Partial Differential Equations

Computational Methods in Applied Sciences

Volume 16

Partial Differential Equations

Modeling and Numerical Simulation

Edited by

Roland Glowinski
University of Houston, TX, USA

and

Pekka Neittaanmäki
University of Jyväskylä, Finland

Editors
Roland Glowinski
Department of Mathematics
University of Houston
USA
roland@math.uh.edu

Pekka Neittaanmäki
Department of Mathematical Information
Technology
University of Jyväskylä
Finland
pn@mit.jyu.fi

Printed on acid-free paper

9 8 7 6 5 4 3 2 1

springer.com

Dedicated to Olivier Pironneau

Preface

For more than 250 years partial differential equations have been clearly the most important tool available to mankind in order to understand a large variety of phenomena, natural at first and then those originating from human activity and technological development. Mechanics, physics and their engineering applications were the first to benefit from the impact of partial differential equations on modeling and design, but a little less than a century ago the Schrödinger equation was the key opening the door to the application of partial differential equations to quantum chemistry, for small atomic and molecular systems at first, but then for systems of fast growing complexity. The place of partial differential equations in mathematics is a very particular one: initially, the partial differential equations modeling natural phenomena were derived by combining calculus with physical reasoning in order to express conservation laws and principles in partial differential equation form, leading to the wave equation, the heat equation, the equations of elasticity, the Euler and Navier–Stokes equations for fluids, the Maxwell equations of electro-magnetics, etc. It is in order to solve 'constructively' the heat equation that Fourier developed the series bearing his name in the early 19th century; Fourier series (and later integrals) have played (and still play) a fundamental role in both pure and applied mathematics, including many areas quite remote from partial differential equations.

On the other hand, several areas of mathematics such as differential geometry have benefited from their interactions with partial differential equations. The need for a better understanding of the properties of the solution of these equations has been a driver for both the mathematical investigation of their existence, uniqueness, regularity, and other properties, and the development of constructive methods to approximate these solutions. Numerical methods for the approximate solution of partial differential equations were invented, developed and applied to real life situations long before the advance (in the mid-forties) of digital computers; let us mention among these early methods: finite differences, Galerkin, Courant finite element, and a variety of iterative methods. However, the exponential growth in speed and memory of digital

computers has been at the origin of an explosive development of numerical mathematics, leading itself to applications of size and complexity unthinkable a not so long time ago.

There has been simultaneity in the progress achieved on both the theory and the numerics of partial differential equations, each feeding the other one: indeed, methods for proving the existence of solutions have lead to numerical methods for the actual computation of these solutions; on the other hand, conjectures on mathematical properties of solutions have been verified first computationally providing thus a justification for further analytical investigations. Applications of partial differential equations are essentially everywhere since to the areas mentioned above we have to add bio and health sciences, finance, image processing. (It is worth mentioning that today the term partial differential equations has to be taken in a broader sense than let say fifty years ago in order to include partial differential inequalities, which are of fundamental importance in, for example, the modeling of non-smooth phenomena.)

From the above comments, it is quite obvious that the "world of partial differential equations" is a very large and complex one, and, therefore, quite difficult to explore. Not surprisingly, the many aspects of partial differential equations (theory, modeling and computation) have motivated a huge number of publications (books, articles, conference proceedings, websites). Concerning books, most of them are necessarily specialized (unless elementary) with topics such as elliptic equations, parabolic equations, Navier–Stokes equations, Maxwell equations, to name some of the most popular ones. We think thus that there is a need for books on partial differential equations addressing at a reasonably advanced level a variety of topics. From a practical point of view, the diversity we mentioned above implies that such books have to be necessarily multi-authors. We think that the present volume is an answer to such a need since it contains the contributions of experts of international reputation on a quite diverse selection of topics all partial differential equation related, ranging from well-established ones in mechanics and physics to very recent ones in micro-electronics and finance. In all these contributions the emphasis has been on the modeling and computational aspects.

This volume is structured as follows: In Part I, discontinuous Galerkin and mixed finite element methods are applied to a variety of linear and nonlinear problems, including the Stokes problem from fluid mechanics and fully nonlinear elliptic equations of the Monge-Ampère type. Part II is dedicated to the numerical solution of linear and nonlinear hyperbolic problems. In Part III one discusses the solution by domain decomposition methods of scattering problems for wave models and of electronic structure related nonlinear variational problems. Part IV is devoted to various issues concerning the modeling and simulation of fluid mechanics phenomena involving free surfaces and moving boundaries. The finite difference solution of a problem from spectral geometry has also been included in this part. Part V is dedicated to inverse problems. Finally, in Part VI one addresses the parabolic variational inequalities based modeling and simulation of finance related processes.

Some of the issues discussed in this volume have been addressed at the international conference taking place in Helsinki during fall 2005 to honor Olivier Pironneau on the occasion of his 60th anniversary. Additional material has been included in order to broaden the scope of the volume.

Special acknowledgements are due to Marja-Leena Rantalainen from University of Jyväskylä for her most constructive role in the various stages of this project.

Houston and Jyväskylä   *Roland Glowinski*
Pekka Neittaanmäki

Contents

List of Contributors

Yves Achdou
UFR Mathématiques
Université Paris 7
Case 7012
FR-75251 Paris Cedex 05
France
achdou@math.jussieu.fr

Maxime Barrault
EDF R&D
1 avenue du Général de Gaulle
92141 Clamart Cedex
France
maxime.barrault@edf.fr

Guy Benctueux
EDF R&D
1 avenue du Général de Gaulle
92141 Clamart Cedex
France
guy.bencteux@edf.fr

Andrea Bonito
Department of Mathematics
University of Maryland
College Park, MD 20742-4015
USA
andrea.bonito@epfl.ch

Jean-Marc Brun
CEMAGREF/ITAP
FR-34095 Montpellier
France
jean-marc.brun@cemagref.fr

Alexandre Caboussat
Department of Mathematics
University of Houston
Houston, TX 77204-3008
USA
caboussat@math.uh.edu

Eric Cancès
CERMICS
Ecole Nationale des Ponts
et Chaussées
6 & 8 avenue Blaise Pascal
Cité Descartes
77455 Marne-La-Vallée Cedex 2
France
cances@cermics.enpc.fr

Edward J. Dean
University of Houston
Department of Mathematics
4800 Calhoun
Houston, TX 77004
USA
dean@math.uh.edu

Jean-Charles Gilbert
INRIA
Domaine de Voluceau-Roquencourt
BP 105
FR-78153 Le Chesnay Cedex
France
`Jean-Charles.Gilbert@inria.fr`

Vivette Girault
Laboratoire Jacques-Louis Lions
Université Pierre et Marie Curie
Case 187, 4 Place Jussieu
FR-75252 Paris Cedex 05
France
`girault@ann.jussieu.fr`

Roland Glowinski
University of Houston
Department of Mathematics
4800 Calhoun
Houston, TX 77004
USA
`roland@math.uh.edu`

William H. Hager
Department of Mathematics
University of Florida
Gainesville, FL 32611-8105
USA
`hager@math.ufl.edu`

Jian Hao
Department of Mathematics
University of Houston
Houston, TX 77204-3008
USA
`jianh@math.uh.edu`

Oubay Hassan
Civil and Computational
Engineering Centre
University of Wales-Swansea
Swansea SA2 8PP
Wales
UK
`O.Hassan@swansea.ac.uk`

Samuli Ikonen
Nordea Markets
FI-00020 Nordea
Finland
`Samuli.Ikonen@nordea.com`

Pierre-Marie Jacquart
Dassault Aviation
78, Quai Marcel Dassault
Cedex 300, Saint-Cloud 92552
France
`pierre-marie.jacquart@dassault-`
`aviation.fr`

Patrick Joly
INRIA
Domaine de Voluceau-Roquencourt
BP 105
FR-78153 Le Chesnay Cedex
France
`Patrick.Joly@inria.fr`

Yuri Kuznetsov
University of Houston
Department of Mathematics
4800 Calhoun
Houston, TX 77004
USA
`kuz@math.uh.edu`

Alexander Lapin
Kazan State University
Department of Computational
Mathematics and Cybernetics
18 Kremlyovskaya St.
Kazan 420008
Russia
`alapin@ksu.ru`

Serguei Lapin
University of Houston
Department of Mathematics
4800 Calhoun Rd
Houston, TX 77204
USA
`slapin@math.uh.edu`

Claude Le Bris
CERMICS
6&8 Avenue Blaise Pascal
Cité Descartes
FR-77455 Marne-la-Vallée Cedex 02
France
`lebris@cermics.enpc.fr`

Bertrand Maury
Laboratoire de Mathématiques
Université Paris-Sud
FR-91405 Orsay Cedex
France
`Bertrand.Maury@math.u-psud.fr`

Bijan Mohammadi
Mathematics and Modeling Institute
Université de Montpellier II
CC 51
FR-34095 Montpellier
France
`Bijan.Mohammadi@math.`
`univ-montp2.fr`

Ken Morgan
Civil and Computational
Engineering Centre
University of Wales-Swansea
Swansea SA2 8PP
Wales
UK
`K.Morgan@swansea.ac.uk`

Pekka Neittaanmäki
University of Jyväskylä
Department of Mathematical
Information Technology
P.O. Box 35 (Agora)
FI-40014, Jyväskylä
Finland
`pn@mit.jyu.fi`

Tsorng-Whay Pan
Department of Mathematics
University of Houston
Houston, TX 77204-3008
USA
`pan@math.uh.edu`

Jacques Periaux
University of Jyväskylä
Department of Mathematical
Information Technology
P.O. Box 35
FI-40014 University of Jyväskylä
Finland
`jperiaux@free.fr`

Marco Picasso
Institute of Analysis &
Scientific Computing
Ecole Polytechnique
Fédérale de Lausanne
1015 Lausanne
Switzerland
`marco.picasso@epfl.ch`

Jacques Rappaz
Institut d'Analyse et Calcul
Scientifique
Bat. de mathématiques, Station 8
Ecole Polytechnique Fédérale de
Lausanne
CH-1015 Lausanne
Switzerland
`jacques.rappaz@epfl.ch`

Doreen Rosenstrauch
The Texas Heart Institute & The
University of Texas Health Science
Center at Houston
Houston, TX 77030
USA
`Doreen.Rosenstrauch@uth.tmc.edu`

Richard Sanders
University of Houston
Department of Mathematics
4800 Calhoun
Houston, TX 77004
USA
`sanders@math.uh.edu`

Igor Sazanov
Civil and Computational
Engineering Centre
University of Wales-Swansea
Swansea SA2 8PP
Wales
UK
i.sazonov@swansea.ac.uk

Danny C. Sorensen
Rice University
Department of Computational
& Applied Mathematics
Houston, TX, 77251-1892
USA
sorensen@rice.edu

Allen M. Tesdall
Fields Institute
Toronto,ON M5T 3J1
and
Department of
Mathematics
University of Houston
Houston, TX 77204
USA
atesdall@fields.utoronto.ca

Dan Tiba
Romanian Academy

Institute of Mathematics
P.O. Box 1-764
RO-014700 Bucharest
Romania
dan.tiba@imar.ro

Jari Toivanen
Department of Mathematical
Information Technology
P.O. Box 35 (Agora)
FI-40014 University of Jyväskylä,
Finland
Jari.Toivanen@mit.jyu.fi

Nigel P. Weatherill
Civil and Computational
Engineering Centre
University of Wales-Swansea
Swansea SA2 8PP
Wales
UK
N.P.Weatherill@swansea.ac.uk

Mary F. Wheeler
Institute for Computational
Engineering & Sciences (ICES)
University of Texas at Austin
Austin, TX 78712
USA
mfw@ices.utexas.edu

Part I

Discontinuous Galerkin and Mixed Finite Element Methods

Discontinuous Galerkin Methods

Vivette Girault[1] and Mary F. Wheeler[2]

[1] Laboratoire Jacques-Louis Lions, Université Pierre et Marie Curie, Paris VI,
    FR-75252 Paris cedex 05, France `girault@ann.jussieu.fr`
[2] Institute for Computational Engineering and Sciences (ICES),
    University of Texas at Austin, Austin, TX 78712, USA `mfw@ices.utexas.edu`

Summary. In this article, we describe some simple and commonly used discontinuous Galerkin methods for elliptic, Stokes and convection-diffusion problems. We illustrate these methods by numerical experiments.

