Optim. Appl. **31**(3), 309–333 (2005)
2. Borzì, A. and Schulz, V.: Multigrid methods for PDE optimization. SIAM Review **51**(2), 361–395 (2009)
3. Borzì, A. and Kunisch, K. and Kwak, D. Y.: Accuracy and convergence properties of the finite difference multigrid solution of an optimal control optimality system. SIAM J. Control Optim. **41**(5), 1477–1497 (2002)
4. Brandt, A.: Multi-level adaptive solutions to boundary-value problems. Math. Comp. **31**(138), 333–390 (1977)
5. Kelley, C.T.: Iterative Methods for Optimization. Kluwer, New York (1987)

6. Lass, O. and Vallejos, M. and Borzì, A. and Douglas, C.C.: Implementation and analysis of multigrid schemes with finite elements for elliptic optimal control problems. J. Computing **84**(1-2), 27–48 (2009)
7. Lewis, R.M. and Nash, S.: Model problems for the multigrid optimization of systems governed by differential equations. SIAM J. Sci. Comput. **26**(6), 1811–1837 (2005)
8. Nash, S.: A multigrid approach to discretized optimization problems. Optim. Methods Softw. **14**(1-2), 99–116 (2000)
9. Nocedal, J. and Wright, S.J.: Numerical optimization. Kluwer, New York (1999)
10. Oh, S. and Milstein, A. and Bouman, C. and Webb, K.J.: A general framework for nonlinear multigrid inversion. IEEE Trans. Image Process. **14**(1), 125–140 (2005)
11. Oh, S. and Bouman, C. and Webb, K.J.: Multigrid tomographic inversion with variable resolution data and image spaces. IEEE Trans. Image Process. **15**(9), 2805–2819 (2006)
12. Vallejos, M. and Borzì, A.: Multigrid optimization methods for linear and bilinear elliptic optimal control problems. J. Computing **82**(1), 31–52 (2008)

Modelling the New Soil Improvement Method Biogrout: Extension to 3D

W.K. van Wijngaarden, F.J. Vermolen, G.A.M. van Meurs, and C. Vuik

Abstract Biogrout is a new soil improvement method based on microbial induced carbonate precipitation. Bacteria and reactants are flushed through the soil, resulting in calcium carbonate precipitation and consequent soil reinforcement. A mathematical model was created to describe the process. The model contains the concentrations of the dissolved species that are present in the precipitation reaction. These concentrations can be solved from a convection-dispersion-reaction equation with a variable porosity. Other model equations involve the concentrations of the bacteria and of the solid calcium carbonate, the decreasing porosity (due to precipitation) and the flow. The partial differential equations are solved by the Standard Galerkin Finite Element Method. The subject of this paper is the extension of the mathematical model to 3D.

1 Introduction

Biogrout is a new soil reinforcement method based on microbial induced carbonate precipitation [7]. Bacteria are placed and subsequently reactants (urea ($CO(NH_2)_2$) and calcium chloride ($CaCl_2$)) are flushed through the soil, resulting in calcium carbonate ($CaCO_3$) precipitation, causing an increase in strength and stiffness of the soil.

Biogrout can be applied to a wide variety of situations, in which it is desirable to change the properties of the subsoil [2]. We briefly mention the following examples

W.K. van Wijngaarden (✉) and 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 and 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

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_96,

- reinforcement of the soil underneath railway-tracks;
- soil stabilization prior to tunnelling;
- reinforcement of dunes to decrease effects of wave erosion, and hence to protect delicate coastlines;
- prevention of liquefaction of the subsoil resulting from earthquakes.

The Biogrout process consists of two parts: the microbial induced production of carbonate (CO_3^{2-}) due to the hydrolysis of urea (with ammonium (NH_4^+) as a side-product) and the precipitation of calcium carbonate. In [7], the corresponding reaction equations are given. Combining these reactions gives the overall Biogrout reaction equation:

$$CO(NH_2)_2(aq) + Ca^{2+}(aq) + 2H_2O(l) \rightarrow 2NH_4^+(aq) + CaCO_3(s). \qquad (1)$$

The solid calcium carbonate strengthens the subsoil by connecting the sand grains. As a result of the precipitation of calcium carbonate, the porosity and the permeability of the soil decrease. This phenomenon influences the flow.

In [5] a model has been derived to describe the Biogrout process. Thus far, only simulations for 1D and 2D configurations have been done. In this paper, a simulation will be carried out for a 3D configuration.

This paper contains the following sections. Section 2 summarizes the model for the Biogrout process that was derived in [5]. Section 3 is devoted to the numerical methods, used to solve the model equations. Section 4 contains some computer simulations for a 3D configuration and in Sect. 5 conclusions and discussions can be found.

2 The Mathematical Model

In this section, the (differential) equations that are needed to describe the Biogrout process are given, together with a short explanation. In [5] the derivation can be found. These (differential) equations were derived in respect with the following assumptions:

- Only dissolved species do react;
- The biochemical reaction of the Biogrout process is the only reaction that takes place and this reaction is governed by reaction (1);
- The concentration of the bacteria is constant in time and homogeneous;
- Calcium carbonate is not transported but it precipitates on the matrix of the porous medium;
- The precipitation of calcium carbonate has no influence on the total volume of the fluid over the entire domain of computation;
- The flow is incompressible;
- The viscosity is constant.

The biochemical reaction of the Biogrout process is given by (1). We will start by giving the differential equations for the aqueous species in this equation.

The differential equation for the concentration of urea is given by:

$$\theta \frac{\partial C^{urea}}{\partial t} = \nabla \cdot (\theta \mathsf{D} \cdot \nabla C^{urea}) - \mathbf{q} \cdot \nabla C^{urea} - \theta r. \tag{2}$$

In this equation, θ is the porosity, C^{urea} is the dissolved concentration of urea, D is the dispersion tensor, $\mathbf{v}$ is the pore water velocity and r is the reaction rate of the production of calcium carbonate, which is a non-linear function of the urea concentration and the time.

The term at the left-hand side represents the accumulation. The first term at the right-hand side represents the effect of dispersion and diffusion, the second term models advection and the last term stands for the biochemical reaction. The minus-sign comes from the fact that urea is consumed at the same rate as calcium carbonate is formed, see (1).

In three dimensions, the coefficients of the dispersion tensor D equal $D_{ij} = (\alpha_L - \alpha_T)\frac{v_i v_j}{|\mathbf{v}|} + \delta_{ij}\alpha_T \sum_i \frac{v_i^2}{|\mathbf{v}|}$, see [8]. The quantity α_L is the longitudinal dispersivity and α_T is the transverse dispersivity.

Analogously, we have the following differential equation for the concentrations of calcium and ammonium:

$$\theta \frac{\partial C^{\mathrm{Ca}^{2+}}}{\partial t} = \nabla \cdot (\theta \mathsf{D} \cdot \nabla C^{\mathrm{Ca}^{2+}}) - \mathbf{q} \cdot \nabla C^{\mathrm{Ca}^{2+}} - \theta r, \tag{3}$$

$$\theta \frac{\partial C^{\mathrm{NH}_4^+}}{\partial t} = \nabla \cdot (\theta \mathsf{D} \cdot \nabla C^{\mathrm{NH}_4^+}) - \mathbf{q} \cdot \nabla C^{\mathrm{NH}_4^+} + 2\theta r. \tag{4}$$

Note the $+2$ in the biochemical reaction term in the differential equation for ammonium: for each produced mole of calcium carbonate, two moles of ammonium are generated.

For the non-aqueous species in reaction (1), calcium carbonate, we have the following differential equation:

$$\frac{\partial C^{CaCO_3}}{\partial t} = m_{CaCO_3} \theta r. \tag{5}$$

In this equation, m_{CaCO_3} is the molar mass of calcium carbonate and is used to convert number of molecules (moles) into mass (kilograms). The right-hand side of this differential equation only contains the reaction term since it has been assumed that calcium carbonate is not transported.

We have the following relation between the concentration of calcium carbonate and the porosity:

$$\theta(t) = \theta(0) - \frac{C^{CaCO_3}(t) - C^{CaCO_3}(0)}{\rho_{CaCO_3}}, \tag{6}$$

where ρ_{CaCO_3} is the density of calcium carbonate.

The flow is calculated from Darcy's Law, given in [8]:

$$q_x = -\frac{k_x}{\mu}\frac{\partial p}{\partial x},\ q_y = -\frac{k_y}{\mu}\frac{\partial p}{\partial y},\ q_z = -\frac{k_z}{\mu}\left(\frac{\partial p}{\partial z} + \rho g\right). \tag{7}$$

In Darcy's Law, p is the pressure, k_i is the intrinsic permeability in the various coordinate directions ($i = x, y, z$), μ is the viscosity that is assumed to be constant in the Biogrout case, ρ is the density of the solution and g is the gravitational constant.

The intrinsic permeability k is determined, using the Kozeny–Carman relation: an empirical relation between the intrinsic permeability and the porosity that is commonly used in ground water flow modelling (see [1]):

$$k = \frac{(d_m)^2}{180}\frac{\theta^3}{(1-\theta)^2}. \tag{8}$$

In this relation, d_m is the mean particle size of the subsurface medium. If the porosity is small, it might be that the pores are not connected. Hence, the permeability is zero. This phenomenon is not directly incorporated in the Kozeny–Carman relation. Since in our simulations the porosity is not that small, we assume that the Kozeny–Carman relation is a good relation between the permeability and the porosity.

The density of the solution (at 20 °C), ρ, will be calculated with the following experimental relation:

$$\rho = 1000 + 15.4996C^{urea} + 86.7338C^{Ca^{2+}} + 15.8991C^{NH_4^+}. \tag{9}$$

For the pressure, the following differential equation was derived in [5] by the use of Darcy's Law (7):

$$-\nabla\cdot\left(\frac{k}{\mu}\left(\nabla p + \rho g\mathbf{e_z}\right)\right) = \frac{m_{CaCO_3}}{\rho_{CaCO_3}}\theta r. \tag{10}$$

Differential equation (2), (3), (4), (5) and (10) contain the reaction rate r of the biochemical reaction (1). This rate decreases in time as is shown in experiments, see [6]. In [5] a linear reduction had been assumed, combined with Monod kinetics, [3]. In this paper, we will combine Monod kinetics with an exponential reduction, since this is commonly used as a first approximation (see [4]):

$$r = v_{max}\frac{C^{urea}}{K_m + C^{urea}}e^{-bt}. \tag{11}$$

In this equation, v_{max} is the initial activity, K_m is the saturation constant and b is some constant, representing the reduction in bacterial activity in the course of time.

As initial conditions, the concentration of calcium carbonate, urea, calcium and ammonium are equal to zero and the porosity equals θ_0.

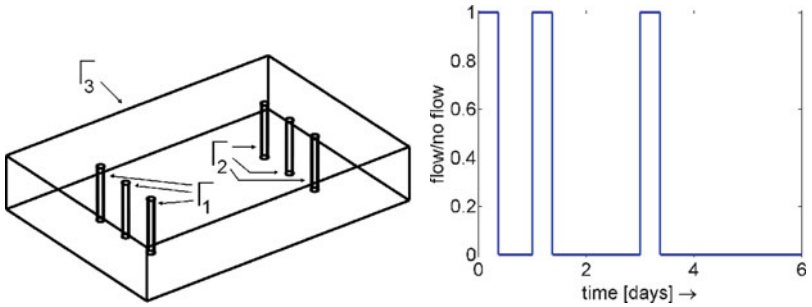


Fig. 1 Experimental set-up. *Left*: Configuration, *Right*: Flow strategy

Table 1 Boundary conditions for the pressure and the concentration of urea, calcium and ammonium

	p	C^{urea}	$C^{Ca^{2+}}$	$C^{NH_4^+}$
Γ_1(during injection)	$-\frac{k}{\mu}(\nabla p + \rho g \mathbf{e_z}) \cdot \mathbf{n} = q_{in}$	$C^{urea} = c_{in}$	$C^{Ca^{2+}} = c_{in}$	$C^{NH_4^+} = 0$
Γ_1(during rest)	$p = p_2 + \int_z^{1.5} \rho g \bar{z} d\bar{z}$	$\frac{\partial C^{urea}}{\partial n} = 0$	$\frac{\partial C^{Ca^{2+}}}{\partial n} = 0$	$\frac{\partial C^{NH_4^+}}{\partial n} = 0$
Γ_2	$p = p_2 + \int_z^{1.5} \rho g \bar{z} d\bar{z}$	$\frac{\partial C^{urea}}{\partial n} = 0$	$\frac{\partial C^{Ca^{2+}}}{\partial n} = 0$	$\frac{\partial C^{NH_4^+}}{\partial n} = 0$
Γ_3	$-\frac{k}{\mu}(\nabla p + \rho g \mathbf{e_z}) \cdot \mathbf{n} = 0$	$\frac{\partial C^{urea}}{\partial n} = 0$	$\frac{\partial C^{Ca^{2+}}}{\partial n} = 0$	$\frac{\partial C^{NH_4^+}}{\partial n} = 0$

As a model experiment, a container (8 m×5.6 m×1.5 m) has been taken, having closed boundaries (represented by boundary Γ_3). In this container injection and extraction wells have been placed (Fig. 1). The injection wells are represented by boundary Γ_1, whereas the extraction wells are represented by boundary Γ_2. The following flow strategy has been chosen: there are three batches, starting with nine hours of injection and no injection during the rest of the batch. The duration of the batches is respectively 1, 2 and 3 days, see Fig. 1.

Table 1 displays the boundary conditions that are chosen.

Since we have the same differential equation, initial condition and boundary conditions for both the concentration of urea and calcium chloride, these concentrations are equal. Hence it is sufficient to calculate only the urea concentration.

3 Numerical Method

The differential equations for the pressure, the velocity and the concentration of the aqueous species are solved by the Standard Galerkin Finite Element Method. The weak formulation is derived by multiplication by a test function $\eta \in H^1(\Omega)$ and integration over the domain Ω. For the time integration, an IMEX-scheme is used: all components are solved implicitly, except for the porosity θ, the intrinsic permeability k and the density of the solution ρ. Solving the differential equation for the pressure, the reaction rate r is also computed explicitly. While solving

the differential equation for the urea concentration, Newton's method is used, because of the non-linearity in the reaction term (11). The Newton-Cotes quadrature rules have been used for the approximation of the element matrices and vectors. Tetrahedral elements have been used, in combination with linear basis functions.