1 Introduction and Preliminaries

Discontinuous Galerkin (DG) methods use discontinuous piece-wise polynomial spaces to approximate the solution of PDE's in variational form. The concept of discontinuous space approximations was introduced in the early 70's, probably starting with the work of Nitsche [Nit71] in 1971 on domain decomposition and followed by a number of important contributions such as the work of Babuška and Zlamal [BZ73], Crouzeix and Raviart [CR73], Rachford and Wheeler [RW74], Oden and Wellford [OW75], Douglas and Dupont [DD76], Baker [Bak77], Wheeler [Whe78], Arnold [Arn79, Arn82] and Wheeler and Darlow [WD80]. Afterward, interest in DG methods for elliptic problems declined probably because computing facilities at that time were not sufficient to solve efficiently such schemes. By the end of the 90's, the thesis of Baumann [Bau97] and the spectacular increase in computing power, triggered a renewal of interest in discontinuous Galerkin methods for elliptic and parabolic problems. The work of Baumann was followed by numerous publications such as Oden, Babuška and Baumann [OBB98], Baumann and Oden [BO99], Rivière et al. [RWG99, RWG01], Rivière [Riv00], Arnold et al. [ABCM02], among many others. Research on DG methods is now a very active field.

In the meantime, discontinuous methods were applied extensively to hyperbolic problems [Bey94, BOP96]. One of the first is the upwind scheme introduced by Reed and Hill in their report [RH73] on neutron transport in 1973. The first numerical analysis was done by Lesaint and Raviart [LR74] in 1974 for the transport equation and by Girault and Raviart [GR79] in 1982 for the Navier–Stokes equations. We refer to the books by Pironneau [Pir89] and by Girault and Raviart [GR86] for a thorough study of this upwind scheme.

DG methods have many advantages over continuous methods. The discontinuity of their functions allow the use of non-conforming grids and variable degree of polynomials on adjacent elements. They are locally mass conservative on each element. Their mass matrix in time-dependent problems is block diagonal. They are particularly well-adapted to problems with discontinuous coefficients and can effectively capture discontinuities in the solution. They can impose essential boundary conditions weakly without the use of a multiplier and thus can be applied to domain decomposition without involving multipliers. They can be applied to incompressible elasticity problems. They can be easily coupled with continuous methods.

On the negative side, they are expensive, because they require many degrees of freedom and for this reason, efficient solvers using DG methods for elliptic or parabolic problems are still the object of research.

In this article, we present a survey on some simple DG methods for elliptic, flow and transport problems. We concentrate essentially on IIPG, SIPG, NIPG, OBB-DG and the upwind DG of Lesaint and Raviart. There is no space to present all DG methods and for this reason, we have left out the more sophisticated schemes such as Local Discontinuous Galerkin (LDG) methods for which we refer to Arnold et al. [ABCM02].

This article is organized as follows. In Section 2, we derive the equations on which number of DG methods are based when applied to simple model problems. Section 3 is devoted to the approximation of a Darcy flow. In Section 4, we describe some DG methods for an incompressible Stokes flow. A convection-diffusion equation is approximated in Section 5. Section 6 is devoted to numerical experiments performed at the Institute for Computational Engineering and Sciences, UT Austin.

In the sequel, we shall use the following functional notation. Let Ω be a domain in $\mathbb{R}^d$, where d is the dimension. For an integer $m \geq 1$, $H^m(\Omega)$ denotes the Sobolev space defined recursively by

$$H^m(\Omega) = \{v \in H^{m-1}(\Omega); \nabla v \in H^{m-1}(\Omega)^d\},$$

and we set

$$H^0(\Omega) = L^2(\Omega),$$

equipped with the norm

$$\|v\|_{L^2(\Omega)} = \left(\int_\Omega |v|^2 \, d\mathbf{x} \right)^{\frac{1}{2}}.$$

For fluid pressure and other variables defined up to an additive constant, it is useful in theory to fix the constant by imposing the zero mean value and, therefore, we use the space

$$L_0^2(\Omega) = \left\{ v \in L^2(\Omega); \int_\Omega v \, d\mathbf{x} = 0 \right\}.$$

2 An Elementary Derivation of Some Simple DG Methods

In this section, we use very simple examples to derive the equations that are at the basis of IIPG, SIPG, NIPG, OBB-DG methods and the upwind DG method of Lesaint–Raviart. In each example, we work out the equations on a plane domain Ω, with boundary $\partial\Omega$, partitioned into two non-overlapping subdomains Ω_1 and Ω_2 with interface Γ_{12}, and to fix ideas we assume that each subdomain has part of its boundary on $\partial\Omega$.

2.1 The General Idea for Elliptic Problems

Consider the Laplace equation with a homogeneous Dirichlet boundary condition in Ω and with data in $L^2(\Omega)$:

$$-\Delta u = f \quad \text{in } \Omega, \quad u = 0 \quad \text{on } \partial\Omega. \tag{1}$$

Let v be a test function that is sufficiently smooth in each Ω_i, but does not belong necessarily to $H^1(\Omega)$. If we multiply both sides of the first equation in (1) by v, apply Green's formula in each Ω_i, and assume that the solution u is smooth enough, we obtain:

$$\sum_{i=1}^{2} \left(\int_{\Omega_i} \nabla u \cdot \nabla v \, d\mathbf{x} - \int_{\partial\Omega_i} (\nabla u \cdot \mathbf{n}_i)|_{\Omega_i} \, v|_{\Omega_i} \, d\sigma \right) = \int_{\Omega} f v \, d\mathbf{x}, \tag{2}$$

where $\mathbf{n}_i$ denotes the unit normal to $\partial\Omega_i$, exterior to Ω_i. If u has sufficient smoothness, then the trace of $\nabla u \cdot \mathbf{n}_i$ on the interface has the same absolute value, but opposite signs, on Γ_{12} when coming either from Ω_1 or from Ω_2. As the change in sign comes from the normal vector, we choose once and for all the normal's orientation on Γ_{12}; for example, we choose the orientation of $\mathbf{n}_1$. Therefore, setting $\mathbf{n}_e = \mathbf{n}_1$, denoting by $\mathbf{n}_\Omega$ the exterior normal to $\partial\Omega$, denoting by $[v]_e$ and $\{v\}_e$ the jump and average of the trace of v across Γ_{12}:

$$[v]_e = v|_{\Omega_1} - v|_{\Omega_2}, \quad \{v\}_e = \frac{1}{2}\left(v|_{\Omega_1} + v|_{\Omega_2}\right),$$

and using the identity

$$\forall a_1, a_2, b_1, b_2 \in \mathbb{R}, \quad a_1 b_1 - a_2 b_2 = \frac{1}{2}\left[(a_1 + a_2)(b_1 - b_2) + (a_1 - a_2)(b_1 + b_2)\right],$$

(2) becomes

$$\sum_{i=1}^{2} \left(\int_{\Omega_i} \nabla u \cdot \nabla v \, d\mathbf{x} - \int_{\partial\Omega_i \setminus \Gamma_{12}} (\nabla u \cdot \mathbf{n}_\Omega) v \, d\sigma \right) - \int_{\Gamma_{12}} \{\nabla u \cdot \mathbf{n}_e\}_e [v]_e \, d\sigma$$

$$= \int_{\Omega} f v \, d\mathbf{x}. \tag{3}$$

The discontinuous Galerkin method called IIPG is based on (3). It uses the regularity of the normal derivative of u. If, in addition, we want to use the regularity of u and its zero boundary value, then we can add or subtract the following terms to the left-hand side of (3):

$$\int_{\Gamma_{12}} \{\nabla v \cdot \mathbf{n}_e\}_e [u]_e \, d\sigma, \quad \int_{\partial\Omega_i \backslash \Gamma_{12}} (\nabla v \cdot \mathbf{n}_\Omega) u \, d\sigma, \quad i = 1, 2.$$

Since these terms are zero, the resulting equation is equivalent to (3). The discontinuous Galerkin method called SIPG is based on subtraction of these terms:

$$\sum_{i=1}^{2} \left(\int_{\Omega_i} \nabla u \cdot \nabla v \, d\mathbf{x} - \int_{\partial\Omega_i \backslash \Gamma_{12}} \left((\nabla u \cdot \mathbf{n}_\Omega) v + (\nabla v \cdot \mathbf{n}_\Omega) u \right) d\sigma \right)$$
$$- \int_{\Gamma_{12}} \left(\{\nabla u \cdot \mathbf{n}_e\}_e [v]_e + \{\nabla v \cdot \mathbf{n}_e\}_e [u]_e \right) d\sigma = \int_{\Omega} f v \, d\mathbf{x}, \quad (4)$$

and the discontinuous Galerkin methods called NIPG and OBB-DG are based on addition of this term:

$$\sum_{i=1}^{2} \left(\int_{\Omega_i} \nabla u \cdot \nabla v \, d\mathbf{x} - \int_{\partial\Omega_i \backslash \Gamma_{12}} \left((\nabla u \cdot \mathbf{n}_\Omega) v - (\nabla v \cdot \mathbf{n}_\Omega) u \right) d\sigma \right)$$
$$- \int_{\Gamma_{12}} \left(\{\nabla u \cdot \mathbf{n}_e\}_e [v]_e - \{\nabla v \cdot \mathbf{n}_e\}_e [u]_e \right) d\sigma = \int_{\Omega} f v \, d\mathbf{x}. \quad (5)$$

In fact, the OBB-DG formulation is precisely (5).