Since the differential equation for the concentration of calcium carbonate, (5), is an ordinary differential equation (in each grid point), it is not necessary to use the Finite Element Method. For the time integration, an IMEX-scheme is used: solving all components implicitly, except for the porosity.

At each time step, the differential equations for the following components are solved successively: the pressure, the flow and the concentration of urea, calcium, ammonium and calcium carbonate. For more details, see [5].

Finally, the porosity (θ), the intrinsic permeability (k) and the density of the fluid (ρ) are recalculated with (6), (8) and (9), respectively. Also the boundary conditions are updated.

Since the porosity, the permeability and the density of the solution (may) vary, at each time step all the matrices are rebuilt. That means, calculate for each element a 4×4 element matrix and add them to the large matrix. This is done for 10 different matrices + the number of Newton-iterations, since in each Newton-iteration a new matrix is built.

4 Results

In this section, the results of the simulation with the model for a 3D configuration are shown. Matlab has been used to do the numerical simulations. The linear systems are solved by a direct method. The time step $\Delta t = 1\,\mathrm{h}$, $q_{in} = 2.29 \times 10^{-4}\,\mathrm{m\,s^{-1}}$, $v_{max} = 1.621 \times 10^{-2}\,\mathrm{mol\,m^{-3}\,s^{-1}}$ and $b = 7.15 \times 10^{-6}\,\mathrm{s^{-1}}$. The values of the other constants can be found in [5].

All the three batches start with nine hours of injection with inflow velocity q_{in}. During injection, the amount of urea in the domain increases, although this phenomenon is diminished by the hydrolysis of urea. During rest, the total amount of urea decreases, due to the hydrolysis of urea. The reaction rate (11) decreases in time. As a consequence, the total amount of urea decreases slower during the period of rest as time proceeds.

The urea/calcium chloride solution is heavier than water and is also heavier than the solution of the reaction product ammonium chloride as can be seen from formula (9). As a result, in the lower parts of the domain a higher urea and calcium chloride concentration are expected. This will result in a higher calcium carbonate concentration in the lower parts of the domain. Figure 2 confirms these expectations and also gives some quantitative details.

At each time step, new matrices are built, since porosity, permeability and density of the solution (may) vary. In this paper, the relation between the CPU time for the building part and for the solving part has been investigated. Seven different meshes have been taken, with increasing number of elements. With each mesh 10 time steps

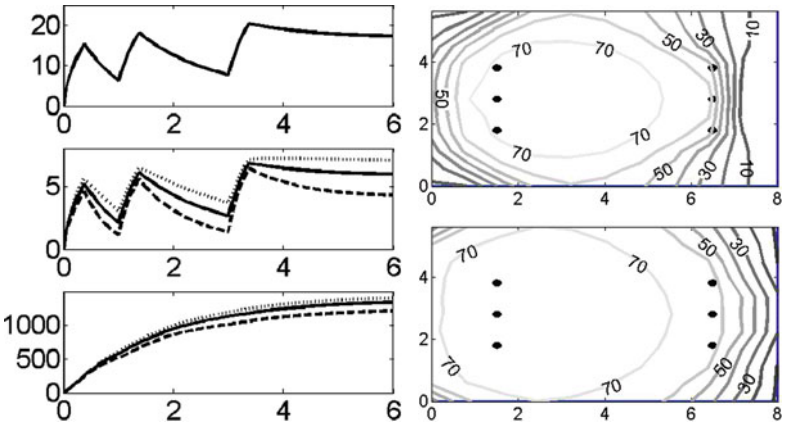


Fig. 2 Some results of the 3D model experiment. *Top left*: total amount of urea (kmol) in time (days) in the whole domain; *Middle left*: amount of urea (kmol) in time (days) in several parts of the domain: —— upper part, — middle part, ··· lower part; *Bottom left*: the amount of calcium carbonate (kg) in the same parts of the domain: —— upper part, — middle part, ··· lower part. *Top right*: a contour plot of the calcium carbonate concentration (kg/m^3) after the three batches at z = 1.5 m (top domain), x[m] and y[m] on the x-axis and y-axis; *Bottom right*: a contour plot of the calcium carbonate concentration (kg/m^3) after the three batches at z = 0 m (bottom domain), x[m] and y[m] on the x-axis and y-axis

Table 2 CPU time per time step, subdivided in the building part and the solving part for seven different meshes, with increasing number of elements and the relative error

Number of Elements (approximately)	CPU time			Percentage Solving part	Relative Error
	Per time step (s)	Building part (s)	Solving part (s)		
2500	0.344	0.242	0.102	30%	24%
5000	0.715	0.459	0.255	36%	15%
10000	1.58	0.921	0.661	42%	10%
20000	4.28	1.88	2.39	56%	6.3%
40000	13.9	3.80	10.1	73%	3.5%
80000	46.8	8.23	38.6	82%	1.1%
160000	182	17.0	165	91%	(0%)

have been taken, registering the average CPU time per time step and the average CPU time per time step for the building part and the solving part. The results can be found in Table 2. This table also contains the percentage solving time/total time.

From this table, it can be seen that, if the number of elements increases with a factor 2, so does the CPU time for the building part. This is what is expected: for each element a 4×4 element matrix is created and is added to the large matrix. If the number of elements doubles, the amount of work doubles, too.

If the number of elements doubles, the amount of solving work increases with a factor 2.5, 2.6, 3.6, 4.2, 3.8 and 4.3, respectively. So the amount of work increases with more than a factor 2, what can also be expected from the analysis of a band

matrix solver. For a mesh with 2500 elements only 30% of the CPU time is spent in the solving part. For a mesh with 160000 elements this is even 91%. If the number of elements increases further, it will be necessary to use an iterative method instead of a direct method.

The discretization error is $O(\Delta x^2 + \Delta t)$. If the number of elements is increased with a factor 2, Δx^2 is decreased with a factor $2^{2/3}$. If the time step is also decreased with a factor $2^{2/3}$, then, in the limit, the error should decrease with a factor $2^{2/3} (\approx 1.6)$. The last column of Table 2 contains the relative error in the concentration after six hours in an arbitrary point in the domain with respect to the finest mesh. For the coarsest mesh, a time step of $\Delta t = 0.5$ h has been taken and this time step has been decreased while doubling the number of elements. The relative error decreases with a factor 1.6, 1.5, 1.6, 1.8 and 3.2, respectively. So in the limit, the error decreases with even more than a factor 1.6.

5 Conclusions and Discussion

An extension to 3D of the Biogrout model has been made. The results of the numerical simulation with the 3D configuration with three injection lances and three extraction lances look promising. Also the error analysis gives a good result.

For a small number of elements, building matrices takes more CPU time than solving the matrix vector systems. For a large number of elements it is the other way around. In building matrices, the amount of work increases linearly with the number of elements. If the number of elements increases further, it will be necessary to use an iterative method instead of a direct method.

References

1. Bear, J.: Dynamics of fluids in porous media, Dover Publications, New York (1972)
2. DeJong, J.T., Mortensen, B.M., Martinez, B.C., Nelson, D.C.: Bio-mediated soil improvement. Ecol. Eng. (2009) doi:10.1016/j.ecoleng.2008.12.029
3. Monod, J.: The growth of bacterial cultures, Annu. Rev. Microbiol. **3**, 371–394 (1949)
4. Pruitt, K.M., Kamau, D.N.: Mathematical models of bacterial growth, inhibition and death under combined stress conditions, J. Ind. Microbiol. **12**, 221–231 (1993)
5. 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. Report at DIAM 09-09, Delft University of Technology, Faculty of Electrical and Engineering, Mathematics and Computer Science, Delft Institute of Applied Mathematics, the Netherlands (2009)
6. Whiffin, V.S.: Microbial $CaCO_3$ Precipitation for the production of Biocement. Ph.D thesis. Murdoch University, Perth, Australia (2004)
7. Whiffin, V.S., van Paassen, L.A., Harkes, M.P.: Microbial carbonate precipitation as a soil improvement technique. Geomicrobiol. J., **24:5**, 417–423 (2007)
8. Zheng, C., Bennett, G D. : Applied contaminant transport modeling, Van Nostrand Reinhold, New York (1995)

Angle Conditions for Discrete Maximum Principles in Higher-Order FEM

Tomáš Vejchodský

Abstract This contribution reviews the general theory of the discrete Green's function and presents a numerical experiment indicating that the discrete maximum principle (DMP) fails to hold in the case of Poisson problem on any uniform triangulation of a triangular domain for orders of approximation three and higher. This extends the result [Computing 27, 145–154 (1981)] that the Laplace equation discretized by the higher-order FEM satisfies the DMP on a patch of triangular elements in exceptional cases only.

1 Introduction

The discrete maximum principle (DMP) is important in practice, because it guarantees nonnegativity of approximations of naturally nonnegative quantities like temperature, concentration, density, etc. Its theoretical significance lies in its connection with the uniform convergence of the finite element approximations [5]. In contrast to the lowest-order finite element method (FEM), the DMP for the higher-order FEM in dimension two and higher is not well understood, yet.

A stronger version of the DMP for the Laplace equation discretized by higher-order finite elements was studied by Höhn and Mittelman in [8]. This stronger version requires the validity of the DMP on all vertex patches (union of elements sharing a vertex) in the triangulation. They find that the quadratic elements do not satisfy the stronger DMP unless the triangulation is very special (e.g., all equilateral triangles) and that the restrictions for cubic elements are even more severe.

In the present contribution we briefly review the general theory about the discrete Green's function (DGF) and the standard DMP for the Poisson problem. Then we present a numerical experiment indicating that the standard DMP is not satisfied on any uniform triangulation for the finite elements of order three and higher.

T. Vejchodský
Institute of Mathematics, Academy of Sciences, Žitná 25, CZ–115 67 Prague 1, Czech Republic
e-mail: vejchod@math.cas.cz

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_97,

2 Model Problem and Its FEM Discretization

First, we briefly introduce the Poisson problem and its discretization by the FEM. The main purpose of this section is to settle down the notation.

Let $\Omega \subset \mathbb{R}^d$ be a Lipschitz domain. The classical and the weak formulations of the Poisson problem reads as follows:

$$\text{Find } u \in C^2(\Omega) \cap C(\overline{\Omega}) \text{ such that} \quad -\Delta u = f \quad \text{in } \Omega, \text{ and} \quad u = 0 \text{ on } \partial\Omega. \tag{1}$$

$$\text{Find } u \in H^1_0(\Omega) \text{ such that} \quad a(u,v) = \mathscr{F}(v) \quad \forall v \in H^1_0(\Omega), \tag{2}$$

where $a(u,v) = \int_\Omega \nabla u \cdot \nabla v \,\mathrm{d}x$ and $\mathscr{F}(v) = \int_\Omega f v \,\mathrm{d}x$. We require $f \in C(\Omega)$ for the classical formulation and $f \in L^2(\Omega)$ for the weak one.

In order to discretize problem (2) by the Galerkin method, we introduce a finite dimensional subspace V_h of $H^1_0(\Omega)$. We assume that $V_h \subset C(\overline{\Omega})$. The Galerkin solution $u_h \in V_h$ is given by the requirement

$$a(u_h, v_h) = \mathscr{F}(v_h) \quad \forall v_h \in V_h. \tag{3}$$

Considering a basis $\varphi_1, \varphi_2, \ldots, \varphi_N$ of V_h, we can express $u_h = \sum_{i=1}^N z_i \varphi_i$ and verify that problem (3) is equivalent to the system $Az = F$ of linear algebraic equations, where the stiffness matrix $A \in \mathbb{R}^{N\times N}$ has entries $a_{ij} = a(\varphi_j, \varphi_i)$, the load vector $F \in \mathbb{R}^N$ has entries $F_i = \mathscr{F}(\varphi_i)$, and $z = (z_1, z_2, \ldots, z_N)^\top$.

The FEM can be seen as a special case of the Galerkin method, where the space V_h is chosen in a special way such that the stiffness matrix A is sparse. The particular choice of V_h is not important at this point and it will be specified later on.

3 Discrete Maximum Principle

Theorem 1 below is an equivalent formulation of the standard maximum principle due to E. Hopf [9] applied to problem (1). Similarly, Theorem 2 presents the same principle for the weak solution.

Theorem 1. *Let u be a classical solution to (1). If $f \geq 0$ in Ω then $u \geq 0$ in Ω.*

Theorem 2. *Let u be a weak solution to (2). If $f \geq 0$ a.e. in Ω then $u \geq 0$ a.e. in Ω.*

The same result for the the Galerkin solution $u_h \in V_h$ is known as the DMP. Unfortunately, it is not valid in general and various conditions for its validity are studied.

Definition 1. Let the finite dimensional space V_h be fixed. We say that discretization (3) satisfies the discrete maximum principle (DMP) if the solution $u_h \in V_h$ is nonnegative in Ω for any $f \in L^2(\Omega)$, $f \geq 0$ a.e. in Ω.

A usefull tool for investigation of the DMP especially for the higher-order FEM is the so-called discrete Green's function (DGF) which was already introduced in [3, 6]. For any $y \in \Omega$ let us define the DGF $G_{h,y} \in V_h$ as the unique function satisfying

$$a(v_h, G_{h,y}) = v_h(y) \quad \forall v_h \in V_h. \tag{4}$$

This definition together with (3) implies the representation formula

$$u_h(y) = \mathscr{F}(G_{h,y}) = \int_\Omega f(x) G_h(x, y)\,\mathrm{d}x \quad \forall y \in \Omega,$$

where we use the usual notation $G_h(x, y) = G_{h,y}(x)$. This representation formula immediately proves the following theorem.

Theorem 3. *The discretization (3) satisfies the DMP if and only if* $G_h(x, y) \geq 0$ *for all* $(x, y) \in \Omega^2$.

Interestingly, the DGF G_h can be expressed in terms of a basis of V_h [12]:

$$G_h(x, y) = \sum_{i=1}^{N} \sum_{j=1}^{N} (A^{-1})_{ij} \varphi_i(x) \varphi_j(y) \quad \forall (x, y) \in \Omega^2, \tag{5}$$

where $(A^{-1})_{ij}$ stand for entries of the inverse of the stiffness matrix A. Let us remark that a special case of this formula, where the basis is formed by the eigenvectors of the discrete Laplacian was already presented in [3]. Further, we remark that the concept of the DGF is relevant even for more general problems [13, 14]. However, in the case of nonhomogeneous Dirichlet boundary conditions the boundary Green's function has to be introduced [4]. General formula (5) is used below to analyze the nonnegativity of the DGF and consequently the validity of the DMP.

4 Nonnegativity of the DGF for the Lowest-Order FEM

The analysis of nonnegativity of expression (5) simplifies if the basis functions $\varphi_1, \varphi_2, \ldots, \varphi_N$ of V_h have the following property

$$\sum_{i=1}^{N} z_i \varphi_i \geq 0 \quad \text{in } \Omega \quad \Leftrightarrow \quad z_i \geq 0 \quad \forall i = 1, 2, \ldots, N. \tag{6}$$

This property is typically satisfied for the lowest-order finite elements such as linear functions on simplices and multilinear functions on blocks (Cartesian products of intervals). Before we state the following well-known theorem, we recall that a square matrix A is monotone if it is nonsingular and $A^{-1} \geq 0$ (i.e., all entries of A^{-1} are nonnegative).

Theorem 4. *Let the basis functions $\varphi_1, \varphi_2, \ldots, \varphi_N$ of V_h have property (6). Then the discretization (3) satisfies the DMP if and only if the stiffness matrix A is monotone.*

Proof. It follows immediately from assumption (6), formula (5), and Theorem 3.

If the off-diagonal entries of the stiffness matrix A are nonpositive then A is M-matrix and, hence, monotone. The nonpositivity of the off-diagonal entries can be guaranteed by various geometric conditions on finite element meshes like the nonobtuseness condition for simplicial meshes [1] or the nonnarrowness condition for rectangular finite elements [2]. However, these conditions could be too restrictive, because it suffices to have the stiffness matrix monotone and not M-matrix. An experiment indicating how much the nonobtuseness condition for triangles can be weaken is described in Sect. 6 and its results are presented in Fig. 2 (top-left).

5 Nonnegativity of the DGF for the Higher-Order FEM

Let us investigate the case of the higher-order FEM in more details. For simplicity let us consider two dimensional Poisson problem (1) in a polygonal domain Ω. We define the finite element space as $V_h = \{v \in H_0^1(\Omega) : v|_K \in \mathbb{P}^p(K) \forall K \in \mathscr{T}_h\}$, where $\mathscr{T}_h$ is a face-to-face triangulation of Ω and $\mathbb{P}^p(K)$ stands for the space of polynomials of degree at most p on the triangle K.

The standard basis of V_h consists of N^{V} vertex (piecewise linear) functions $\varphi_1, \varphi_2, \ldots, \varphi_{N^{\mathrm{V}}}$ and of $N - N^{\mathrm{V}}$ higher-order basis functions $\varphi_{N^{\mathrm{V}}+1}$, $\varphi_{N^{\mathrm{V}}+2}, \ldots, \varphi_N$, see e.g., [11]. The vertex functions are the usual piecewise linear "hat" functions. Thus, if B_j, $j = 1, 2, \ldots, N^{\mathrm{V}}$, denote the interior vertices of the triangulation $\mathscr{T}_h$ then the vertex functions satisfy $\varphi_i(B_j) = \delta_{ij}$, $i, j = 1, 2, \ldots, N^{\mathrm{V}}$.

The vertex and the higher-order (non-vertex) basis functions yield a natural 2×2 block structure of the stiffness matrix and its inverse

$$A = \begin{pmatrix} A^{\mathrm{VV}} & A^{\mathrm{VN}} \\ A^{\mathrm{NV}} & A^{\mathrm{NN}} \end{pmatrix}, \qquad A^{-1} = \begin{pmatrix} S^{-1} & -(A^{\mathrm{VV}})^{-1} A^{\mathrm{VN}} R^{-1} \\ -(A^{\mathrm{NN}})^{-1} A^{\mathrm{NV}} S^{-1} & R^{-1} \end{pmatrix},$$

where $A^{\mathrm{VV}} \in \mathbb{R}^{N^{\mathrm{V}} \times N^{\mathrm{V}}}$, $A^{\mathrm{NN}} \in \mathbb{R}^{(N-N^{\mathrm{V}}) \times (N-N^{\mathrm{V}})}$, etc., $S = A^{\mathrm{VV}} - A^{\mathrm{VN}}(A^{\mathrm{NN}})^{-1} A^{\mathrm{NV}}$, and $R = A^{\mathrm{NN}} - A^{\mathrm{NV}}(A^{\mathrm{VV}})^{-1} A^{\mathrm{VN}}$.

The Schur complement S has the following interesting property. Let B_i and B_j, $i, j = 1, 2, \ldots, N^{\mathrm{V}}$, be two interior vertices of the triangulation $\mathscr{T}_h$. Since $\varphi_i(B_j) = \delta_{ij}$ and due to (5) we obtain

$$G_h(B_i, B_j) = (A^{-1})_{ij} \varphi_i(B_i) \varphi_j(B_j) = (A^{-1})_{ij} = (S^{-1})_{ij}. \tag{7}$$

Hence, the values of the DGF at the vertices of $\mathscr{T}_h$ coincide with the entries of S^{-1}. Furthermore, the DGF has a natural structure given by the Cartesian product of the

mesh $\mathscr{T}_h$ with itself. In particular, if K and L are two elements from $\mathscr{T}_h$ and ι_K and ι_L denote the sets of indices of basis functions supported in K and L, respectively, i.e., $\iota_K = \{i : \mathrm{meas}(K \cap \mathrm{supp}\,\varphi_i) > 0\}$, then the DGF restricted to $K \times L$ is given by

$$G_h|_{K\times L}(x, y) = \sum_{i\in\iota_K} \sum_{j\in\iota_L} (A^{-1})_{ij} \varphi_i|_K(x) \varphi_j|_L(y), \quad (x, y) \in K \times L. \tag{8}$$

This formula contains a small number of basis functions and we use it for fast evaluation of the DGF at a given point.

6 Numerical Experiment

In this experiment we test nonnegativity of the DGF on uniform meshes. We consider Poisson problem (1) on a triangle Ω. The finite element mesh is constructed by three successive uniform (red) refinements of Ω, see Fig. 1 (left).

To speed up the test of the nonnegativity of the DGF, we first check the values at vertices, using the Schur complement S, see (7). If S is monotone, it remains to verify the nonnegativity at the other points. We proceed by inspection of all pairs of elements $K, L \in \mathscr{T}_h$ using formula (8). Function $G_h|_{K\times L}$ is a polynomial. The test of nonnegativity of a multivariate polynomial is a complicated task (connected with the 17th Hilbert's problem [10]). Therefore, we sample the values of $G_h|_{K\times L}$ in a number of points $(x^K_{k\ell}, x^L_{mn}) \in K \times L$, where the sample point $x^K_{k\ell}$ has barycentric coordinates $(k, \ell, M - k - \ell)/M$, $0 \le k + \ell \le M$, see Fig. 1 (right). The total number of sample points in an element is $(M + 1)(M + 2)/2$. To ensure that the number of sample points is sufficient, we always perform a series of computations starting with $M = 8$ and doubling M until the results do not change.

Figure 2 presents the results. Each point in a panel corresponds to a pair of angles α and β, which represent the vertex angles of the triangle Ω. The color of this point is given by the properties of the DGF. If the DGF is nonnegative at all vertices and

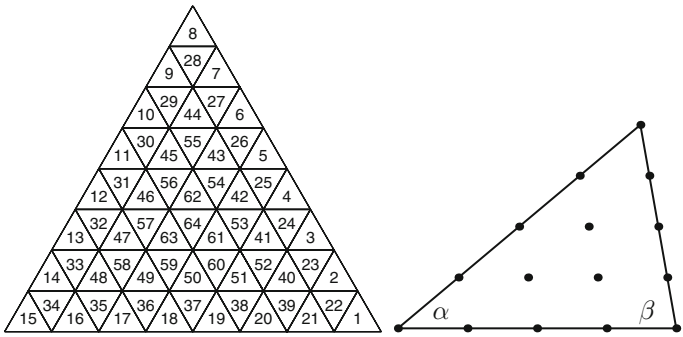


Fig. 1 A uniform mesh with 64 triangles enumerated in a spiral way (*left*). A triangular element characterized by a pair of of angles α and β with sample points for $M = 4$ (*right*)

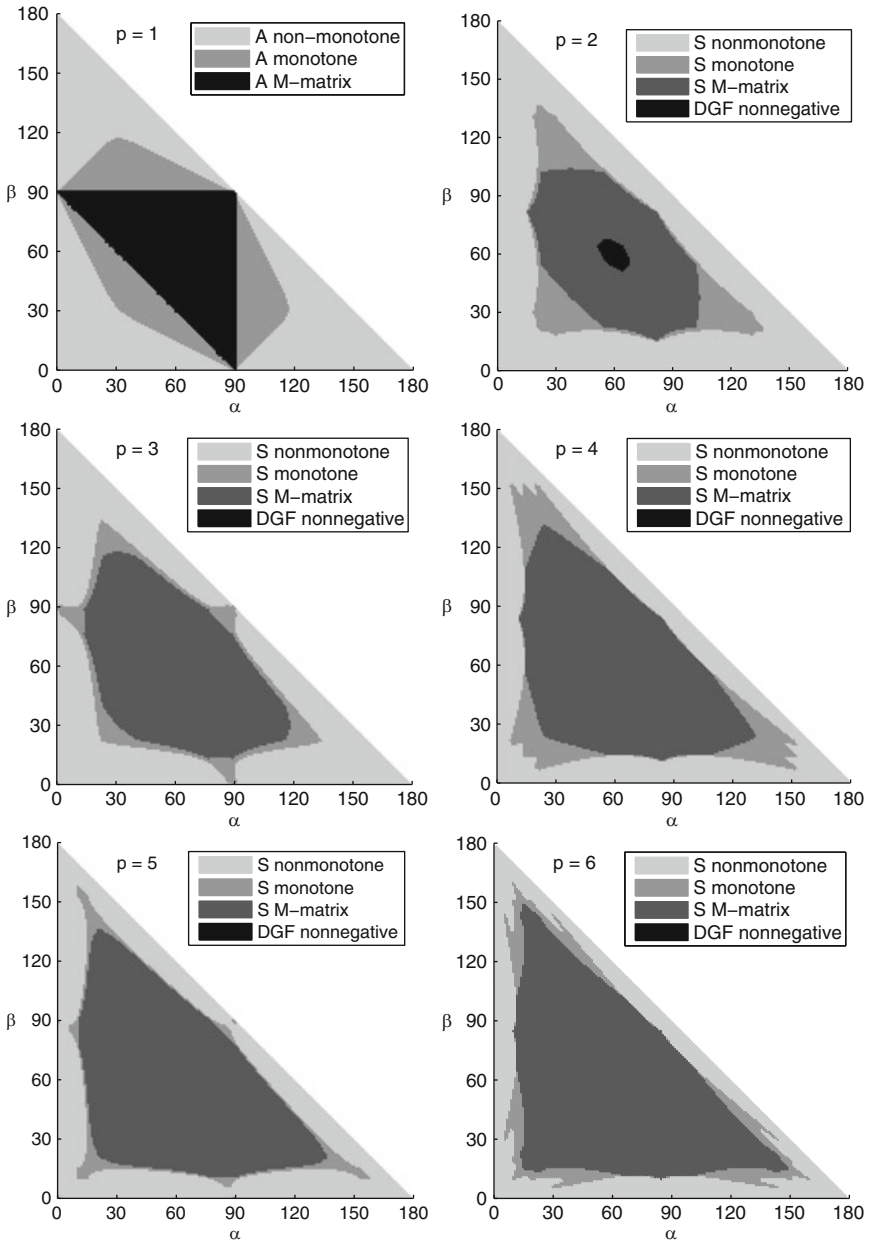


Fig. 2 The nonnegativity of the DGF and its dependence on the angles in the triangulation for orders $p = 1, 2, \ldots, 6$

at all sample points then the color is black. This is the only case when the DMP is hopefully satisfied. If the DGF is not nonnegative then we distinguish three more cases. (i) The DGF is negative in a sample point and S is M-matrix (dark gray). (ii) The DGF is negative in a sample point and S is monotone but not M-matrix (lighter gray). (iii) The DGF is negative in a vertex, i.e., S is nonmonotone (lightest gray).

The above description, however, applies for higher-order elements only ($p \geq 2$). The case of linear elements ($p = 1$) is exceptional, because just the vertex values of the DGF are relevant for its nonnegativity. Due to Theorem 4, we distinguish in the top-left panel of Fig. 2 the cases (a) A is nonmonotone, (b) A is monotone but not M-matrix, (c) A is M-matrix. Notice that the DMP is satisfied in cases (b) and (c).

Clear conclusion from Fig. 2 is that the DGF has negative values for all tested pairs of angles for orders $p \geq 3$. However, if we look on vertex values of the DGF only, we observe that the area of this region increases with p. The increase is not monotone but in principle the higher polynomial degree p we use the wider range of angles can be used in order to keep the vertex values of the DGF nonnegative.

The only polynomial degrees allowing the DMP on uniform meshes are $p = 1$ and $p = 2$. For the case $p = 1$ (see Sect. 4 above) the black area in the top-left panel of Fig. 2 clearly shows that the stiffness matrix A is M-matrix provided the maximal angle is at most 90°. In addition, we observe that the stiffness matrix can be monotone even if the maximal angle is about 117°. In the case $p = 2$ the DMP is satisfied only if all the angles are close to 60°. We also check the nonnegativity of the DGF for meshes finer than the mesh sketched in Fig. 1 (left). The results on meshes one and two times refined are exactly the same as those presented in Fig. 2.