Clearly, the contribution of the surface integrals to the left-hand side of (5) is anti-symmetric and hence the left-hand side of (5) is non-negative when $v = u$. The left-hand side of (4) is symmetric, but there is no reason why it should be non-negative and the left-hand side of (3) has no symmetry and no positivity. The left-hand side of (5) can be made positive when $v = u$ by adding to it the jump terms

$$\frac{1}{|\Gamma_{12}|} \int_{\Gamma_{12}} [u]_e [v]_e \, d\sigma + \sum_{i=1}^{2} \frac{1}{|\partial\Omega_i \backslash \Gamma_{12}|} \int_{\partial\Omega_i \backslash \Gamma_{12}} uv \, d\sigma,$$

where for any set S, $|S|$ denotes the measure of S. But, of course, this will not do for (3) and (4). However, considering that all these formulations will be applied to functions in finite-dimensional spaces, we expect to make (3) and (4) positive by incorporating into the jump terms adequate parameters. Thus we add

$$J_0(u, v) = \frac{\sigma_{12}}{|\Gamma_{12}|} \int_{\Gamma_{12}} [u]_e [v]_e \, d\sigma + \sum_{i=1}^{2} \frac{\sigma_i}{|\partial\Omega_i \backslash \Gamma_{12}|} \int_{\partial\Omega_i \backslash \Gamma_{12}} uv \, d\sigma, \quad (6)$$

Fig. 1. Jumps and averages: the jump on an interior edge is given by $[v] = v|_{E_1} - v|_{E_2}$ and on a boundary edge by $[v] = v|_{E_1}$; the averages are respectively given by $v = \frac{1}{2}(v|_{E_1} + v|_{E_2})$ and $v = v|_{E_1}$. The unit normal to γ_a is $\mathbf{n}_a$

.

where σ_{12} and σ_i are suitable non-negative parameters. Summing up, the IIPG, SIPG, NIPG and OBB-DG formulations read:

$$\sum_{i=1}^{2} \left(\int_{\Omega_i} \nabla u \cdot \nabla v \, d\mathbf{x} - \int_{\partial \Omega_i \setminus \Gamma_{12}} \left((\nabla u \cdot \mathbf{n}_\Omega)v + \varepsilon(\nabla v \cdot \mathbf{n}_\Omega)u \right) d\sigma \right)$$
$$- \int_{\Gamma_{12}} \left(\{\nabla u \cdot \mathbf{n}_e\}_e [v]_e + \varepsilon \{\nabla v \cdot \mathbf{n}_e\}_e [u]_e \right) d\sigma + J_0(u, v) = \int_{\Omega} fv \, d\mathbf{x}, \quad (7)$$

with $\varepsilon = 0$ for IIPG, $\varepsilon = 1$ for SIPG and $\varepsilon = -1$ for NIPG and OBB-DG, $\sigma_i = \sigma_{12} = 1$ for NIPG, $\sigma_i = \sigma_{12} = 0$ for OBB-DG and σ_i and σ_{12} are well chosen positive parameters for IIPG and SIPG. An example of jumps and average for a non-conforming mesh are shown in Figure 1.

Remark 1. The NIPG and OBB-DG formulations differ only on the presence or absence of jump terms. It turns out that in several cases, such as in Section 3, the jump terms are not necessary, but they can be added to enhance convergence. However, there are cases, such as in Section 4, where OBB-DG seems sub-optimal without jumps.

Remark 2. As the normal derivative of the solution has no jumps, it is also possible to add jumps involving this normal derivative (cf. [Dar80, WD80]):

$$|\Gamma_{12}| \int_{\Gamma_{12}} [\nabla u \cdot \mathbf{n}]_e [\nabla v \cdot \mathbf{n}]_e \, d\sigma.$$

The resulting equation is still equivalent to (3).

Finally, let us examine a Laplace equation with mixed non-homogeneous Dirichlet–Neumann boundary conditions. As an example, we replace (1) by

$$-\Delta u = f \text{ in } \Omega, \quad u = g_1 \text{ on } \partial \Omega_1 \setminus \Gamma_{12}, \quad \nabla u \cdot \mathbf{n}_\Omega = g_2 \text{ on } \partial \Omega_2 \setminus \Gamma_{12}. \quad (8)$$

In this case, we suppress from J_0 the boundary term on $\partial\Omega_2 \setminus \Gamma_{12}$:

$$J_0(u,v) = \frac{\sigma_{12}}{|\Gamma_{12}|} \int_{\Gamma_{12}} [u]_e [v]_e \, d\sigma + \frac{\sigma_1}{|\partial\Omega_1 \setminus \Gamma_{12}|} \int_{\partial\Omega_1 \setminus \Gamma_{12}} uv \, d\sigma, \qquad (9)$$

and the IIPG, SIPG, NIPG and OBB-DG formulations become:

$$\sum_{i=1}^{2} \int_{\Omega_i} \nabla u \cdot \nabla v \, d\mathbf{x} - \int_{\partial\Omega_1 \setminus \Gamma_{12}} \left((\nabla u \cdot \mathbf{n}_\Omega)v + \varepsilon(\nabla v \cdot \mathbf{n}_\Omega)u \right) d\sigma$$

$$- \int_{\Gamma_{12}} \left(\{\nabla u \cdot \mathbf{n}_e\}_e [v]_e + \varepsilon\{\nabla v \cdot \mathbf{n}_e\}_e [u]_e \right) d\sigma + J_0(u,v)$$

$$= \int_\Omega fv \, d\mathbf{x} + \int_{\partial\Omega_2 \setminus \Gamma_{12}} g_2 v \, d\sigma - \varepsilon \int_{\partial\Omega_1 \setminus \Gamma_{12}} g_1(\nabla v \cdot \mathbf{n}_\Omega) \, d\sigma$$

$$+ \frac{\sigma_1}{|\partial\Omega_1 \setminus \Gamma_{12}|} \int_{\partial\Omega_1 \setminus \Gamma_{12}} g_1 v \, d\sigma, \quad (10)$$
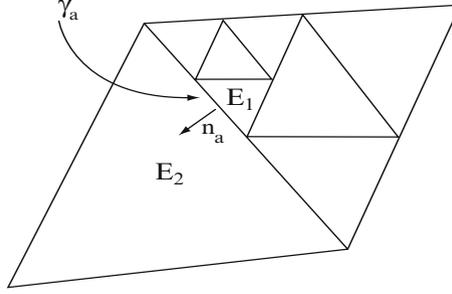
with the same values of ε, σ_1 and σ_{12} as in (7).

2.2 The General Idea for the Stokes Problem

Consider the incompressible Stokes problem in Ω with data $\mathbf{f}$ in $L^2(\Omega)^2$:

$$-\mu\Delta\mathbf{u} + \nabla p = \mathbf{f}, \quad \operatorname{div}\mathbf{u} = 0 \text{ in } \Omega, \quad \mathbf{u} = \mathbf{0} \text{ on } \partial\Omega, \qquad (11)$$

where the viscosity parameter μ is a given positive constant. This is a typical problem with a linear constraint (the zero divergence) and a Lagrange multiplier (the pressure p).

For treating the pressure term and divergence constraint, we take again a test function $\mathbf{v}$ that is not necessarily globally smooth, but has smooth components in each Ω_i, and assuming the pressure p is sufficiently smooth, we apply Green's formula in each Ω_i:

$$\int_\Omega (\nabla p) \cdot \mathbf{v} \, d\mathbf{x} = \sum_{i=1}^{2} \left(-\int_{\Omega_i} p \operatorname{div}\mathbf{v} \, d\mathbf{x} + \int_{\partial\Omega_i \setminus \Gamma_{12}} p(\mathbf{v} \cdot \mathbf{n}_\Omega) \, d\sigma \right)$$

$$+ \int_{\Gamma_{12}} \{p\}_e [\mathbf{v}]_e \cdot \mathbf{n}_e \, d\sigma. \quad (12)$$

We apply the same formula to the divergence constraint. Thus combining (12) with (7), we have the following IIPG, SIPG, NIPG and OBB-DG formulations for the Stokes problem (11):

$$\sum_{i=1}^{2} \mu \left(\int_{\Omega_i} \nabla \mathbf{u} : \nabla \mathbf{v} \, d\mathbf{x} - \int_{\partial\Omega_i \backslash \Gamma_{12}} \big((\nabla \mathbf{u} \cdot \mathbf{n}_\Omega) \mathbf{v} + \varepsilon (\nabla \mathbf{v} \cdot \mathbf{n}_\Omega) \mathbf{u} \big) \, d\sigma \right)$$

$$- \int_{\Gamma_{12}} \mu \big(\{\nabla \mathbf{u} \cdot \mathbf{n}_e\}_e [\mathbf{v}]_e + \varepsilon \{\nabla \mathbf{v} \cdot \mathbf{n}_e\}_e [\mathbf{u}]_e \big) \, d\sigma + \mu J_0(\mathbf{u}, \mathbf{v})$$

$$+ \sum_{i=1}^{2} \left(-\int_{\Omega_i} p \operatorname{div} \mathbf{v} \, d\mathbf{x} + \int_{\partial\Omega_i \backslash \Gamma_{12}} p (\mathbf{v} \cdot \mathbf{n}_\Omega) \, d\sigma \right) + \int_{\Gamma_{12}} \{p\}_e [\mathbf{v}]_e \cdot \mathbf{n}_e \, d\sigma$$

$$= \int_\Omega \mathbf{f} \cdot \mathbf{v} \, d\mathbf{x}, \tag{13}$$

$$\sum_{i=1}^{2} \left(\int_{\Omega_i} q \operatorname{div} \mathbf{u} \, d\mathbf{x} - \int_{\partial\Omega_i \backslash \Gamma_{12}} q (\mathbf{u} \cdot \mathbf{n}_\Omega) \, d\sigma \right) - \int_{\Gamma_{12}} \{q\}_e [\mathbf{u}]_e \cdot \mathbf{n}_e \, d\sigma = 0, \tag{14}$$

with the interpretation for the parameters ε and σ of the formula (7).