It might be of further interest to see how the DGF really looks like. For illustration we choose $p = 3$ and $\alpha = \beta = 60°$. For these values the DGF is nonnegative in the vertices and negative somewhere in between. The graph of the function $G_h(x, y)$, $(x, y) \in \Omega^2$, is difficult to visualize, because it is a five dimensional object. However, each pair of elements $K_i \in \mathscr{T}_h$ and $K_j \in \mathscr{T}_h$ corresponds to a point in a plane and the color of this point can be chosen according to some characteristic of the DGF restricted to the polytope $K_i \times K_j$. The left panel of Fig. 3 presents the mean values of G_h over $K_i \times K_j$. The right panel illustrates the negative part of the minimum of G_h in $K_i \times K_j$, i.e., $(\min_{K_i \times K_j} G_h)^-$, where $\chi^- = (|\chi| - \chi)/2$. Both these quantities are approximated using the sample points as described above. The used triangulation together with indices of elements is shown in Fig. 1 (left). Notice that the elements with indices 1–39 are adjacent to the boundary of Ω while the elements 40–64 are interior. The right panel of Fig. 3 clearly shows that the DGF is negative in polytopes $K_i \times K_j$, where K_i and K_j are both adjacent to the boundary and they are neighbors to each other including the case $K_i = K_j$. Another choice of angles α and β leads, however, to the negativity of the DGF for more pairs K_i, K_j.

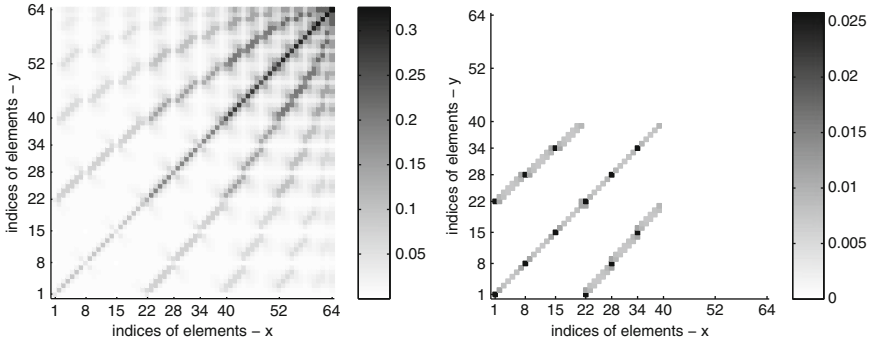


Fig. 3 A visualization of the entire DGF. A point with coordinates i, j corresponds to a pair of elements K_i, K_j. The color of this point represents the mean value (*left*) and the negative part of the minimum (*right*) of G_h in $K_i \times K_j$

7 Conclusions

We discussed the nonnegativity of the DGF and equivalently the validity of the DMP for Galerkin solutions of Poisson problem (1) with homogeneous Dirichlet boundary conditions. Results of the performed experiment indicate that the DGF is not nonnegative on uniform meshes for all shapes of triangular elements for the order three and higher. The quadratic elements yield nonnegative DGF for triangles close to equilateral ones.

The results also indicate that the DGF is negative in the areas close to the boundary. In accordance with [7] we could speculate that the nonnegativity of the DGF is not primarily determined by the angles in the triangulation but by the way how the boundary is resolved. In addition, the domain, where the DGF is negative, is relatively small with respect to the entire Ω^2 and it lies close to the boundary. This means that a nonnegative f corrupting the DMP (Definition 1) must have great values in an element close to the boundary and small values in the interior of Ω (like an approximation of the Dirac delta function). Such data are rare in practice, however. This leads us to another generalization of the (continuous) maximum principle from Theorem 2. If $f \geq 0$ is given, we may ask how must the mesh look like in order to obtain the nonnegative finite element solution. Up to the author's knowledge, this question was not considered in the literature, yet.

A possible remedy of the failure of the DMP for higher-order elements could be a modification of the higher-order basis functions based on the exact eigenfunctions of the Laplacian. This approach was successfully applied in [6] for 1D elliptic problems. A generalization to higher dimension is still an unsolved problem.

Acknowledgements The author acknowledges the support of the Czech Science Foundation, Grant no. 102/07/0496, and of the Czech Academy of Sciences, Grant no. IAA100760702, and Institutional Research Plan no. AV0Z10190503.

References

1. Brandts, J., Korotov, S., Křížek, M.: Dissection of the path-simplex in $\mathbf{R}^n$ into n path-subsimplices. Linear Algebra Appl. **421**, 382–393 (2007)
2. Christie, I., Hall, C.: The maximum principle for bilinear elements. Internat. J. Numer. Methods Engrg. **20**, 549–553 (1984)
3. Ciarlet, P.G.: Discrete variational Green's function. I. Aequationes Math. **4**, 74–82 (1970)
4. Ciarlet, P.G.: Discrete maximum principle for finite-difference operators. Aequationes Math. **4**, 338–352 (1970)
5. Ciarlet, P.G., Raviart, P.A.: Maximum principle and uniform convergence for the finite element method. Comput. Methods Appl. Mech. Engrg. **2**, 17–31 (1973)
6. Ciarlet, P.G., Varga, R.S.: Discrete variational Green's function. II. One dimensional problem. Numer. Math. **16**, 115–128 (1970)
7. Drăgănescu, A., Dupont, T.F., Scott, L.R.: Failure of the discrete maximum principle for an elliptic finite element problem. Math. Comp. **74**, 1–23 (2005)
8. Höhn, W., Mittelmann, H.-D.: Some remarks on the discrete maximum-principle for finite elements of higher order. Computing **27**, 145–154 (1981)
9. Hopf, E.: Elementäre Bemerkungen über die Lösungen partieller Differentialgleichungen zweiter Ordnung vom elliptischen Typus. Sitzungsberichte Preussische Akademie der Wissenschaften, Berlin, 147–152 (1927)
10. Prestel, A., Delzell, C. N.: Positive polynomials: From Hilbert's 17th problem to real algebra. Springer, Berlin (2001)
11. Šolín, P., Segeth, K., Doležel, I.: Higher-order finite element methods. Chapman & Hall/CRC, Boca Raton, FL (2004)
12. Vejchodský, T., Šolín, P.: Discrete maximum principle for higher-order finite elements in 1D. Math. Comp. **76**, 1833–1846 (2007)
13. Vejchodský, T., Šolín, P.: Discrete maximum principle for a 1D problem with piecewise-constant coefficients solved by hp-FEM. J. Numer. Math. **15**, 233–243 (2007)
14. Vejchodský, T., Šolín, P.: Discrete maximum principle for Poisson equation with mixed boundary conditions solved by hp-FEM. Adv. Appl. Math. Mech. **1**, 201–214 (2009)

Unsteady High Order Residual Distribution Schemes with Applications to Linearised Euler Equations

N. Villedieu, L. Koloszar, T. Quintino, and H. Deconinck

Abstract This article is dedicated to the design of high order residual distributive schemes for unsteady problems. We use a space-time strategy, which means that the time is considered as a third dimension. To achieve high order both in space and in time, we use prismatic elements having $(k+1)$ levels, each level being a P^k element. The first section is dedicated to the deign of space-time schemes on such elements. The second section presents the performances on different type of problems. In particular, we look at a discontinuous problem on Euler equations and two problems of propagation of sound using Linearised Euler equations.

1 Generalities and Notations

We describe a class of compact methods to approximate the unsteady solution of

$$\frac{\partial u}{\partial t} + \nabla \cdot F(u) = S \quad \forall (x, y, t) \in \Omega_t = \Omega \times [0; t_f] \tag{1}$$

To solve the unsteady system (1) of m equations we consider that time is a third dimension. So, the domain Ω_t is descretised by a succession of prismatic elements. We first build a triangulation τ_h of the spatial domain Ω, with averaged mesh spacing h. Each of these triangles are P^k elements, and in each element, we construct the conformal sub-triangulation composed of k^2 triangles. We denote by T_s the generic sub-element. Until then we have constructed a high order discretisation of space. To get high order discretisation in time we use prisms with $k+1$ levels, each level being a P^k element of τ_h. As an example, on Fig. 1 are plotted a P^1 and a P^2 prismatic element. For any given function u, its restriction on the prism K is defined by:

$$u^h = \sum_{l=n-k+1}^{n+1} H^l(t) \sum_{i \in T} \psi_i(x, y)\, u_i^l \ , \tag{2}$$

N. Villedieu (✉), L. Koloszar, T. Quintino, and H. Deconinck
Von Karman Institute, Chausse de Waterloo, 72; 1640 Rhode-St-Gense; Belgium
e-mail: villedie@vki.ac.be, koloszar@vki.ac.be

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_98,

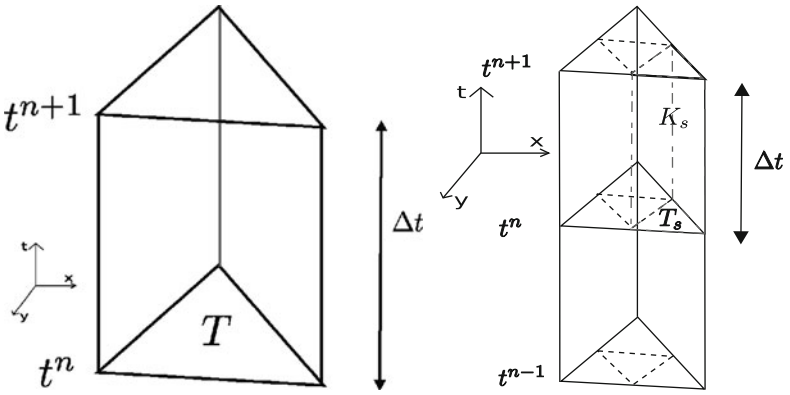


Fig. 1 Space-time element: P^1 element (*left*) and P^2 element (*right*) with subdivision, definition of Δt

where u_i^l is the value of u^h at node i and time $t_l : u_i^l = u^h(x_i, y_i, t_l)$ and $\psi_i(x, y)$ denotes the (mesh dependent) continuous $k-th$ order Lagrangian basis function. H^l is the $1D$ $k-th$ order basis function of level l. In each space-time element the $(u^{n-l})_{k-1\leq l\leq 0}$ are considered as known and u^{n+1} is the unknown which we compute using the process of a steady problem:

1. We compute the vectorial residual on each space-time sub-prism between n and $n+1$:
$$\Phi^{K_s} = \int_{t_n}^{t_{n+1}} \int_{T_s} \left(\frac{\partial u}{\partial t} + \nabla \cdot F - S \right) d\Omega \, dt \tag{3}$$
2. We distribute the residual to the nodes of the sub-prism K_s. To respect the physical meaning of time, we would like to distribute only to the nodes of the level $n+1$. The consistency of the scheme is ensured by a constaint on the time step called past-shield condition (for more details we refer to [6]). Under this condition it is possible to distribute the residual Φ^{K_s} only to the nodes of the level $n+1$:
$$\begin{cases} \Phi_i^n & = 0 \\ \Phi_i^{n+1} & = \beta_i \Phi^{K_s}, \quad \sum_{i \in T_s} \beta_i = I \end{cases}$$
where I is the identity matrix of size $m \times m$
3. We obtain the following nodal equation that is solved by pseudo-time iterations:
$$\sum_{K_s, i \in K_s} \Phi_i^{n+1, K_s} = 0 \tag{4}$$

2 Linear Schemes

In this paper we make use of the following two upwind linear schemes

ST-LDA scheme This scheme is multidimensional upwind[1] and linearity preserving[2] and its residual is defined by:

$$\Phi_i^{K_s} = \Phi_i^{\mathrm{LDA}} = \beta_i^{\mathrm{LDA}} \Phi^{K_s}, \quad \beta_i^{\mathrm{LDA}} = \tilde{k}_i^{n+1,+} (\sum_{j \in T_s} \tilde{k}_j^{n+1,+})^{-1} \tag{7}$$

ST-N scheme is the multidimensional upwind scheme defined by

$$\phi_i^{\text{ST-N}} = \phi_i^{ST-LDA} + d_i^{ST-N} \tag{8}$$

$$d_i^{\text{ST-N}} = \sum_{j \in T_s} \tilde{k}_i^{+} \widetilde{N} \tilde{k}_j^{+} (u_i^{n+1} - u_j^{n+1}) \tag{9}$$

$$\widetilde{N} = (\sum_{j \in T_s} \tilde{k}_j^{+})^{-1} \tag{10}$$

The N scheme is monotone[3] but only first order because it is not linearity preserving.

[1] A scheme is multidimensional upwind if $\beta_i = 0$ when $\tilde{k}_i^{n+1,+} = 0$ where $\tilde{k}_i^{n+1}$ is the upwind matrix of the level $n+1$:

$$\tilde{k}_i^{n+1} = (\frac{1}{2} \frac{\partial F(u^{*,n+1})}{\partial u} \cdot \mathbf{n}_i) \Delta t_1 + \frac{|T_s|}{3} I \tag{5}$$

with $u^{*,n+1}$ an arbitrary average of $u^h(t_{n+1})$ over T_s, $\Delta t_1 = t^{n+1} - t^n$, $|T_s|$ the area of the sub-triangle T_s, I the indentity matrix and $\mathbf{n}_i$ is the inward normal to the face of T_s opposite to node i, the normal is scaled on the length of this face. And its positive part $\tilde{k}_i^{n+1,+}$ is defined by $\tilde{k}_j^{n+1,+} = R \Lambda^+ R^{-1}$, R and Λ^+ being respectively the matrix of the right eigenvectors and of the positive eigenvalues of $\tilde{k}_i^{n+1}$.

[2] Linearity preserving and Accuracy: In the steady case the condition to get $k+1$th order schemes is that (see [2] for details)

$$\Phi_j^{K_s} = \mathscr{O}(h^{k+2}) \tag{6}$$

For the kth degree polynomial approximation (2) we get $\Phi^{K_s} = \mathscr{O}(h^{k+2})$, hence the accuracy condition is also expressed by $\Phi_j^{K_s} = \mathscr{O}(\Phi^{K_s})$ meaning that the distribution coefficient should be bounded (Linearity preserving condition).

[3] Monotonicity: The rigorous definition of monotonicity for RD schemes resorts to the theory of positive coefficients, see [2,8] for details. In this paper we will define a scheme as being monotone if, in practical computations, it gives a non-oscillatory approximations of discontinuities. In particular, we are interested in schemes for which, across a discontinuity, $\Phi_j^{K_s} \times \Phi_j^{\mathrm{M}} \geq 0$, for some first order monotone splitting Φ_j^{M}.

3 Nonlinear Schemes

To combine high order of accuracy and monotonicity, we must use a non-linear splitting. There are several ways of doing this. Here, we only consider the non-linear limitation of the ditribution coefficient of the N scheme.