2.3 Upwinding in a Transport Problem: General Idea

Consider the simple transport problem in Ω:

$$c + \mathbf{u} \cdot \nabla c = f \quad \text{in } \Omega, \tag{15}$$

where f belongs to $L^2(\Omega)$ and $\mathbf{u}$ is a sufficiently smooth vector-valued function that satisfies

$$\operatorname{div} \mathbf{u} = 0 \text{ in } \Omega, \quad \mathbf{u} \cdot \mathbf{n}_\Omega = 0 \text{ on } \partial\Omega. \tag{16}$$

Recall the notation

$$\mathbf{u} \cdot \nabla c = \sum_{i=1}^{2} u_i \frac{\partial c}{\partial x_i},$$

and note that when the functions involved are sufficiently smooth, Green's formula and (16) yield

$$\int_\Omega (\mathbf{u} \cdot \nabla c) c \, d\mathbf{x} = 0. \tag{17}$$

For the applications we have in mind, let us assume that c is sufficiently smooth in each Ω_i, but is not necessarily in $H^1(\Omega)$. Then, we must give a meaning to the product $\mathbf{u} \cdot \nabla c$. From the following identity and the fact that the divergence of $\mathbf{u}$ is zero:

$$\operatorname{div}(c\mathbf{u}) = c(\operatorname{div} \mathbf{u}) + \mathbf{u} \cdot \nabla c = \mathbf{u} \cdot \nabla c,$$

and we derive for any smooth function φ with compact support in Ω

$$\langle \mathbf{u} \cdot \nabla c, \varphi \rangle = \langle \operatorname{div}(c\mathbf{u}), \varphi \rangle = -\langle c\mathbf{u}, \nabla \varphi \rangle = -\int_\Omega (c\mathbf{u}) \cdot \nabla \varphi \, d\mathbf{x}$$

$$= -\sum_{i=1}^{2} \int_{\Omega_i} (c\mathbf{u}) \cdot \nabla \varphi \, d\mathbf{x}. \tag{18}$$

We use the last equality to define $\mathbf{u} \cdot \nabla c$ in the sense of distributions.

Now, we wish to extend this definition to functions $\mathbf{u}$ and φ that are not necessarily smooth. Then, we take again a test function v that is sufficiently smooth in each Ω_i, but may not be in $H^1(\Omega)$. Applying Green's formula to the last equality in (18) in each Ω_i and using the fact that $\mathbf{u}$ has zero divergence, we define:

$$\int_\Omega (\mathbf{u} \cdot \nabla c) v \, d\mathbf{x} := \sum_{i=1}^2 \left(\int_{\Omega_i} (\mathbf{u} \cdot \nabla c) v \, d\mathbf{x} - \int_{\partial\Omega_i} c(\mathbf{u} \cdot \mathbf{n}) v \, d\sigma \right). \qquad (19)$$

In order to introduce an upwinding into this formula, we consider each Ω_i and the portion of its boundary where the flow driven by $\mathbf{u}$ enters Ω_i, i.e., where $\{\mathbf{u}\} \cdot \mathbf{n}_i < 0$. We set

$$(\partial\Omega_i)_- = \{\mathbf{x} \in \partial\Omega_i; \{\mathbf{u}\} \cdot \mathbf{n}_i(\mathbf{x}) < 0\}. \qquad (20)$$

Then we replace (19) by

$$\int_\Omega (\mathbf{u} \cdot \nabla c) v \, d\mathbf{x} := \sum_{i=1}^2 \left(\int_{\Omega_i} (\mathbf{u} \cdot \nabla c) v \, d\mathbf{x} - \int_{(\partial\Omega_i)_-} \{\mathbf{u}\} \cdot \mathbf{n}_i (c^{\text{int}} - c^{\text{ext}}) v^{\text{int}} \, d\sigma \right), \qquad (21)$$

where the superscript int (resp. ext) refers to the interior (resp. exterior) trace of the function in Ω_i, and on the part of $(\partial\Omega_i)_-$ that lies on $\partial\Omega$, $c^{\text{ext}} = 0$ and $\{\mathbf{u}\} = \mathbf{u}$. This is a straightforward extension of the Lesaint–Raviart upwind scheme.

Finally, we wish to extend (21) to the case where $\mathbf{u}$ satisfies (14) instead of (16), while preserving some property analogous to (17). Keeping in mind the identity:

$$\int_\Omega (\mathbf{u} \cdot \nabla c) c \, d\mathbf{x} + \frac{1}{2} \int_\Omega (\text{div } \mathbf{u}) c^2 \, d\mathbf{x} - \frac{1}{2} \int_{\partial\Omega} (\mathbf{u} \cdot \mathbf{n}) c^2 \, d\sigma = 0, \qquad (22)$$

that holds if c and $\mathbf{u}$ are sufficiently smooth, we replace (21) by:

$$\int_\Omega (\mathbf{u} \cdot \nabla c) v \, d\mathbf{x} := \sum_{i=1}^2 \left(\int_{\Omega_i} \left(\mathbf{u} \cdot \nabla c + \frac{1}{2}(\text{div } \mathbf{u}) c \right) v \, d\mathbf{x} \right.$$

$$\left. - \frac{1}{2} \int_{\partial\Omega_i \setminus \Gamma_{12}} (\mathbf{u} \cdot \mathbf{n}_\Omega) cv \, d\sigma - \int_{(\partial\Omega_i)_-} \{\mathbf{u}\} \cdot \mathbf{n}_i (c^{\text{int}} - c^{\text{ext}}) v^{\text{int}} \, d\sigma \right)$$

$$- \frac{1}{2} \int_{\Gamma_{12}} [\mathbf{u}]_e \cdot \mathbf{n}_e \{cv\}_e \, d\sigma. \qquad (23)$$

This is the upwind formulation proposed and analyzed by Rivière et al. [GRW05].

3 DG Approximation of an Elliptic Problem

Let Ω be a polygon in dimension $d = 2$ or a Lipschitz polyhedron in dimension $d = 3$, with boundary $\partial\Omega$ partitioned into two disjoint parts: $\partial\Omega = \Gamma_D \cup \Gamma_N$, with polygonal boundaries if $d = 3$. For simplicity, we assume that $|\Gamma_D|$ is positive. Consider the continuity equation for Darcy flow in pressure form in Ω:

$$-\operatorname{div}(\mathbf{K}\nabla p) = f, \quad \text{in } \Omega, \tag{24}$$

$$p = g_1, \quad \text{on } \Gamma_D, \tag{25}$$

$$\mathbf{K}\nabla p \cdot \mathbf{n}_\Omega = g_2, \quad \text{on } \Gamma_N, \tag{26}$$

where $\mathbf{n}_\Omega$ is the unit normal vector to $\partial\Omega$, exterior to Ω, and the permeability $\mathbf{K}$ is a uniformly bounded, positive definite symmetric tensor, that is allowed to vary in space. For $f \in L^2(\Omega)$, $g_1 \in H^{1/2}(\Gamma_D)$ and $g_2 \in L^2(\Gamma_N)$, system (24)–(26) has a unique solution $p \in H^1(\Omega)$ and we assume that p is sufficiently regular to guarantee the consistency of the schemes below.

Let $\mathcal{E}_h$ be a regular family of triangulations of $\overline{\Omega}$ consisting of triangles (or tetrahedra if $d = 3$) E of maximum diameter h, and such that no face or side of ∂E intersects both Γ_D and Γ_N. It is regular in the sense of Ciarlet [Cia91]: There exists a constant $\gamma > 0$, independent of h, such that

$$\forall E \in \mathcal{E}_h, \quad \frac{h_E}{\varrho_E} = \gamma_E \leq \gamma, \tag{27}$$

where h_E denotes the diameter of E (bounded above by h) and ϱ_E denotes the diameter of the ball inscribed in E.

To simplify the discussion, we assume that $\mathcal{E}_h$ is conforming, but most results in this section remain valid for non-conforming grids as well as for quadrilateral (or hexahedral if $d = 3$) grids. We denote by Γ_h the set of all interior edges (or faces if $d = 3$) of $\mathcal{E}_h$ and by $\Gamma_{h,D}$ (resp. $\Gamma_{h,N}$) the set of all edges or faces of $\mathcal{E}_h$ that lie on Γ_D (resp. Γ_N). The elements E of $\mathcal{E}_h$ are numbered and denoted by E_i, say for $1 \leq i \leq P_h$. With any edge or face e of Γ_h shared by E_i and E_j with $i < j$, we associate once and for all the unit normal vector $\mathbf{n}_e$ directed from E_i to E_j and we define the jump $[\varphi]_e$ and average $\{\varphi\}_e$ of a function φ by:

$$[\varphi]_e = \varphi|_{E_i} - \varphi|_{E_j}, \quad \{\varphi\}_e = \frac{1}{2}(\varphi|_{E_i} + \varphi|_{E_j}).$$

If $e \subset \partial\Omega$, then $\mathbf{n}_e = \mathbf{n}_\Omega$ and the jump and average of φ coincide with the trace of φ.

Considering the differential operator in (24), we define the "discontinuous" space:

$$H^1(\mathcal{E}_h) = \{v \in L^2(\Omega); \forall E \in \mathcal{E}_h, v|_E \in H^1(E)\},$$

equipped with the "broken" semi-norm

$$\|\mathbf{K}^{\frac{1}{2}}\nabla v\|_{L^2(\mathcal{E}_h)} = \left[\sum_{E\in\mathcal{E}_h}\|\mathbf{K}^{\frac{1}{2}}\nabla v\|_{L^2(E)}^2\right]^{\frac{1}{2}}, \tag{28}$$

and norm (for which it is a Hilbert space)

$$\|v\|_{H^1(\mathcal{E}_h)} = \left(\|v\|_{L^2(\Omega)}^2 + \|\mathbf{K}^{\frac{1}{2}}\nabla v\|_{L^2(\mathcal{E}_h)}^2\right)^{\frac{1}{2}}.$$

In view of (9), we define the jump bilinear form

$$J_0(u,v) = \sum_{e\in\Gamma_h\cup\Gamma_{h,D}}\frac{\sigma_e}{h_e}\int_e [u]_e[v]_e\,d\sigma, \tag{29}$$

where h_e denotes the diameter of e, and each σ_e is a suitable non-negative parameter. It is convenient to define also the mesh-dependent semi-norm

$$\|v\|_{H^1(\mathcal{E}_h)} = \left(\|\mathbf{K}^{\frac{1}{2}}\nabla v\|_{L^2(\mathcal{E}_h)}^2 + J_0(v,v)\right)^{\frac{1}{2}}. \tag{30}$$

Now, we choose an integer $k \geq 1$ and we discretize $H^1(\mathcal{E}_h)$ with the finite element space

$$X_h = \{v\in L^2(\Omega) : \forall E\in\mathcal{E}_h,\ v|_E\in\mathbb{P}_k(E)\}. \tag{31}$$

It is possible to let k vary from one element to the next, but for simplicity we keep the same k. Then, keeping in mind (10), we discretize (24)–(26) by the following discrete system: Find $p_h\in X_h$ such that for all $q_h\in X_h$,

$$\sum_{E\in\mathcal{E}_h}\int_E \mathbf{K}\nabla p_h\cdot\nabla q_h\,d\mathbf{x}$$

$$-\sum_{e\in\Gamma_h\cup\Gamma_{h,D}}\int_e\left(\{\mathbf{K}\nabla p_h\cdot\mathbf{n}_e\}_e[q_h]_e + \varepsilon\{\mathbf{K}\nabla q_h\cdot\mathbf{n}_e\}_e[p_h]_e\right)d\sigma + J_0(p_h,q_h)$$

$$= \int_\Omega fq_h\,d\mathbf{x} + \int_{\Gamma_N} g_2 q_h\,d\sigma - \varepsilon\sum_{e\in\Gamma_{h,D}}\int_e g_1(\mathbf{K}\nabla q_h\cdot\mathbf{n}_\Omega)\,d\sigma$$

$$+\sum_{e\in\Gamma_{h,D}}\frac{\sigma_e}{h_e}\int_e g_1 q_h\,d\sigma, \tag{32}$$

with $\varepsilon = 1$ for SIPG, $\varepsilon = 0$ for IIPG and $\varepsilon = -1$ for NIPG and OBB-DG; and for each e, $\sigma_e = 1$ for NIPG, $\sigma_e = 0$ for OBB-DG and again σ_e is a well chosen positive parameter for IIPG and SIPG.