ST-NLim scheme We limit the β_i^N of the N scheme.[4] In scalar this scheme is defined by:

$$\phi_i^{K_s} = \phi_i^{\text{Nlim}} = \beta_i^{\text{Nlim}} \Phi^{K_s}, \quad \beta_i^{Nlim} = \beta_i^{\text{N},+} / \sum_{j \in K_s} \beta_j^{\text{Nlim},+} \tag{11}$$

with $\beta_j^{\text{N}} = \phi_i^{\text{N}} / \Phi_i^{K_s}$. The Nlim scheme verifies both the monotonicity requirement ($\Phi_i^{T_s} \times \Phi_i^{\text{N}} \geq 0$), and the accuracy condition (6) (β_i^{Nlim} bounded).

4 Results

We test these schemes on several test cases. The ST-LDA(P^k) will be tested on acoustic problems using Linearised Euler Equations. For more details on the implementation of LEE for RDS we refer to the work of Koloszar et al. [4]. The goal of the last test case of this article is to test the monotonicity of Nlim(P^k).

4.1 Gauss Pulse

The first test case is the propagation of a Gauss pulse on $[-20; 20] \times [-20; 20]$ in a quiescent flow. We simulate this with the formulation of Linearised Euler equations of Bailly et al. [3]. The density and pressure of the mean flow are $\rho_0 = 1.225$ and $p_0 = 101325$ yielding the speed of sound of the mean flow $c_0 = 340$. The definition of this perturbation is:

$$\begin{cases} \rho & = e^{-(ln2)\frac{x^2+y^2}{9}} = f_{pulse}(x, y) \\ \rho_0 u & = 0 \\ \rho_0 v & = 0 \\ p & = {C_0}^2 f_{pulse}(x, y) \end{cases} \tag{12}$$

[4] In scalar the distribution coefficients of N scheme can be defined by $\beta_j^{\text{N}} = \Phi_i^{\text{N}} / \Phi_i^{K_s}$. In the case of non-linear system of equations, this definition is not any more valid. So, we use the wave decomposition proposed by Abgrall to demonstrate the monotonicity of N scheme [1]. For more details on this methodology we refer to [6].

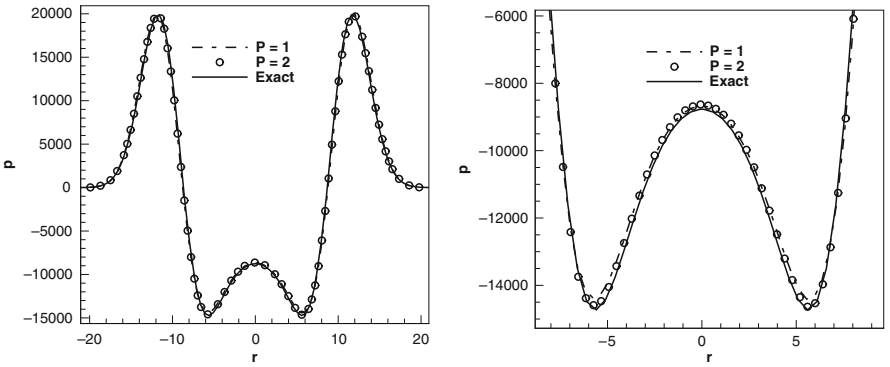


Fig. 2 Propagation of a Gauss pulse: Slice at $t = 0.03$, comparison between ST-LDA(P^1) and ST-LDA(P^2)

We perform this test case on a mesh of 40×40 degrees of freedom. On Fig. 2, we compare the results obtained with linear and quadratic elements on meshes having exactly the same number of degrees of freedom. Both schemes show a good agreement with the analytical solution. Even if the results obtained with ST-LDA(P^1) and ST-LDA(P^2) are very similar, we can already see a slight improvement brought by the quadratic discretisation. In fact, to capture well the acoustic waves, ST-LDA(P^2) needs as much degrees of freedom per wave length as ST-LDA(P^1). And with this amount, the result is already good with ST-LDA(P^1). The derivation of the Fourier analysis showing this is in [4, 5].

4.2 Monopole

We consider now, a monopole in an uniform flow at $M = 0.5$. The source is defined by:

$$S = f(x, y)\sin(\omega t)\begin{bmatrix}1\\0\\0\\1\end{bmatrix}, \quad f(x, y) = \epsilon \exp^{-\alpha((x-x_s)^2+(y-y_s)^2} \tag{13}$$

In this test case, the location of the monopole (x_s, y_s) is $(0, 0)$, the amplitude ϵ is set to 0.5, the thickness of the source is $\frac{\ln 2}{2}$ and the angular frequency is $\omega = \frac{2\pi}{30}$. The computational domain is $[-100; 100] \times [-100; 100]$ and the mesh has 200×200 degrees of freedom. We first plot, on Fig. 3, the pressure iso-lines obtained with ST-LDA(P^2) at $t = 270$ and we compare the result obtained with ST-LDA(P^1) and ST-LDA(P^2) to a reference solution on a slice done at $y = 0$. The reference

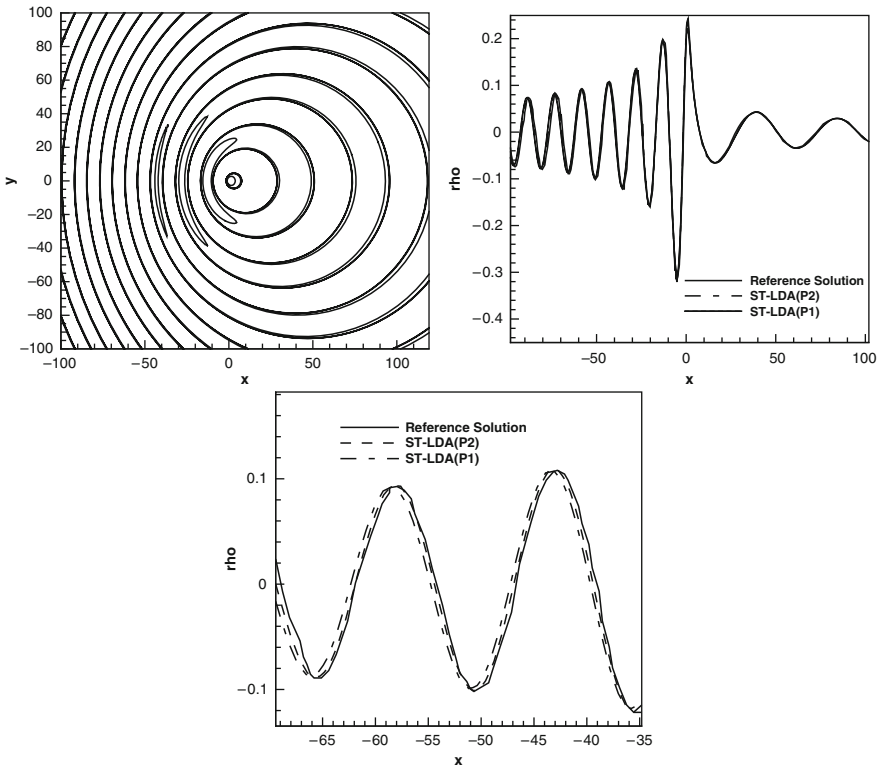


Fig. 3 Propagation of a monopole ($t = 270$): pressure iso-lines (*left*); slice at $y = 0.0$, comparison between the reference solution, ST-LDA(P^1) and ST-LDA(P^2) (*right*), zoom on the upstream waves (*bottom*)

solution was obtained by Bailly et al. [3], using a 7th order Dispersion-Relation-Preserving Finite Difference scheme. We can see that there are two acoustic waves propagating. The first propagates upstream with a velocity $1 - M$ and a wavelength $\lambda_{up} = (1 + M)\lambda$ (λ being the wavelength of the monopole). The second propagates downstream with a velocity $1 + M$ and a wavelength $\lambda_{down} = (1 - M)\lambda$. Both ST-LDA($P^1$) and ST-LDA($P^2$) give very satisfactory results. The upstream wave is a bit dispersed by ST-LDA(P^1) whereas ST-LDA(P^2) preserve it very well. This result really shows the ability of RDS because we obtain a similar result as a 7th order finite difference scheme using the same number of degrees of freedom.

4.3 Double Mach Reflection

Now, we want to test the monotonicity and the accuracy of Nlim(P^2). The double Mach reflection test case was first proposed in [9]. It is very interesting to test the

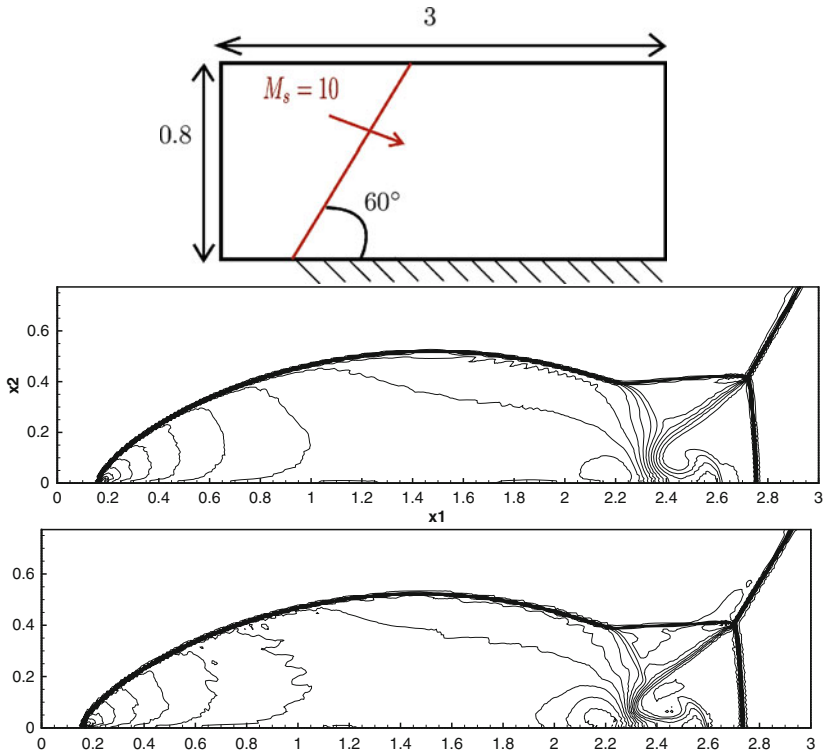


Fig. 4 Double mach reflection: setting of the test case (*top*) density iso-lines at $t = 0.2$ Nlim(P^1) (*middle*) and Nlim(P^2) (*bottom*)

accuracy and the robustness of a scheme. It consists of the interaction of a planar right-moving $M = 10$ shock with a 30° ramp. We consider that the ramp is aligned with the x-axis. The computational domain is $[0; 3] \times [0; 0.8]$ and the ramp start at $x = \frac{1}{6}$. The initial shock forms an angle of 60° with the x-axis as sketched on Fig. 4. On the top boundary, we impose the movement of the shock. We look at the solution at $t = 0.2$ and, on Fig. 4, we plot the solution obtained with ST-Nlim(P^1) and ST-Nlim(P^2) on a mesh with $h = 1/100$ (the P^2 mesh has the same number of degrees of freedom). We can see that all the schemes give a monotone result. Moreover, the shock and the slip line are better resolved with ST-Nlim(P^2).

The real challenge of this test case is to catch the Kelvin–Helmholtz instabilities. To see them, we zoom in on the triple point region. Of course, this first mesh does not allow to see the instabilities. But, if we decrease the mesh spacing to $h = 1/240$ then, with the quadratic discretisation, we can see the instabilities appearing on Fig. 5. If we compare with the results obtained by Ricchiuto in [6] with the same number of degrees of freedom, the instabilities are better resolved.

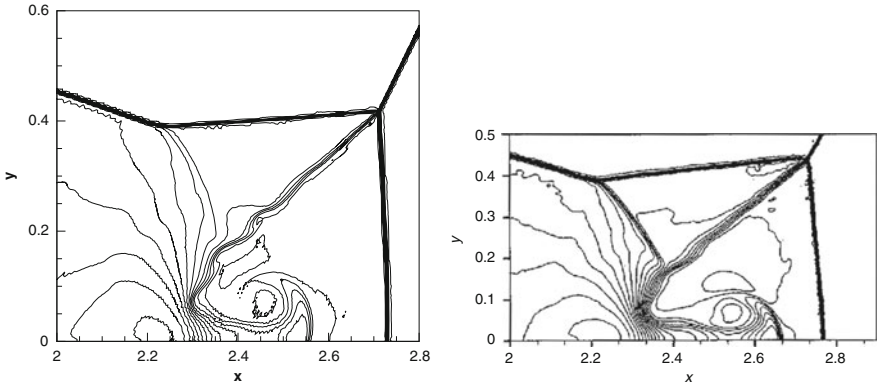


Fig. 5 Double mach reflection: zoom in on the triple point density isolines at $t = 0.2$, solution obtained with Nlim(P^2) (*left*) with mesh spacing $h = 1/240$ (size of the sub-elements); Nlim(P^1) of Ricchiuto (*right*) $h = 1/240$

5 Conclusion

In this article, we have presented the extension of the space-time schemes of Ricchiuto et al. [7] to systems of equations. We have shown that even if in general, upwind schemes are not used to simulate the propagation of sound, RDS are performing very well on Linearised Euler equations. Moreover, even if on simple test cases the solution obtained with ST-LDA(P^1) and ST-LDA(P^2) are very similar, the improvement brought by quadratic elements is more visible on more complex problems such as the propagation of a monopole. Finally, we have designed some high order monotone schemes. Here again, the better resolution due to quadratic elements was obvious and we have shown that ST-LN(P^2, P^1) can be considered as monotone.