Remark 3. Let E be an element of $\mathcal{E}_h$ with no edge (or face) e on $\partial\Omega$. Taking $q_h = \chi_E$, the characteristic function of E in (32), we easily derive the discrete mass balance relation where $\mathbf{n}_E$ denotes the unit normal exterior to E:

$$-\sum_{e\in\partial E}\int_e\{\mathbf{K}\nabla p_h\}\cdot\mathbf{n}_E\,d\sigma + \sum_{e\in\partial E}\frac{\sigma_e}{h_e}\int_e(p_h^{\text{int}} - p_h^{\text{ext}})\,d\sigma = \int_E f\,d\mathbf{x}.$$

3.1 Numerical Analysis

To simplify the discussion, we introduce the bilinear form defined for any pair of functions p and q in $X_h + H^s(\Omega)$ with $s > \frac{3}{2}$ (so that the integrals over e are well-defined):

$$a_h(p,q) = \sum_{E \in \mathcal{E}_h} \int_E \mathbf{K}\nabla p \cdot \nabla q \, d\mathbf{x}$$

$$- \sum_{e \in \Gamma_h \cup \Gamma_{h,D}} \int_e \left(\{\mathbf{K}\nabla p \cdot \mathbf{n}_e\}_e [q]_e + \varepsilon \{\mathbf{K}\nabla q \cdot \mathbf{n}_e\}_e [p]_e \right) d\sigma. \quad (33)$$

Clearly, for NIPG,

$$a_h(q_h, q_h) + J_0(q_h, q_h) = \|q_h\|_{H^1(\mathcal{E}_h)}^2, \quad (34)$$

and, therefore, (32) has a unique solution. For IIPG and SIPG [Whe78, DSW04], an argument on finite-dimensional spaces (cf. [GSWY]) shows that for each e there exists a constant c_e, independent of h, but depending on k, the regularity constant γ of (27) and the maximum and minimum eigenvalues of $\mathbf{K}$ on the elements adjacent to e, such that for all p_h and q_h in X_h

$$\left| \sum_{e \in \Gamma_h \cup \Gamma_{h,D}} \int_e \{\mathbf{K}\nabla p_h \cdot \mathbf{n}_e\}_e [q_h]_e \, d\sigma \right|$$

$$\leq \|\mathbf{K}^{\frac{1}{2}} \nabla p_h\|_{L^2(\mathcal{E}_h)} \left(\sum_{e \in \Gamma_h \cup \Gamma_{h,D}} \frac{c_e}{h_e} \|[q_h]\|_{L^2(e)}^2 \right)^{\frac{1}{2}}. \quad (35)$$

The assumptions on $\mathbf{K}$ imply that the constants c_e can be bounded above independently of h and e and, therefore, applying Young's inequality, we can choose constants σ_e, uniformly bounded above and below with respect to h:

$$\forall e \in \Gamma_h \cup \Gamma_{h,D}, \quad 1 \leq \sigma_0 \leq \sigma_e \leq \sigma_m, \quad (36)$$

such that (for instance)

$$\left| \sum_{e \in \Gamma_h \cup \Gamma_{h,D}} \int_e \{\mathbf{K}\nabla q_h \cdot \mathbf{n}_e\}_e [q_h]_e \, d\sigma \right| \leq \frac{1}{4} \|q_h\|_{H^1(\mathcal{E}_h)}^2. \quad (37)$$

With this choice of penalty parameters σ_e, the system (32) for IIPG and SIPG has a unique solution. Furthermore, there exist two positive constants α and M, independent of h such that for all p_h and q_h in X_h

$$|a_h(p_h, q_h)| + |J_0(p_h, q_h)| \leq M \|p_h\|_{H^1(\mathcal{E}_h)} \|q_h\|_{H^1(\mathcal{E}_h)},$$

$$a_h(q_h, q_h) + J_0(q_h, q_h) \geq \alpha \|q_h\|_{H^1(\mathcal{E}_h)}^2. \quad (38)$$

This analysis cannot be applied to establish the solvability of OBB-DG, because the term J_0 is missing. If $k \geq 2$, one can show directly for OBB-DG that (32) has a unique solution cf. [RWG01], but the second part of (38) does not hold. When $k = 1$, there is a counter-example that shows that (32) is not well-posed (cf. [OBB98]). For this reason, OBB-DG is only applied when $k \geq 2$.

With the above choice of penalty parameters σ_e, a standard error analysis allows to prove optimal a priori error estimates in the norm $\|\cdot\|_{H^1(\mathcal{E}_h)}$ for IIPG, SIPG and NIPG: if the exact solution p of (24)–(26) belongs to $H^{k+1}(\Omega)$, then for the three methods

$$\|p_h - p\|_{H^1(\mathcal{E}_h)} = \mathrm{O}(h^k).$$

The same result holds for OBB-DG, but the proof is more subtle. The difficulty lies in estimating the term

$$T = \sum_{e \in \Gamma_h \cup \Gamma_{h,D}} \int_e \{\mathbf{K}\nabla(p - R_h p) \cdot \mathbf{n}_e\}_e [q_h]_e \, d\sigma,$$

where R_h is an interpolation operator in X_h and $q_h \in X_h$ is an arbitrary test function. If we had jumps, we would write as in the cases of IIPG, SIPG and NIPG:

$$|T| \leq \sum_{e \in \Gamma_h \cup \Gamma_{h,D}} \left(\frac{h_e}{\sigma_e}\right)^{\frac{1}{2}} \|\{\mathbf{K}\nabla(p - R_h p) \cdot \mathbf{n}_e\}_e\|_{L^2(e)} \left(\frac{\sigma_e}{h_e}\right)^{\frac{1}{2}} \|[q_h]_e\|_{L^2(e)}.$$

With a standard interpolation operator, owing to the factor $h_e^{\frac{1}{2}}$, the term

$$\left(\frac{h_e}{\sigma_e}\right)^{\frac{1}{2}} \|\{\mathbf{K}\nabla(p - R_h p) \cdot \mathbf{n}_e\}_e\|_{L^2(e)} = \mathrm{O}(h^k).$$

Here we have no jumps and the only way in which we can recover the factor $h_e^{\frac{1}{2}}$ is by constructing an interpolation operator R_h such that

$$\int_e \{\mathbf{K}\nabla(p - R_h p) \cdot \mathbf{n}_e\}_e \, d\sigma = 0.$$

If this is the case, then we can write

$$T = \sum_{e \in \Gamma_h \cup \Gamma_{h,D}} \int_e \{\mathbf{K}\nabla(p - R_h p) \cdot \mathbf{n}_e\}_e ([q_h]_e - c_e) \, d\sigma,$$

where the number c_e is chosen so that

$$\|[q_h]_e - c_e\|_{L^2(e)} \leq C\left(h_{E_i}^{\frac{1}{2}} \|\nabla q_h\|_{L^2(E_i)} + h_{E_j}^{\frac{1}{2}} \|\nabla q_h\|_{L^2(E_j)}\right),$$

and E_i and E_j are the elements adjacent to e. This interpolation operator is constructed in [RWG01], for $k \geq 2$. When $k = 1$, there are not enough degrees of freedom for its construction.

When the solution of (24)–(26) belongs to $H^2(\Omega)$ for all sufficiently smooth data (this holds, for example, when $\mathbf{K}$ and g_1 are sufficiently smooth and Γ_D is the whole boundary), then a duality argument shows that the error for SIPG in the L^2 norm has a higher order:

$$\|p_h - p\|_{L^2(\Omega)} = \mathrm{O}(h^{k+1}). \tag{39}$$

More generally, if there exists $s \in \left]\frac{3}{2}, 1\right]$ such that the solution of (24)–(26) belongs to $H^{1+s}(\Omega)$ for all correspondingly smooth data then (cf. [RWG01])

$$\|p_h - p\|_{L^2(\Omega)} = \mathrm{O}(h^{k+s}).$$

This result follows from the symmetry of a_h. For the other methods, which are not symmetric, the same duality argument (cf. [RWG01]) does not yield any increase in order, namely all we have is

$$\|p_h - p\|_{L^2(\Omega)} = \mathrm{O}(h^k). \tag{40}$$

Nevertheless, numerical results for NIPG and OBB-DG tend to prove that (39) holds if k is an odd integer, but so far we have no proof of this result.

Remark 4. The choice of penalty parameters for IIPG and SIPG is not straightforward. If chosen too small, the stability properties in (38) may be lost. But if chosen too large, the matrix of system (32) may become ill-conditioned.

Remark 5. One cannot prove basic inequalities on the functions of X_h, such as Poincaré's Inequality, without adding jumps to the broken norm; i.e., the gradients in each element are not sufficient to control the L^2 norm. With jumps, one can prove Poincaré–Friedrich's inequalities, Sobolev inequalities, Korn's inequalities and trace inequalities. For Poincaré–Friedrich's inequalities and Korn's inequalities, we refer to the very good contributions of Brenner [Bre03, Bre04]. The Sobolev and trace inequalities can be derived by using similar arguments (cf. [GRW05]). Note that, by virtue of Poincaré's Inequality, (40) can be established directly for IIPG, SIPG and NIPG without having to assume that the solution of (24)–(26) has extra smoothness for all smooth data.

4 DG Approximation of an Incompressible Stokes Problem

Let us revert to the problem (11) on a connected polygonal or polyhedral domain:

$$-\mu \Delta \mathbf{u} + \nabla p = \mathbf{f}, \quad \operatorname{div} \mathbf{u} = 0 \text{ in } \Omega, \quad \mathbf{u} = \mathbf{0} \text{ on } \partial\Omega.$$

For a given force $\mathbf{f} \in L^2(\Omega)^d$, this problem has a unique solution $\mathbf{u} \in H_0^1(\Omega)^d$ and $p \in L_0^2(\Omega)$ (cf., for instance, [Tem79, GR86]). In fact, the solution is more regular and the scheme below is consistent (cf. [Gri85, Dau89]).

In view of the operator and boundary condition in (11), the relevant spaces here are $H^1(\mathcal{E}_h)^d$ and $L_0^2(\Omega)$, and the set $\Gamma_{h,N}$ is empty. The definition of J_0 is extended straightforwardly to vectors and the permeability tensor is replaced by the identity multiplied by the viscosity. Thus, the semi-norms (28) and (30) are replaced by

$$\|\nabla \mathbf{v}\|_{L^2(\mathcal{E}_h)} = \left[\sum_{E \in \mathcal{E}_h} \|\nabla \mathbf{v}\|_{L^2(E)}^2 \right]^{\frac{1}{2}}, \tag{41}$$

$$\|\mathbf{v}\|_{H^1(\mathcal{E}_h)} = \mu^{\frac{1}{2}} \left(\|\nabla \mathbf{v}\|_{L^2(\mathcal{E}_h)}^2 + J_0(\mathbf{v}, \mathbf{v}) \right)^{\frac{1}{2}}. \tag{42}$$

Again, we choose an integer $k \geq 1$ and we discretize $H^1(\mathcal{E}_h)^d$ and $L_0^2(\Omega)$ with the finite element spaces

$$\mathbf{X}_h = \{\mathbf{v} \in L^2(\Omega)^d : \forall E \in \mathcal{E}_h, \ \mathbf{v}|_E \in \mathbb{P}_k(E)^d\}, \tag{43}$$

$$M_h = \{q \in L_0^2(\Omega) : \forall E \in \mathcal{E}_h, \ q|_E \in \mathbb{P}_{k-1}(E)\}. \tag{44}$$

The choice $\mathbb{P}_{k-1}$ for the discrete pressure, one degree less than the velocity, is suggested by the fact that L^2 is the natural norm for the pressure. Keeping in mind (13) and (14), we discretize (11) by the following discrete system: Find $\mathbf{u}_h \in \mathbf{X}_h$ and $p_h \in M_h$ satisfying for all $\mathbf{v}_h \in \mathbf{X}_h$ and $q_h \in M_h$:

$$\mu \sum_{E \in \mathcal{E}_h} \int_E \nabla \mathbf{u}_h : \nabla \mathbf{v}_h \, d\mathbf{x}$$

$$- \mu \sum_{e \in \Gamma_h \cup \partial\Omega} \int_e \left(\{\nabla \mathbf{u}_h \cdot \mathbf{n}_e\}_e [\mathbf{v}_h]_e + \varepsilon \{\nabla \mathbf{v}_h \cdot \mathbf{n}_e\}_e [\mathbf{u}_h]_e \right) d\sigma + \mu J_0(\mathbf{u}_h, \mathbf{v}_h)$$

$$- \sum_{E \in \mathcal{E}_h} \int_E p_h \operatorname{div} \mathbf{v}_h \, d\mathbf{x} + \sum_{e \in \Gamma_h \cup \partial\Omega} \int_e \{p_h\}_e [\mathbf{v}_h]_e \cdot \mathbf{n}_e \, d\sigma = \int_\Omega \mathbf{f} \cdot \mathbf{v}_h \, d\mathbf{x},$$
$$\tag{45}$$

$$\sum_{E \in \mathcal{E}_h} \int_E q_h \operatorname{div} \mathbf{u}_h \, d\mathbf{x} - \sum_{e \in \Gamma_h \cup \partial\Omega} \int_e \{q_h\}_e [\mathbf{u}_h]_e \cdot \mathbf{n}_e \, d\sigma = 0, \tag{46}$$

with the interpretation for the parameters ε and σ of formula (7).

Let a_h and b_h denote the bilinear forms

$$a_h(\mathbf{u}, \mathbf{v}) = \mu \sum_{E \in \mathcal{E}_h} \int_E \nabla \mathbf{u} : \nabla \mathbf{v} \, d\mathbf{x}$$

$$- \mu \sum_{e \in \Gamma_h \cup \partial \Omega} \int_e \left(\{\nabla \mathbf{u} \cdot \mathbf{n}_e\}_e [\mathbf{v}]_e + \varepsilon \{\nabla \mathbf{v} \cdot \mathbf{n}_e\}_e [\mathbf{u}]_e \right) d\sigma, \quad (47)$$

$$b_h(\mathbf{v}, q) = \sum_{E \in \mathcal{E}_h} \int_E q \operatorname{div} \mathbf{v} \, d\mathbf{x} - \sum_{e \in \Gamma_h \cup \partial \Omega} \int_e \{q\}_e [\mathbf{v}]_e \cdot \mathbf{n}_e \, d\sigma. \quad (48)$$

Clearly, the properties of a_h listed in the previous section are valid here and, therefore, existence and uniqueness of $\mathbf{u}_h$ hold for IIPG and SIPG if the penalty parameters σ_e are well-chosen; they hold unconditionally for NIPG and they hold for OBB-DG if $k \geq 2$. But existence and uniqueness of p_h is not straightforward because it is the consequence of the uniform "inf-sup" condition, that is now a standard tool in studying problems with a linear constraint (cf. [Bab73, Bre74]): There is a constant $\beta^* > 0$ independent of h such that

$$\inf_{q_h \in M_h} \sup_{\mathbf{v}_h \in \mathbf{X}_h} \frac{b_h(\mathbf{v}_h, q_h)}{\|\mathbf{v}_h\|_{H^1(\mathcal{E}_h)} \|q_h\|_{L^2(\Omega)}} \geq \beta^*. \quad (49)$$

By using the Raviart–Thomas interpolation operator (cf. [RT75, GR86]), we can readily show that (49) holds for IIPG, SIPG, NIPG and OBB-DG (cf., for instance, [SST03]). Hence the four schemes have a unique solution. However, in order to derive optimal error estimates, we have to bound the term $b_h(\mathbf{v}_h, p - \rho_h p)$, where ρ_h is a suitable approximation operator, for instance, a local L^2 projection on each E, and $\mathbf{v}_h$ is an arbitrary test function in $\mathbf{X}_h$. It is easy to prove that if $p \in H^k(\mathcal{E}_h)$ then

$$|b_h(\mathbf{v}_h, p - \rho_h p)| \leq C h^k \left(\sum_{e \in \Gamma_h \cup \partial \Omega} \frac{1}{h_e} \|[\mathbf{v}_h]\|_{L^2(e)}^2 + \|\nabla \mathbf{v}_h\|_{L^2(\mathcal{E}_h)}^2 \right)^{\frac{1}{2}}.$$

As J_0 is zero for OBB-DG, we cannot obtain a good estimate for this method: it does not seem to be well-adapted to this formulation of the Stokes problem.

On the other hand, we can obtain optimal error estimates for IIPG, SIPG, NIPG: if the exact solution $(\mathbf{u}, p)$ of the problem (11) belongs to $H^{k+1}(\Omega)^d \times H^k(\Omega)$, then for the three methods

$$\|\mathbf{u}_h - \mathbf{u}\|_{H^1(\mathcal{E}_h)} + \|p_h - p\|_{L^2(\Omega)} = \mathrm{O}(h^k). \quad (50)$$

Remark 6. Let E be an element as in Remark 3. Taking first $q_h = \chi_E$ in (46) and next the i-th component of $\mathbf{v}_h$, $v_{h,i} = \chi_E$ in (45), we obtain the discrete mass balance relations:

$$\int_E \operatorname{div} \mathbf{u}_h \, d\mathbf{x} - \frac{1}{2} \sum_{e \in \partial E} \int_e (\mathbf{u}_h^{\mathrm{int}} - \mathbf{u}_h^{\mathrm{ext}}) \cdot \mathbf{n}_E \, d\sigma = 0,$$

$$-\mu \sum_{e \in \partial E} \int_e \{\nabla u_{h,i}\} \cdot \mathbf{n}_E \, d\sigma + \mu \sum_{e \in \partial E} \frac{\sigma_e}{h_e} \int_e (u_{h,i}^{\mathrm{int}} - u_{h,i}^{\mathrm{ext}}) \, d\sigma = \int_E f_i \, d\mathbf{x}.$$

5 DG Approximation of a Convection-Diffusion Equation

Consider the convection-diffusion equation combining (24) and (15) in the domain Ω of the previous sections:

$$-\operatorname{div}(\mathbf{K}\nabla c) + \mathbf{u}\cdot\nabla c = f, \quad \text{in } \Omega, \tag{51}$$

$$\mathbf{K}\nabla c\cdot\mathbf{n}_\Omega = 0, \quad \text{on } \partial\Omega, \tag{52}$$

where f belongs to $L_0^2(\Omega)$, the tensor $\mathbf{K}$ satisfies the assumptions listed in Section 3 and $\mathbf{u}$ satisfies (16):

$$\operatorname{div}\mathbf{u} = 0 \text{ in } \Omega, \quad \mathbf{u}\cdot\mathbf{n}_\Omega = 0 \text{ on } \partial\Omega.$$

This problem has a solution $c \in H^1(\Omega)$, unique up to an additive constant under mild restrictions on the velocity $\mathbf{u}$, for instance, when $\mathbf{u}$ belongs to $H^1(\Omega)^d$. We propose to discretize it with a DG method when $\mathbf{u}$ is replaced by the solution $\mathbf{u}_h \in \mathbf{X}_h$ of a flow problem that satisfies $b_h(\mathbf{u}_h, q_h) = 0$ for all $q_h \in M_h$:

$$\sum_{E\in\mathcal{E}_h}\int_E q_h \operatorname{div}\mathbf{u}_h\,d\mathbf{x} - \sum_{e\in\Gamma_h\cup\partial\Omega}\int_e \{q_h\}_e[\mathbf{u}_h]_e\cdot\mathbf{n}_e\,d\sigma = 0.$$

For an integer $\ell \geq 1$, we define

$$Y_h = \{c \in L^2(\Omega) : \forall E \in \mathcal{E}_h,\ c|_E \in \mathbb{P}_\ell(E)\}. \tag{53}$$

In view of (23) and (32), we discretize (51)–(52) by: Find $c_h \in Y_h$ such that for all $v_h \in Y_h$:

$$\sum_{E\in\mathcal{E}_h}\int_E \mathbf{K}\nabla c_h\cdot\nabla v_h\,d\mathbf{x}$$

$$-\sum_{e\in\Gamma_h\cup\partial\Omega}\int_e\left(\{\mathbf{K}\nabla c_h\cdot\mathbf{n}_e\}_e[v_h]_e + \varepsilon\{\mathbf{K}\nabla v_h\cdot\mathbf{n}_e\}_e[c_h]_e\right)d\sigma + J_0(c_h, v_h)$$

$$+\sum_{E\in\mathcal{E}_h}\int_E\left(\mathbf{u}_h\cdot\nabla c_h + \frac{1}{2}(\operatorname{div}\mathbf{u}_h)c_h\right)v_h\,d\mathbf{x} - \frac{1}{2}\sum_{e\in\Gamma_h\cup\partial\Omega}\int_e[\mathbf{u}_h]_e\cdot\mathbf{n}_e\{c_h v_h\}_e\,d\sigma$$

$$-\sum_{E\in\mathcal{E}_h}\int_{(\partial E)_-}\{\mathbf{u}_h\}\cdot\mathbf{n}_E(c_h^{\text{int}} - c_h^{\text{ext}})v_h^{\text{int}}\,d\sigma = \int_\Omega f v_h\,d\mathbf{x}, \tag{54}$$

where $(\partial E)_-$ is defined by (20)

$$(\partial E)_- = \{\mathbf{x} \in \partial E : \{\mathbf{u}_h\}\cdot\mathbf{n}_E(\mathbf{x}) < 0\},$$

and the parameters ε and σ_e are the same as previously.