References

1. Abgrall, R., Mezine, M.: Construction of second-order accurate monotone and stable residual distribution schemes for steady flow problems. J. Comput. Phys., **195**, pp 474–507 (2004)
2. Abgrall, R., Roe, P.L.: High order fluctuation schemes on triangular meshes. J. Sci. Comput., **19(3)**, 3–36 (2003)
3. Bailly, C., Juvé, D.: Numerical Solution of Acoustic Propagation Problems Linearized Euler's Equations. In: 6th AIAA/CEAS Aeroacoustic Conference (2009)
4. Koloszar, L., Villedieu, N., Quintino, T., Rambaud, P., Anthoine, J.: Application of Residual Distribution Method for wave propagation. In proceeding of 15th AIAA/CEAS Aeroacoustics Conference (30th AIAA Aeroacoustics Conference), Miami (2009)
5. Koloszar, L., Villedieu, N., Quintino, T., Rambaud, P., Anthoine, J.: Artificial Numerical Damping for Linearized Euler Equation implemented in Residual Distribution Method. In Proceedings of the 3rd European Conference for Aero-Space Sciences

6. Ricchiuto, M.: Construction and analysis of compact residual discretizations for conservation laws on unstructured meshes. PhD thesis Université Libre de Bruxelles (2005)
7. Ricchiuto, M., Abgrall, R., Deconinck H.: Construction of very high order residual distributive schemes for unsteady scalar advection: preliminary results. VKI LS 2003-05 33rd computational fluid dynamics-novel methods for solving convection dominated systems (2003)
8. Ricchiuto, M., Villedieu, N., Abgrall, R., Deconinck, H.: High-order residual distribution schemes: discontinuity capturing crosswind dissipation and extension to advection-diffusion, VKI LS on Higher Order Discretization Methods for Computational Physics, Von Karman Institute for Fluid Dynamics (2005)
9. Woodward, W.A., Colella, P.: The numerical simulation of two-dimensional flows with strong shocks. In J. Comput. Phys. **54**, pp 115–173 (1984)

Implicit–Explicit Backward Difference Formulae Discontinuous Galerkin Finite Element Methods for Convection–Diffusion Problems

Miloslav Vlasák and Vít Dolejší

Abstract We deal with a numerical solution of a scalar nonstationary convection–diffusion equation with a nonlinear convection and a linear diffusion. We carry out the space semi-discretization with the aid of the symmetric interior penalty Galerkin (SIPG) method and the time discretization by backward difference formulae (BDF) and suitable linearization of nonlinear convective term. The resulting scheme is unconditionally stable, has a high order of accuracy with respect to space and time coordinates and requires solutions of linear algebraic problems at each time step. We derive a priori error estimates in the $L^\infty(L^2)$-norm up to the order 6 in time.

1 Introduction

We numerically solve a nonstationary nonlinear convection–diffusion equation, which represents a model problem for the system of the compressible Navier–Stokes equations. The class of *discontinuous Galerkin* (DG) methods seems to be one of the most promising candidates to construct high order accurate schemes for solving of convection–diffusion problems. For a survey about DG methods, see [1] or [2]. An analysis of DG methods was presented in many papers, see, e.g., [3, 4, 8, 9].

In [4] we carried out the space semi-discretization of the scalar convection–diffusion equation with the aid of the *discontinuous Galerkin finite element* method and derived a priori error estimates. Within this contribution, we deal with the time discretization of the resulting system of ordinary differential equations. This paper can be viewed as extension of [5], where we presented a formulation of the general order (BDF DG) and derived error estimates up to the order 3. Here we extend this result up to the order 6, which is the highest achievable order due to the stability properties of BDF. For the details about BDF see [7] and [6].

M. Vlasák (✉) and V. Dolejší
Charles University Prague, Faculty of Mathematics and Physics, Sokolovská 83, Prague, Czech Republic
e-mail: vlasakmila@gmail.com, dolejsi@karlin.mff.cuni.cz

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_99,

2 Continuous Problem

Let $\Omega \subset R^d$ ($d = 2$ or 3) be a bounded polyhedral domain and $T > 0$. We set $Q_T = \Omega \times (0,T)$. By $\overline{\Omega}$ and $\partial\Omega$ we denote the closure and boundary of Ω, respectively. Let us consider the following *initial-boundary value problem*: Find $u : Q_T \to R$ such that

$$\frac{\partial u}{\partial t} + \nabla \cdot \mathbf{f}(u) = \varepsilon\, \Delta u + g \quad \text{in } Q_T, \tag{1}$$

$$u\,\big|_{\partial\Omega\times(0,T)} = u_D, \tag{2}$$

$$u(x,0) = u^0(x), \quad x \in \Omega. \tag{3}$$

In (1)–(3), $\mathbf{f} = (f_1, \ldots, f_d)$, $f_s \in C^2(R)$, $f_s(0) = 0$, $s = 1, \ldots, d$ represents convective terms, $\varepsilon > 0$ plays a role of viscosity, $g \in C([0,T]; L^2(\Omega))$ represents volume sources. The Dirichlet boundary condition is given over $\partial\Omega$ by u_D, which is the trace of some $u^* \in C([0,T]; H^1(\Omega)) \cap L^\infty(Q_T)$ is given over $\partial\Omega \times (0,T)$ and $u^0 \in L^2(\Omega)$ is an initial condition. We use the standard notation for Lebesgue, Sobolev and Bochner function spaces (see, e.g., [10]).

In order to introduce the concept of a weak solution, we define the forms

$$(u,w) = \int_\Omega uw\,\mathrm{d}x, \quad u,\, w \in L^2(\Omega),$$

$$a(u,w) = \varepsilon \int_\Omega \nabla u \cdot \nabla w\,\mathrm{d}x, \quad u,\, w \in H^1(\Omega),$$

$$b(u,w) = \int_\Omega \nabla \cdot \mathbf{f}(u)\, w\,\mathrm{d}x, \quad u \in H^1(\Omega) \cap L^\infty(\Omega),\ w \in L^2(\Omega),$$

Definition 1. We say that a function u is a *weak solution* of (1)–(3) if the following conditions are satisfied

$$\text{a)}\ \ u - u^* \in L^2(0,T; H_0^1(\Omega)), \quad u \in L^\infty(Q_T), \tag{4}$$

$$\text{b)}\ \ \frac{\mathrm{d}}{\mathrm{d}t}(u(t),w) + b(u(t),w) + a(u(t),w) = (g(t),w)$$

for all $w \in H_0^1(\Omega)$ in the sense of distributions on $(0,T)$,

$$\text{c)}\ \ u(0) = u^0 \quad \text{in } \Omega.$$

By $u(t)$ we denote the function on Ω such that $u(t)(x) = u(x,t)$, $x \in \Omega$.

With the aid of techniques from [11] and [12], it is possible to prove that there exists a unique weak solution. We shall assume that the weak solution u is sufficiently regular, namely,

$$u \in W^{1,\infty}(0,T; H^{p+1}(\Omega)) \cap W^{k,\infty}(0,T; H^1(\Omega)), \quad u^{(k+1)} \in L^\infty(0,T; L^2(\Omega)).$$

where $u^{(k)} = \partial^k u/\partial t^k$, an integer $p \geq 1$ will denote a given degree of polynomial approximationsin in space and $k = 1, \ldots, 6$ desired order of convergence in time. Such a solution satisfies problem (1)–(3) pointwise.

3 Space Semi-Discretization

We discretize problem (4) in space with the aid of the *discontinuous Galerkin finite element method* with *symmetric treatment of stabilization terms* and *interior and boundary penalties*. This approach is called the SIPG variant of the DGFE method, see [1]. We derived the space discretization of (1)–(3) by the SIPG variant of DGFE method in [4] hence here we present only the final expressions.

Let $\mathscr{T}_h$ ($h > 0$) be a partition of the domain Ω into a finite number of closed d-dimensional mutually disjoint simplices K i.e., $\overline{\Omega} = \bigcup_{K \in \mathscr{T}_h} K$. By ∂K we denote the boundary of element $K \in \mathscr{T}_h$ and set $h_K = \operatorname{diam}(K)$, $h = \max_{K \in \mathscr{T}_h} h_K$. We set Γ the faces of $\mathscr{T}_h$ ($\Gamma = \bigcup_{K \in \mathscr{T}_h} \partial K$). For the error estimates we assume the mesh be regular.

Furthermore, we use the following notation: $\boldsymbol{n} = (n_1, \ldots, n_d)$ – a normal vector to Γ which is well defined almost everywhere (on $\partial\Omega$ we use outer normal, inside of Ω we use one (arbitrary but fixed) direction at every point of Γ).

We use well known *broken Sobolev spaces* $H^s(\Omega, \mathscr{T}_h) = \{w; w|_K \in H^s(K) \; \forall \; K \in \mathscr{T}_h\}$. For $w \in H^1(\Omega, \mathscr{T}_h)$, we introduce the following notation on $\Gamma \setminus \partial\Omega$: $w_R(x) = \lim_{\delta \to 0+} w(x + \delta \boldsymbol{n})$, $w_L(x) = \lim_{\delta \to 0-} w(x + \delta \boldsymbol{n})$, $\langle w \rangle = \frac{1}{2}(w_R + w_L)$, $[w] = w_L - w_R$ and on $\partial\Omega$ we put $w_L(x) = \lim_{\delta \to 0-} w(x + \delta \boldsymbol{n})$, $\langle w \rangle = w_L$, $[w] = w_L$.

For $u, w \in H^2(\Omega, \mathscr{T}_h)$ we set

$$A_h(u, w) = \sum_{K \in \mathscr{T}_h} \int_K \nabla u \cdot \nabla w \, \mathrm{d}x - \int_\Gamma \Big(\langle \nabla u \rangle \cdot \boldsymbol{n}[w] + \langle \nabla w \rangle \cdot \boldsymbol{n}[u] \Big) \, \mathrm{d}S \tag{5}$$
$$+ \int_\Gamma \sigma [u] \, [w] \, \mathrm{d}S$$

$$b_h(u, w) = \int_{\Gamma \setminus \partial\Omega} H(u_L, u_R, \boldsymbol{n}) \, [w] \, \mathrm{d}S + \int_{\partial\Omega} H(u_L, u_D, \boldsymbol{n}) \, w_L \, \mathrm{d}S \tag{6}$$
$$- \sum_{K \in \mathscr{T}_h} \int_K \mathbf{f}(u) \cdot \nabla w \, \mathrm{d}x, \quad u, \, w \in H^1(\Omega, \mathscr{T}_h), \quad u \in L^\infty(\Omega)$$

$$\ell_h(w)(t) = (g(t), w) - \int_{\partial\Omega} (\nabla w \cdot \boldsymbol{n} \, u_D(t) - \sigma u_D(t) \, w) \, \mathrm{d}S. \tag{7}$$

The penalty parameter function σ in (5) and (7) along the face $e \subset \Gamma$ is defined by $\sigma|_e = C_W/(h_K + h_{\widetilde{K}})$, $e = K \cap \widetilde{K}$, where C_W is constant large enough to A_h be coercive. The function $H(\cdot, \cdot, \cdot)$ in the face integrals in (6) is called the *numerical flux*, well-known from the finite volume method and it approximates the terms

$\mathbf{f}(u) \cdot \boldsymbol{n}$. We assume the numerical fluxes H be Lipschitz continuous, conservative and consistent. Now we define the space of discontinuous piecewise polynomial functions

$$S_h = S^{p,-1}(\Omega, \mathscr{T}_h) = \{w; w|_K \in P_p(K) \; \forall \; K \in \mathscr{T}_h\}, \tag{8}$$

where $P_p(K)$ denotes the space of all polynomials on K of degree $\leq p$, where the integer $p \geq 1$ is a given degree of approximation.

We find that the exact solution of (4) with property (5) satisfies the identity

$$\left(\frac{\partial u}{\partial t}(t), w_h\right) + \varepsilon A_h(u(t), w_h) + b_h(u(t), w_h) = \ell_h(w_h)\,(t) \tag{9}$$

for all $w_h \in S_h$ and all $t \in (0, T)$.

The (semi)-discrete problem (9) represents a system of ordinary differential equations (ODEs) which is solved by a suitable solver in the next section.

4 Time Discretization

Since problem (9) is stiff, it is necessary to solve it with a method having a large stability domain. BDF represent the most popular approach in the field of the multi-step methods. Since these formulae are implicit and we need somehow to avoid the nonlinearity in our problem we use explicit extrapolation for convective term which leads to a sufficiently stable method which requires a solution of a linear algebraic problem at each time step. To define the fully discrete soltion we set time partition $t_s = s\tau$, $s = 0, \ldots, r$ with time steps $\tau = T/r$. Following the notation from [5] we define fully discrete solution.

Definition 2. We say that the set of functions $U^s \in S_h$, $s = 0, \ldots, r$ is an approximate solution of problem (9) obtained by the k step IMEX BDF DGFE scheme if

$$\left(\sum_{v=0}^{k} \alpha_v U^{s+v}, w\right) + \tau\varepsilon A_h(U^{s+k}, w) + \tau b_h(\hat{U}^{s+k}, w) = \tau \ell_h(w)(t_{s+k}) \;\; \forall w \in S_h,$$

$$\hat{U}^{s+k} = \sum_{v=1}^{k} \beta_v U^{s+k-v} \tag{10}$$

where $U^0, \ldots, U^{k-1}$ are given to start the method.

The choice of coefficients α_v and β_v is described in [5].

5 Error Estimates

Our goal is to analyse the error estimates of the approximate solution U^s, $s = 0, \ldots, r$ obtained by the method (10). In the sequel we use the notation $u^s = u(t_s)$ and $\hat{u}^{s+k} = \sum_{v=1}^{k} \beta_v u^{s+k-v}$, $\xi^s = U^s - \Pi u^s$, $\hat{\xi}^{s+k} = \sum_{v=1}^{k} \beta_v \xi^{s+k-v}$, $\eta^s = \Pi u^s - u^s$ and $e^s = U^s - u^s = \xi^s + \eta^s$, where Π be the A_h projection on S_h.

Let $|||w|||^2 := A_h(w, w)\ \forall w \in H^2(\Omega, \mathscr{T}_h)$ and $\|\cdot\| := \|\cdot\|_{L^2(\Omega)}$.

Lemma 1. *Let u be sufficiently regular. Then*

$$\|\eta^{s+k}\| \le C h^{p+1}, \tag{11}$$

$$\left| \left(\sum_{v=0}^{k} \alpha_v u^{s+v} - \tau \frac{\partial u}{\partial t}(t_{s+k}), w \right) \right| \le C \tau^{k+1} \|w\| \quad \forall w \in S_h, \tag{12}$$

$$\left| \left(\sum_{v=0}^{k} \alpha_v \eta^{s+v}, w \right) \right| \le C \tau h^{p+1} \|w\| \quad \forall w \in S_h, \tag{13}$$

$$\left| b_h(u^{s+k}, w) - b_h(\hat{U}^{s+k}, w) \right| \le C(h^{p+1} + \tau^k + \|\hat{\xi}^{s+k}\|) |||w||| \ \forall w \in S_h \tag{14}$$

Proof. The proof of Lemma 1 can be found in [5].

Lemma 2. *Let operator A satisfy $(Av, w) = A_h(v, w)$ for all $v, w \in S_h$. Let us set sequence of operators γ_j such that $(\tau\varepsilon A + I \sum_{v=0}^{k} \alpha_{k-v} z^v)^{-1} = \sum_{j=0}^{\infty} \gamma_j z^j$ for any complex number z. Then for $k = 1, \ldots, 6$ we have*

$$\|\gamma_j\| \le C, \quad \forall j = 0, 1, \ldots \tag{15}$$

$$\tau \sum_{j=0}^{\infty} |||\gamma_j w|||^2 \le \frac{C}{\varepsilon} \|w\|^2, \quad w \in S_h. \tag{16}$$

Proof. Since A is symmetric and since we can consider γ_j as operator function of $\tau\varepsilon A$ ($\gamma_j = F_j(\tau\varepsilon A)$), we know that γ_j are symmetric too. We can also see that eigenvalues of γ_j are in the form $F_j(\tau\varepsilon\lambda)$, where λ is eigenvalue of A. Let us prove that there exists constant $0 < C(k) < 1$ such that

$$|F_j(\tau\varepsilon\lambda)| \le C(e^{-cj} + (1 - \frac{\tau\varepsilon\lambda}{2})^j), \quad \tau\varepsilon\lambda \le C(k), \tag{17}$$

$$|F_j(\tau\varepsilon\lambda)| \le C \frac{e^{-cj}}{\tau\varepsilon\lambda}, \quad \tau\varepsilon\lambda \ge C(k), \tag{18}$$

where $c > 0$. The proof of (18) can be found in [13, Lemma 10.3]. To prove (17) we could also follow the proof from [13, Lemma 10.3] only the final expressions must be estemated sharper. The former expressions from the [13, Lemma 10.3] are unsuficient for our purpose.

Now we are ready to prove estimates (15) and (16). The estimate (15) directly follows from boundedness of eigenvalues by estimates (17) and (18). We set v_i orthonormal eigenvectors of A and decompose $w = y + z = \sum_i y_i v_i + \sum_i z_i v_i$, where y represents part with eigenvectors with small eigenvalues of $\tau\varepsilon A$ (those that can be estimated by (17)) and z represents part with eigenvectors with large eigenvalues. Then $\tau \sum_{j=0}^{\infty} |||\gamma_j w|||^2 \leq 2\tau \sum_{j=0}^{\infty} |||\gamma_j y|||^2 + 2\tau \sum_{j=0}^{\infty} |||\gamma_j z|||^2$. At first we estimate y part.

$$2\tau \sum_{j=0}^{\infty} |||\gamma_j y|||^2 = 2 \sum_{j=0}^{\infty} \sum_i y_i^2 \tau\lambda_i \|\gamma_j v_i\|^2 \leq C \sum_i y_i^2 \tau\lambda_i \sum_{j=0}^{\infty} (e^{-cj} + (1-\tau\varepsilon\lambda_i)^j)^2$$
$$\leq C \sum_i y_i^2 \tau\lambda_i \sum_{j=0}^{\infty} (e^{-cj} + (1-\tau\varepsilon\lambda_i)^j) = C \sum_i y_i^2 \tau\lambda_i \left(\frac{1}{1-e^{-c}} + \frac{1}{\tau\varepsilon\lambda_i}\right) \leq \frac{C}{\varepsilon} \|y\|^2 \tag{19}$$

And now the similar for z part.

$$2\tau \sum_{j=0}^{\infty} |||\gamma_j z|||^2 = 2 \sum_{j=0}^{\infty} \sum_i z_i^2 \tau\lambda_i \|\gamma_j v_i\|^2 \leq C \sum_i z_i^2 \tau\lambda_i \sum_{j=0}^{\infty} \left(\frac{e^{-cj}}{\tau\varepsilon\lambda_i}\right)^2 \tag{20}$$
$$\leq C \sum_i z_i^2 \tau\lambda_i \sum_{j=0}^{\infty} \frac{e^{-cj}}{\tau\varepsilon\lambda_i} = \frac{C}{\varepsilon} \sum_i z_i^2 \frac{1}{1-e^{-c}} \leq \frac{C}{\varepsilon} \|z\|^2$$

Finally (16) follows from (19), (20) and $\|w\|^2 = \|y\|^2 + \|z\|^2$.

Theorem 1. *Let u be the exact solution of problem (4) satisfying (5). Let the mesh be regular and the numerical fluxes H be Lipschitz continuous, conservative and consistent. Let U^s, $s = 0, \ldots, r$ be the approximate solution defined by (10). Then for $k = 1, \ldots, 6$ we have*

$$\max_{n=0,\ldots,r} \|U^n - u^n\| \leq O\left(h^p + \tau^k + \sum_{v=0}^{k-1} \|U^v - u^v\|\right) e^{TC(1+1/\varepsilon)} \tag{21}$$

Proof. Since $U^n - u^n = \xi^n + \eta^n$, in virtue of Lemma 1, it is sufficient to estimate $\|\xi^n\|$ only. Let us multiply (9) by τ for $t = t_{s+k}$ and substract this equation from (10). Then we have

$$\left(\sum_{v=0}^{k} \alpha_v \xi^{s+v} + \tau\varepsilon A \xi^{s+k}, w\right) = \left(\tau \frac{\partial u}{\partial t}(t_{s+k}) - \sum_{v=0}^{k} \alpha_v u^{s+v}, w\right) \tag{22}$$
$$- \left(\sum_{v=0}^{k} \alpha_v \eta^{s+v}, w\right) + \tau \left(b_h(u^{s+k}, w) - b_h(\hat{U}^{s+k}, w)\right).$$

Setting $w = \gamma_{n-k-s}\xi^n$ and summing over $s = 0, \ldots, n-k$ we obtain

$$\begin{aligned}\|\xi^n\|^2 = &-\left(\sum_{s=0}^{k-1}\sum_{v=0}^{s}\alpha_v\gamma_{n-k-s+v}\xi^s, \xi^n\right) \\ &+\sum_{s=0}^{n-k}\tau\left(b_h(u^{s+k}, \gamma_{n-k-s}\xi^n) - b_h(\hat{U}^{s+k}, \gamma_{n-k-s}\xi^n)\right) \\ &+\sum_{s=0}^{n-k}\left(\tau\frac{\partial u}{\partial t}(t_{s+k}) - \sum_{v=0}^{k}\alpha_v u^{s+v} - \sum_{v=0}^{k}\alpha_v\eta^{s+v}, \gamma_{n-k-s}\xi^n\right).\end{aligned} \tag{23}$$

Applying Lemma 1, Lemma 2, Young's inequality and $\sum_{s=0}^{n-k}\tau \le T$ we obtain

$$\|\xi^n\|^2 \le \frac{1}{2}\|\xi^n\|^2 + C\left(\tau^{2k} + h^{2p+2} + \sum_{v=0}^{k-1}\|\xi^v\|^2\right) + \tau\frac{C}{\varepsilon}\sum_{s=0}^{n-1}\|\xi^s\|^2. \tag{24}$$

Now it is sufficient to apply Gronwall's lemma to obtain the result. □

Remark 1. The estimate (21) cannot be used for $\varepsilon \to 0+$, because it blows up exponentially. The nonlinearity of the convective terms represents a serious obstacle.

Acknowledgements This work is a part of the research project MSM 0021620839 financed by the Ministry of Education of the Czech Republic, and it was partly supported by the Grant No. 10209/B-MAT/MFF of the Grant Agency of the Charles University Prague. The research of M. Vlasák is also supported by the project LC06052 financed by MSMT (Necas Center for Mathematical Modeling).

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 (2002)
2. Cockburn, B., Karniadakis, G.E., Shu, C.W. (eds.): Discontinuous Galerkin methods. Springer, Berlin (2000)
3. Dolejší, V., Feistauer, M., Kucera, V., Sobotíková, V.: An optimal $L^\infty(L^2)$-error estimate of the discontinuous Galerkin approximation of a nonlinear non-stationary convection–diffusion problem. IMA J. Numer. Anal. **28**(3), 496–521 (2008)
4. Dolejší, V., Feistauer, M., Sobotíková, V.: A discontinuous Galerkin method for nonlinear convection–diffusion problems. Comput. Methods Appl. Mech. Engrg. **194**, 2709–2733 (2005)
5. Dolejší, V., Vlasák, M.: Analysis of a BDF – DGFE scheme for nonlinear convection–diffusion problems. Numer. Math. **110**(4), 405–447 (2008)
6. Hairer, E., Wanner, G.: Solving ordinary differential equations II, Stiff and differential-algebraic problems. Springer, Berlin (2002)
7. Hairer, E., Norsett, S.P., Wanner, G.: Solving ordinary differential equations I, Nonstiff problems. No. 8 in Springer Series in Computational Mathematics. Springer, Berlin (2000)
8. Houston, P., Schwab, C., Süli, E.: Discontinuous *hp*-finite element methods for advection–diffusion problems. SIAM J. Numer. Anal. **39**(6), 2133–2163 (2002)

9. Houston, P., Robson, J., Süli, E.: Discontinuous Galerkin finite element approximation of quasilinear elliptic boundary value problems I: The scalar case. IMA J. Numer. Anal. **25**, 726–749 (2005)
10. Kufner, A., John, O., Fuck, S.: Function spaces. Academia, Prague (1977)
11. Lions, P.L.: Mathematical topics in fluid mechanics. Oxford Science Publications (1996)
12. Rektorys, K.: The method of discretization in time and partial differential equations. Reidel, Dodrecht (1982)
13. Thomée, V.: Galerkin finite element methods for parabolic problems. 2nd revised and expanded ed. Berlin: Springer. xii, (2006)

A Cut-Cell Finite-Element Method for a Discontinuous Switch Model for Wound Closure

S.V. Zemskov, F.J. Vermolen, E. Javierre, and C. Vuik

Abstract A mathematical model for epidermal wound healing is considered. The model is based on a moving boundary problem for the wound edge in which the edge moves if a generic epidermal growth factor exceeds a given threshold value. We use a Galerkin finite-element method to solve the equations for the growth factor concentration. The moving boundary (wound edge) is tracked using a level-set method with a local adaptive mesh refinement in the interface region. To deal with the reaction-diffusion equation for the growth factor, a cut-cell method has been implemented. This cut-cell method warrants the integration over a continuous reaction term elementwisely. The results improved with respect to the results that were obtained without the use of the cut-cell method.

1 Introduction

Wound healing or soft tissue regeneration, involves cell migration, the production and decay of growth factors and a (re-)establishment of the vascular network surrounding the area with an increased mitotic activity. Experimental validation of the models of both complicated biological processes is indispensable. The present paper focuses on a very simplified model for wound closure. This model can be used for intra-osseous and epidermal wound healing. Since the thickness of the epidermis is in the order of 1 mm, it suffices to consider a two-dimensional approach

S.V. Zemskov (✉), F.J. Vermolen, and C. Vuik
Delft Institute of Applied Mathematics, Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, Mekelweg 4, 2628 CD Delft, The Netherlands
e-mail: F.J.Vermolen@tudelft.nl

E. Javierre
Centro de Investigación Biomédica en Red en Bioingeniería, Biomateriales y Nanomedicina (CIBER-BBN), Group of Structural Mechanics and Materials Modelling (GEMM), Aragón Institute of Engineering Research (I3A), Universidad de Zaragoza, Aragon Health Sciences Institute. Agustín de Betancourt Building, C/María de Luna 7, Campus Rio Ebro, University of Zaragoza, 50018, Zaragoza, Spain

G. Kreiss et al. (eds.), *Numerical Mathematics and Advanced Applications 2009*,
DOI 10.1007/978-3-642-11795-4_100,

for epidermal closure (re-epithelialization). Hence, we consider a two-dimensional model in the present paper.

When a wound occurs, blood vessels are cut and blood enters the wound. Due to blood coagulation, the wound is temporarily closed and as a result the blood vessels adjacent to the wound are also closed. In due course, contaminants will be removed from the wounded area and the blood vessel network will be restored, but initially due to insufficient blood supply, there will be a low concentration of nutrients which are necessary for cell division and wound healing. Wound healing, if it occurs, proceeds by a combination of several processes: wound contraction (due to pulling forces caused by fibroblasts entering the wound area underneath the epidermal cells), chemotaxis (movement of cells induced by a concentration gradient), neo-vascularization (formation of network of capillaries), synthesis of extracellular matrix proteins, and scar remodeling. Previous models incorporate cell mitosis, cell proliferation, cell death, capillary formation, oxygen supply and growth factor generation. These models contain visco-elasticity problems coupled with reaction-transport equations. We refer to [6] for an overview.

There is a lot of mathematical models in literature for wound healing and wound closure. In this paper, we do not intend to discuss the variety of models, but we aim at a description of the numerical solution method for one class of models: the models with a discontinuous switch mechanism. This model was initially proposed by [1] and enriched with a moving boundary formulation in [8]. A numerical procedure based on the finite-element method with a level set method is used to track the moving wound edge, is presented in [2]. Existence, uniqueness and mathematical properties of solutions of this problem were demonstrated in [7].

The start of the present paper is the introduction of the discontinuous switch model for re-epithelialization. Subsequently, the cut-cell method for an accurate determination of the solution in the vicinity of the interface is presented. Then, the cut-cell method is numerically compared with the classical finite-element method and finally some conclusions are drawn.

2 The Model

In this section the model based on the ideas of [1] is presented. Firstly, the model for the regeneration, decay and transport of a generic growth factor is given, and subsequently the healing process as a result of the presence of the growth factor is described (see [8]). Finally, a description of the coupling of the two models is presented.

We use Ω_1, Ω_2 and Ω_3 to denote the wound itself, the active layer and the outer tissue respectively. The active layer Ω_2 is a ring surrounding the wound region Ω_1. Since the wound is healing, the areas Ω_1, Ω_2 and Ω_3 are functions of time and to be determined as a part of the solution. Far away from the wound, that is at the boundary of the domain of computation, $\partial\Omega$, we assume that there is no transport of growth factor. The wound edge, the interface between the wound (Ω_1) and the active layer (Ω_2), is indicated by $W(t)$ (i.e., $W = \overline{\Omega}_1 \cap \overline{\Omega}_2$).