To simplify, we introduce the form t_h with the upwind approximation of the transport term in (54):

$$t_h(\mathbf{u}_h; v_h, w_h) = \sum_{E \in \mathcal{E}_h} \int_E \left(\mathbf{u}_h \cdot \nabla v_h + \frac{1}{2}(\operatorname{div} \mathbf{u}_h) v_h \right) w_h \, d\mathbf{x}$$

$$- \sum_{E \in \mathcal{E}_h} \int_{(\partial E)_-} \{\mathbf{u}_h\} \cdot \mathbf{n}_E (v_h^{\text{int}} - v_h^{\text{ext}}) w_h^{\text{int}} \, d\sigma - \frac{1}{2} \sum_{e \in \Gamma_h \cup \partial \Omega} \int_e [\mathbf{u}_h]_e \cdot \mathbf{n}_e \{v_h w_h\}_e \, d\sigma.$$
$$(55)$$

This form is positive in the following sense (cf. [GRW05]): for all $v_h \in Y_h$

$$t_h(\mathbf{u}_h; v_h, v_h) = \frac{1}{2} \sum_{E \in \mathcal{E}_h} \| |\{\mathbf{u}_h\} \cdot \mathbf{n}_E|^{\frac{1}{2}} (v_h^{\text{int}} - v_h^{\text{ext}}) \|^2_{L^2((\partial E)_- \setminus \partial \Omega)}$$

$$+ \| |\mathbf{u}_h \cdot \mathbf{n}_\Omega|^{\frac{1}{2}} v_h \|^2_{L^2((\partial \Omega)_-)}, \quad (56)$$

where

$$(\partial \Omega)_- = \{\mathbf{x} \in \partial \Omega : \mathbf{u}_h \cdot \mathbf{n}_\Omega(\mathbf{x}) < 0\}.$$

Therefore, if the penalty parameters σ_e are chosen as in Section 3, we see that system (54) has a solution t_h in Y_h, unique up to an additive constant. In particular, this means that (54) is *compatible* with (51)–(52) and this is an important property, cf. [DSW04].

However, proving a priori error estimates is more delicate, considering that $\mathbf{u}_h$ proceeds from a previous computation. If the error in computing $\mathbf{u}_h$ is measured in the norm (42), then the contribution of $t_h(\mathbf{u}_h; c_h, v_h)$ to the error is estimated as in the Navier–Stokes equations. This requires discrete Sobolev inequalities, and as mentioned in Remark 5, this does not seem to be possible for OBB-DG schemes. On the other hand, for IIPG, SIPG and NIPG, the analysis in [GRW05] carries over here and yields, when $\mathbf{u}$ and c are sufficiently smooth:

$$\| c_h - c \|_{H^1(\mathcal{E}_h)} = \mathrm{O}(h^{\min(k,\ell)}),$$

where k is the exponent in (50).

Remark 7. Let E be an element as in Remark 3. Taking $v_h = \chi_E$ in (54), we obtain the discrete mass balance relation:

$$- \sum_{e \in \partial E} \int_e \{\mathbf{K} \nabla c_h\} \cdot \mathbf{n}_E \, d\sigma + \sum_{e \in \partial E} \frac{\sigma_e}{h_e} \int_e (c_h^{\text{int}} - c_h^{\text{ext}}) \, d\sigma$$

$$+ \frac{1}{2} \left(\int_E (\operatorname{div} \mathbf{u}_h) c_h \, d\mathbf{x} - \frac{1}{2} \sum_{e \in \partial E} \int_e (\mathbf{u}_h^{\text{int}} - \mathbf{u}_h^{\text{ext}}) \cdot \mathbf{n}_E c_h^{\text{int}} \, d\sigma \right)$$

$$+ \sum_{e \in (\partial E)_-} \int_e |\{\mathbf{u}_h\} \cdot \mathbf{n}_E| (c_h^{\text{int}} - c_h^{\text{ext}}) \, d\sigma = \int_E f \, d\mathbf{x}.$$

6 Some Darcy Flow in Porous Media: Numerical Examples

In recent years DG methods have been investigated and applied to a wide collection of fluid and solid mechanics problems arising in many engineering and scientific fields such as aerospace, petroleum, environmental, chemical and biomedical engineering, and earth and life sciences. Since the list of publications is substantial and continues to grow, we include only a few references to illustrate the diversity of applications, [CKS00]. We do provide some numerical examples arising in modeling Darcy flow and transport in porous media in which DG algorithms offer major advantages over traditional conforming finite element and finite difference methods.

Geological media such as aquifers and petroleum reservoirs exhibit a high level of spatial variability at a multiplicity of scales, from the size of individual grains or pores, to facies, stratigraphic and hydrologic units, up to sizes of formations. These problems are of great importance to a number of scientific disciplines that include the management and protection of groundwater resources, the deposition of nuclear wastes, the recovery of hydrocarbons, and the sequestration of excessive carbon dioxide. Numerical simulation of physical flows and chemical reactions in heterogeneous geological media and their interplay is required for understanding as well as designing mitigation strategies for environmental cleanup or optimizing oil and gas production.

DG methods are effective in treating complex geological heterogeneities such as impermeable boundaries or flow faults occuring in the interior of a reservoir. Because of the flexibility of DG, these boundaries do no require special meshing. Instead the face between two internal elements is simply switched to a no flow boundary condition for both neighboring elements. In Figure 2 we show an example of a mesh with 1683 triangular elements, in which the dark
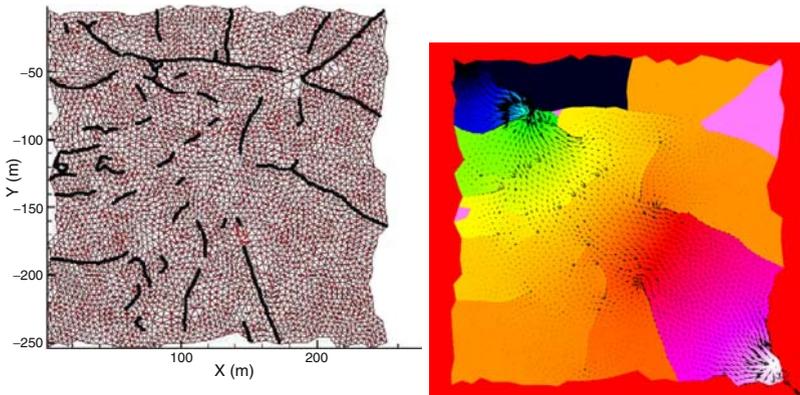
Fig. 2. Mesh with internal boundary conditions (left) and pressure and flux solutions (right)

lines are impermeable boundaries. Also shown is the corresponding pressure and flux solution and the impact of these boundaries is clearly observed.

Another important porous media application where DG could prove to be extremely important is reactive transport. When dealing with general chemistry and transport, it is imperative that the transport operators be monotone and conservative. While a number of monotone finite difference methods have been proposed for structured grids, many of these approaches have not been extended to unstructured grids. With the use of appropriate numerical fluxes, approximate Riemann solvers and stability post-processing (slope-limiting), DG methods can be used to construct discretizations which are conservative and monotone.

A benchmark case in reactive transport is a simulation of a far field nuclear waste management problem [cpl01, cpl]. The problem is characterized by large discontinuous jumps in permeability, effective porosity, and diffusivity, and by the need to model small levels of concentration of the radioactive constituents. The permeability field layers of the subsurface are shown in Figure 3.

For this example the magnitude of the velocity varies greatly in the different layers due to the discontinuities in the permeability of the layers. In addition, in the clay and marl layers, where permeability is small, transport is dominated by molecular diffusion. In the limestone and dogger limestone layers, where permeability is large, transport is dominated by advection and dispersion. This example demonstrates the ability of DG to handle both
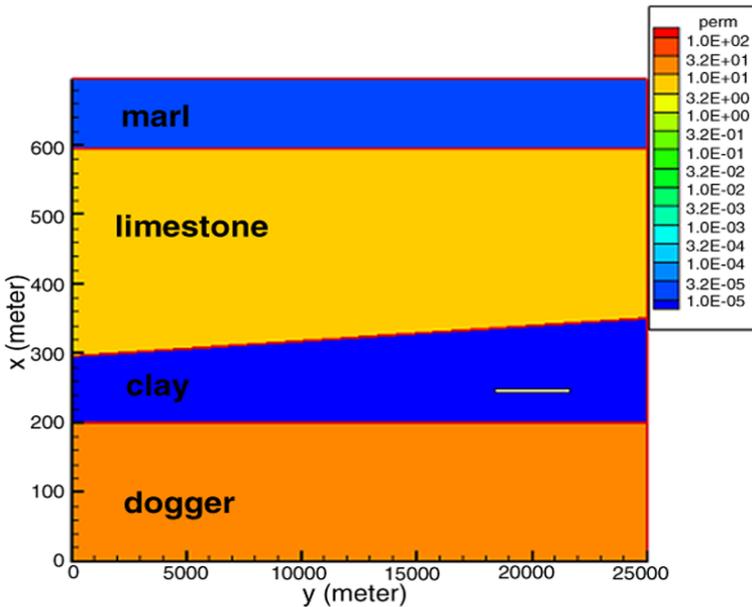
Fig. 3. Permeability field layers in the reactive transport problem
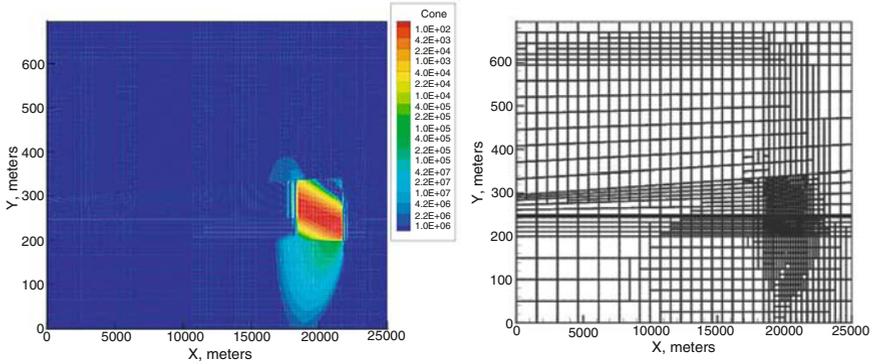
Fig. 4. Simulation of nuclear reactive transport using DG - 1
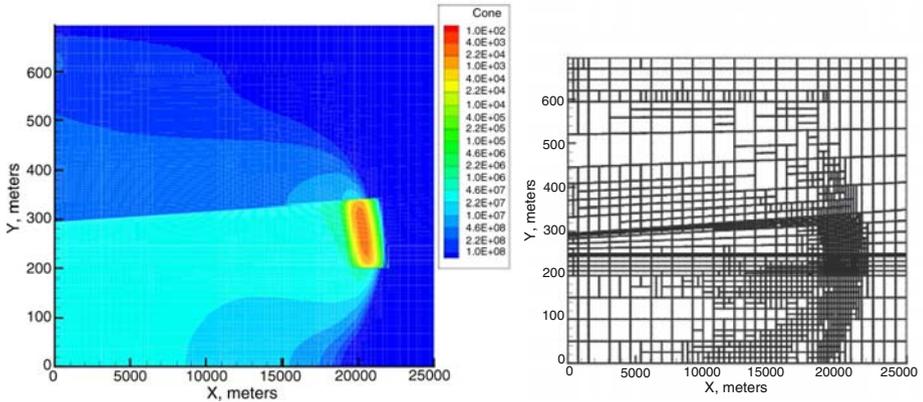
Fig. 5. Simulation of nuclear reactive transport using DG - 2

advection-dominated and diffusion-dominated problems. Figure 4 shows Iodine concentration at 200K years and Figure 5 at 2 million years. The low numerical diffusion of the DG method was also found to be important in this benchmark problem because of the long simulation time, cf. [WESR03]. Details regarding this simulation and several mesh adaptation strategies are discussed in [SW06a, SW06b]. The latter demonstrated that by employing dynamic adaptivity, time-dependent transport could be resolved without slope limiting for both long-term and short-term simulations. Moreover, mass conservation was retained locally during dynamic mesh modification.