Let the total domain of interest be given by $\overline{\Omega} = \cup_{i=1}^{3} \overline{\Omega}_i$, which is Lipschitz, then, following [1], we state the fundamental equation for the transport, production and decay of the growth factor concentration, c, which reads as:

$$\frac{\partial c}{\partial t} - \text{ div } D \text{ grad } c + \lambda c = P \mathbf{1}_{\Omega_2(t)}(\mathbf{x}), \text{ for } (t, \mathbf{x}) \in (0, T] \times \Omega, \tag{1}$$

$$\frac{\partial c}{\partial n} = 0, \text{ for } (t, \mathbf{x}) \in (0, T] \times \partial\Omega, \tag{2}$$

$$\text{where } \mathbf{1}_{\Omega_2(t)}(\mathbf{x}) = \begin{cases} 1, \text{ for } \mathbf{x} \in \Omega_2(t) \\ 0, \text{ for } \mathbf{x} \in \Omega \setminus \Omega_2(t) \end{cases}, \tag{3}$$

As the initial condition, we have

$$c(0, \mathbf{x}) = 0, \text{ for } \mathbf{x} \in \Omega. \tag{4}$$

In the equations, the constants D, P and λ denote the constant diffusion coefficient, production rate constant and the decay coefficient of the growth factor. These constants are non-negative in our parabolic PDE. The growth factor concentration, c, is to be determined. Further, the second and third term in (1) respectively account for growth factor transport and growth factor loss. The right-hand side of (1) accounts for the production of the growth factor. Equation (2) represents the boundary condition and the indicator function $\mathbf{1}_{\Omega_2(t)}(\mathbf{x})$ accounts for the growth factor production taking place in the active layer only.

Healing at a certain location of the interface implies that the inward normal component of the velocity pointing into the wound, v_n, of the interface W is positive. In the present paper we use the assumption from [1] that the interface moves if and only if the growth factor concentration exceeds a threshold concentration $\hat{c}$. This implies that in order to determine whether the wound heals at a certain location on W at a certain time t, one needs to know the growth factor concentration there.

As it has been motivated in [8], we assume that the healing rate is proportional to the local curvature of the wound. Hence, in agreement with (5), the velocity component in the outward (from Ω_1, that is the wound) normal direction is given by

$$v_n = -(\alpha + \beta\kappa) w(c(t, \mathbf{x}) - \hat{c}), \text{ for } (t, \mathbf{x}) \in (0, T] \times W(t), \tag{5}$$

where κ is the local curvature and $\alpha, \beta \geq 0$ are considered as non-negative constants, prohibiting growth of the wound if $\kappa \geq 0$. Further, the function $w(s)$ falls within the class of Heaviside functions, that is $w(s) \in H(s)$, where $H(.)$ represents the family of Heaviside functions, for which we have

$$H : s \to \begin{cases} 0, & \text{if } s < 0, \\ \in [0, 1], & \text{if } s = 0, \\ 1, & \text{if } s > 0. \end{cases} \tag{6}$$

Some models with the same principles as the active layer and / or the discontinuous switch condition can be found in other works, see references in [7]. Further, the existence and uniqueness of solutions in $C^1((0,T);H^1(\Omega)) \cap C^0([0,T);H^1(\Omega))$ was demonstrated and analytic solutions in this function space were constructed in that paper as well.

3 The Method

The mathematical model described falls within the class of moving boundary problems. The position of interface, $W(t)$, has to be determined at each time step t what leads to re-identifying the parts of the computation domain ($\Omega_1(t)$, $\Omega_2(t)$ and Ω_3(t)).

As in [2], the Level Set method [3] is used to follow the evolution of the interface W during the simulation. The interface is identified as the zero level set of a continuous function ϕ which is defined at the initial time $t = 0$ by the following way:

$$\phi(0,\mathbf{x}) = \begin{cases} +\text{dist}(\mathbf{x}, W(0)), & \mathbf{x} \in \Omega_1(0), \\ 0, & \mathbf{x} \in W(0), \\ -\text{dist}(\mathbf{x}, W(0)), & \mathbf{x} \in \overline{\Omega}_2(0) \cap \overline{\Omega}_3(0). \end{cases}$$

Thus, ϕ is defined to be positive inside the wound and negative outside.

Subsequently, a convection equation is solved for the level-set function ϕ in which the velocity at the interface is determined from the local curvature of ϕ at the interface. The velocity is extended onto the entire domain of computation by advection in the appropriate upwind direction. For the following numerical reasons, it is attractive that ϕ is a signed distance function: 1. a reliable inverse interpolation to get the wound edge position, and 2. the straightforward computation of the local curvature. In order to enjoy this property, a reinitialization step is carried out so that $|\nabla\phi| = 1$ in Ω. In this work, a fast-marching method has been selected [4]. More details can be found in [2].

3.1 The Cut-Cell Approach

Since the interface moves only if locally the threshold $\hat{c}$ has been exceeded, an accurate approximation of the concentration at the interface is indispensible. The mathematical model and Level Set method described in the previous sections were implemented by Javierre et al. [2] using a finite-element method with piecewise linear basis functions. A structured triangulation with linear elements is used as a fixed basis mesh. The elements close to the active layer are refined according to certain criteria at each time step.

In the standard finite-element method, the interface concentration is determined by interpolation of the growth factor concentration at the wound edge. The right-hand side of (1) is discontinuous what leads to numerical wiggles at the edges of the active layer. A regularized version χ_ε of the characteristic function χ is used in [2] in order to diminish possible oscillations at the interface. A good approximation of the concentration is crucial for the determination whether or not the interface will move locally.

To eliminate this defect, we propose the use of a cut-cell approach which allows to adapt the existing triangulation to the edges of the active layer at each time step. In application to the model considered, the cut-cell method consists of an additional refinement of FE mesh on the elements intersected by either the interface W or the outer boundary of the active layer $\overline{\Omega}_2 \cap \overline{\Omega}_3$. Such an approach allows to apply the developed technique of numerical integration over new elements where the integrand remains continuous.

The level set function ϕ representing the distance to the interface is defined in each node of the FE mesh (positive for nodes inside the wounded region and negative outside). We have created a module to find elements intersected by a defined level line of ϕ and to perform the subdivision of each element found into triangular sub-elements.

To avoid the appearance in the new refinement of ill-shaped triangles which might be too small with respect to already existing elements, we assume that the distance d between an intersection point and the nearest node fulfils the following condition

$$d < \frac{\min(\Delta_x, \Delta_y)}{10},$$

where Δ_x and Δ_y are horizontal and vertical steps of initial Cartesian mesh respectively. Under such a condition, there are three possible cases for dividing an intersected element (Fig. 1).

Hence, if the intersection point is too close to one of the existing nodes, such an intersection is not registered and we do not add any new point to the set of nodes. In case of subdividing an element into three sub-elements (i.e., by two registered intersection points), one new triangular element is formed immediately by cutting

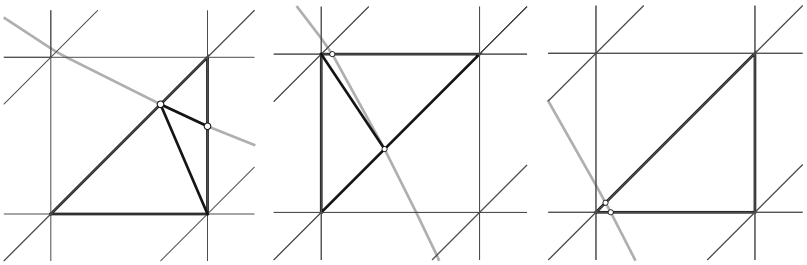


Fig. 1 Intersected element is divided by three sub-elements (*left*), two sub-elements (*center*) or is not divided (*right*)

off from the initial element. From two possible of element divisions, we exclude the variant with the most obtuse angle. During calculations each new point is tested whether it belongs to the element inside the active layer or not and receives the corresponding value.

4 Comparison Between Cut-Cell and Classical Finite-Element Method

To compare the cut-cell method with the classical Galerkin finite-element method, we plot the wound edge concentration profile of an elliptical wound in Fig. 2 for

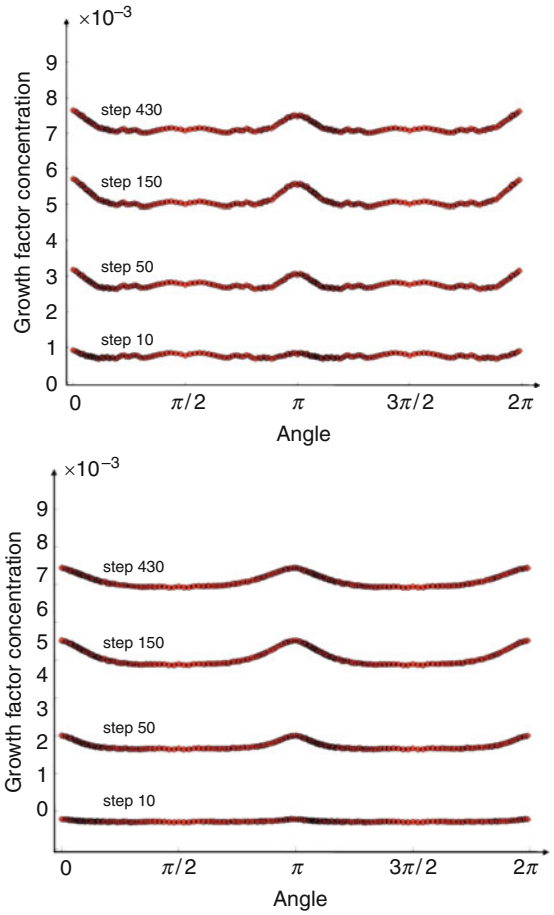


Fig. 2 The epidermal growth factor concentration on the interface after 10, 50, 150 and 430 time steps. At the *top* and *bottom*, the classical FEM and cut-cell method has been used respectively

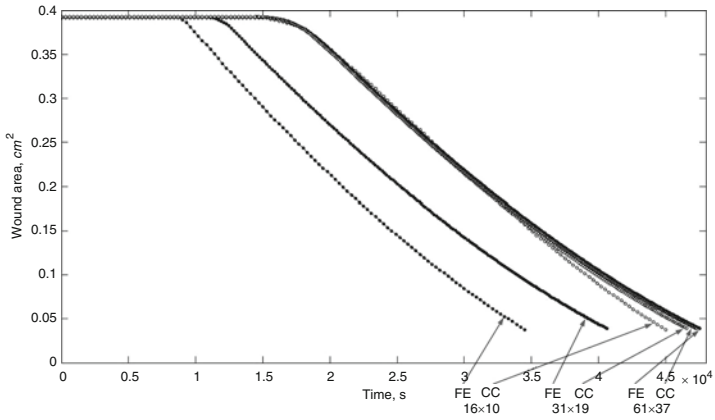


Fig. 3 Change of the wound area during the simulation for 16×10, 31×19 and 61×37 gridnodes for the cut-cell (CC) and standard finite-element method (FE)

the two methods. It can be seen that the peripheral concentration profile in the classical method exhibits an oscillatory behavior due to the interpolation step and due to the integration of a discontinuous reaction term over an element intersected by the wound edge. This will lead to a worse prediction of the wound edge velocity since the concentration on $W(t)$ will oscillate around the threshold concentration $\hat{c}$. The profile from the cut-cell method looks much smoother due to the appropriate integration of the continuous function over the newly formed elements along the interface.

From the more reliable approximations of the interface concentrations, the threshold condition for interface motion can be examined in a more accurate way. Therefore, this results into a more reliable prediction of the interface motion and wound healing kinetics. An example of the evolution of the wound area as a function of time for an elliptic wound is shown in Fig. 3. At the earliest stages, the concentration at the wound edge $W(t)$ has to increase from zero up to the threshold concentration $\hat{c}$. During this stage, the wound edge does not move yet, *that is* $W(t) = W(0)$. As soon as the interface concentration reaches $\hat{c}$, the wound starts to shrink. Further, it can be seen that the standard FEM exhibits a slower convergence behavior than the cut-cell method if the global grid is refined. We note that an adaptive mesh with refinement in the area near the edge was used in all the simulations. From these results, it can be concluded that the cut-cell method gives a significant improvement with respect to the standard finite-element method.

5 Conclusions

The model of epidermal wound healing is improved by using the cut-cell method. The interface points obtained with the cut-cell are used to adapt the triangulation to the wound edge position at each time step. The advantage of such an approach

is that the new subdivision is built without destroying the original FE mesh and altering the level set function.

The results obtained using cut-cell method are significantly better than previous ones. It can be seen clearly that the cut-cell method decreases the oscillatory behavior of the solution.

Acknowledgements Dr. Zemskov and Dr. Javierre acknowlegde the financial support from Senternovem (SHM 08733) and the Spanish Ministery of Science and Innovation (DPI 2009-07514).

References

1. Adam, J.A.: *A Simplified Model of Wound Healing (With Particular Reference to the Critical Size Defect)*, Math. and Comput. Modell., 30 (1999), 23–32
2. Javierre, E., Vermolen, F.J., Vuik, C., van der Zwaag, S.: *A Mathematical Analysis of Physiological and Morphological Aspects of Wound Closure*, J. Math. Biol., 59 (2009), 605–630
3. Osher, S., Sethian, J.A.: *Fronts Propagating with Curvature-Dependent Speed: Algorithms Based on Hamilton-Jacobi Formulations*, J. Comput. Phys., v.79 (1988), 12–49
4. Sethian, J.A.: *Fast Marching Method*, SIAM Rev., 41 (1999), 199–235
5. Sherratt, J.A., Murray, J.D: *Mathematical Analisis of a Basic Model for Epidermal Wound Healing*, J. Math. Biol., v.29 (1991), 389–404
6. Vermolen, F.J., Javierre, E.: *A Suite of Mathematical Models for Wound Contraction, Angiogenesis and Wound Closure*, in: Bioengineering Research of Chronic Wounds, Studies in Mechanobiology, Tissue Engineering and Biomaterials, Springer, Berlin (2009)
7. Vermolen, F.J., Javierre, E.: *On the Construction of Analytic Solutions for a Diffusion-Reaction Equation with a Discontinuous Switch Mechanism*, J. Comp. Appl. Math., 231(2) (2009), 983–1003
8. Vermolen, F.J., van Baaren, E., Adam, J.A.: *A Simplified Model for Growth Factor Induced Healing of Wounds*, Math. and Comput. Modell., 44 (2006), 887–898

Index