The theoretical and computational results obtained for primal DG methods for transport and flow are summarized in Table 1. Two rows provide a comparison of the methods for treating flow problems with highly varying

Table 1. Primal DG for transport

	OBB-DG	NIPG	SIPG	IIPG
Penalty Term	0	≥ 0	$> \sigma_0 > 0$	> 0 and $\ll \sigma_0$
Optimality in $L^2(H^1)$ or H^1	Yes	Yes	Yes	Yes
Optimality in $L^2(L^2)$ or L^2	No	No	Yes	No
Robust probs. with highly var. coeffs.	Yes	Yes	No	Yes
Scalar primary interest(transp.)	No	No	Yes	No
Compatibility Flow Condition	No	No	No	Yes

coefficients and for transport problems in which the scalar variable is of primary interest. These results were obtained from an extensive set of numerical experiments. The studies indicate that the non-symmetric DG formulations are more robust in handling rough coefficients. The symmetric form performs better for treating diffusion/advection/reaction problems since the SIPG form yield optimal L^2 and non-negative norm estimates. The last row summarizes a compatibility condition formulated in [DSW04] in which the objective is to choose a flow field that preserves positive concentrations in reactive transport. The IIPG method is the only primal DG for which this holds.

DG methods are currently being investigated for modeling multiphase flow in porous media, e.g., see [BR04, KR06] for two-phase incompressible and for two and three phases compressible systems see [HF06, Esl05, SW]. While much progress has been made in modeling transport a major disadvantage for DG has been the development of efficient parallel solvers for large linear and nonlinear systems, the pressure equation or a fully implicit formulation for multiphase flow respectively. The development of DG solvers is an active area of research and new domain decomposition approaches are currently being developed, e.g., see [Kan05, Joh05, AA07, Esl05, BR00].

References

[AA07]   P. F. Antonietti and B. Ayuso. Schwarz domain decomposition precon-
           ditioners for discontinuous Galerkin approximations of elliptic problems:
           non-overlapping case. *M2AN Math. Model. Numer. Anal.*, 41(1):21–54,
           2007.

[ABCM02]   D. N. Arnold, F. Brezzi, B. Cockburn, and L. D. Marini. Unified analy-
           sis of discontinuous Galerkin methods for elliptic problems. *SIAM J.
           Numer. Anal.*, 39(5):1749–1779, 2002.

[Arn79]   D. N. Arnold. *An interior penalty finite element method with discontin-
           uous elements.* PhD thesis, University of Chicago, Chicago, IL, 1979.

[Arn82]   D. N. Arnold. An interior penalty finite element method with discon-
           tinuous elements. *SIAM J. Numer. Anal.*, 19(4):742–760, 1982.

[Bab73]   I. Babuška. The finite element method with Lagrangian multipliers.
           Numer. Math., 20:179–192, 1973.

[Bak77]   G. Baker. Finite element methods for elliptic equations using noncon-
           forming elements. *Math. Comp.*, 31:45–59, 1977.
[Bau97]   C. E. Baumann. *An hp-adaptive discontinuous finite element method
           for computational fluid dynamics.* PhD thesis, University of Texas at
           Austin, Austin, TX, 1997.
[Bey94]   K. S. Bey. *An hp-adaptive discontinuous Galerkin method for hyperbolic
           conservative laws.* PhD thesis, University of Texas at Austin, Austin,
           TX, 1994.
[BO99]   C. E. Baumann and J. T. Oden. A discontinuous *hp* finite element
           method for convection-diffusion problems. *Comput. Methods Appl.
           Mech. Engrg.*, 175(3–4):311–341, 1999.
[BOP96]   K. S. Bey, J. T. Oden, and A. Patra. *hp*-version discontinuous Galerkin
           methods for hyperbolic conservation laws. *Compt. Methods Appl. Mech.
           Engrg.*, 133:259–286, 1996.
[BR00]   P. Bastian and V. Reichenberger. Multigrid for higher order disconti-
           uous Galerkin finite elements applied to groundwater flow. Technical
           Report 2000-37, SFB 359, 2000.
[BR04]   P. Bastian and B. Riviere. Discontinuous Galerkin for two-phase flow
           in porous media. Technical Report 2004-28, IWR(SFB 359), University
           of Heidelberg, 2004.
[Bre74]   F. Brezzi. On the existence, uniqueness and approximation of the
           saddle-point problems arising from Lagrangian multipliers. *RAIRO
           Anal. Numér.*, 8:129–151, 1974.
[Bre03]   S. Brenner. Poincaré–Friedrichs inequalities for piecewise h^1 functions.
           SIAM J. Numer. Anal., 41:306–324, 2003.
[Bre04]   S. Brenner. Korn's inequalities for piecewise h^1 vector fields. *Math.
           Comp.*, 73:1067–1087, 2004.
[BZ73]   I. Babuška and M. Zlámal. Nonconforming elements in the finite element
           method with penalty. *SIAM J. Numer. Anal.*, 10:863–875, 1973.
[Cia91]   P. G. Ciarlet. Basic error estimates for elliptic problems. In P. G.
           Ciarlet and J. L. Lions, editors, *Handbook of Numerical Analysis, Vol.
           II*, pages 17–351. North-Holland, Amsterdam, 1991.
[CKS00]   B. Cockburn, G. E. Karniadakis, and C.-W. Shu, editors. *Discontinuous
           Galerkin methods. Theory, computation and applications (Newport, RI,
           1999).* Number 11 in Lecture Notes in Computational Science and
           Engineering. Springer-Verlag, Berlin, 2000.
[cpl]   Couplex1 test case, nuclear waste disposal far field simulation. ANDRA
           (the French National Radioactive Waste Management Agency),
           http://www.andra.fr/couplex/.
[cpl01]   The couplex test cases. ANDRA (the French National Radioactive
           Waste Management Agency), http://www.andra.fr/couplex/, 2001.
[CR73]   M. Crouzeix and P. A. Raviart. Conforming and non-conforming finite
           element methods for solving the stationary Stokes problem. *RAIRO
           Anal. Numér.*, 8:33–76, 1973.
[Dar80]   B. L. Darlow. *An Penalty-Galerkin method for solving the miscible
           displacement problem.* PhD thesis, Rice University, Houston, TX, 1980.
[Dau89]   M. Dauge. Stationary Stokes and Navier–Stokes systems on two
           or three-dimensional domains with corners. *SIAM J. Math. Anal.*,
           20(1):74–97, 1989.

[DD76]   J. Douglas, Jr. and T. Dupont. Interior penalty procedures for elliptic
          and parabolic Galerkin methods. In *Computing Methods in Applied
          Sciences (Second Internat. Sympos., Versailles, 1975)*, number 58 in
          Lecture Notes in Phys., pages 207–216. Springer-Verlag, Berlin, 1976.

[DSW04]   C. Dawson, S. Sun, and M. F. Wheeler. Compatible algorithms for
          coupled flow and transport. *Comput. Methods Appl. Mech. Engrg.*,
          194:2565–2580, 2004.

[Esl05]   O. Eslinger. *Discontinuous Galerkin finite element methods applied to
          two-phase air-water flow problems.* PhD thesis, University of Texas at
          Austin, Austin, TX, 2005.

[GR79]   V. Girault and P.-A. Raviart. An analysis of upwind schemes for the
          Navier–Stokes equations. *SIAM J. Numer. Anal.*, 19(2):312–333, 1979.

[GR86]   V. Girault and P.-A. Raviart. *Finite Element Methods for the Navier–
          Stokes Equations. Theory and Algorithms.* Number 5 in Springer Series
          in Computational Mathematics. Springer-Verlag, Berlin, 1986.

[Gri85]   P. Grisvard. *Elliptic Problems in Nonsmooth Domains.* Number 24 in
          Pitman Monographs and Studies in Mathematics. Pitman, Boston, MA,
          1985.

[GRW05]   V. Girault, B. Rivière, and M. Wheeler. A discontinuous Galerkin
          method with non-overlapping domain decomposition for the Stokes and
          Navier–Stokes problems. *Math. Comp.*, 74:53–84, 2005.

[GSWY]   V. Girault, S. Sun, M. F. Wheeler, and I. Yotov. Coupling discontinuous
          Galerkin and mixed finite element discretizations using mortar finite
          elements. *SIAM J. Numer. Anal.* Submitted Oct. 2006.

[HF06]   H. Hoteit and A. Firoozabadi. Compositional modeling by the combined
          discontinuous Galerkin and mixed methods. *SPE J.*, 11:19–34, 2006.

[Joh05]   K. Johannsen. A symmetric smoother for the nonsymmetric interior
          penalty discontinuous Galerkin discretization. ICES Report 05-23, Uni-
          versity of Texas at Austin, 2005.

[Kan05]   G. Kanschat. Block preconditioners for LDG discretizations of linear
          incompressible flow problems. *J. Sci. Comput.*, 22(1–3):371–384, 2005.

[KR06]   W. Klieber and B. Riviere. Adaptive simulations of two-phase flow by
          discontinuous Galerkin methods. *Comput. Methods Appl. Mech. Engrg.*,
          196(1–3):404–419, 2006.

[LR74]   P. Lesaint and P. A. Raviart. On a finite element method for solving
          the neutron transport equation. In C. deBoor, editor, *Mathematical
          Aspects of Finite Elements in Partial Differential Equations*, pages 89–
          123. Academic Press, 1974.

[Nit71]   J. A. Nitsche. Über ein Variationsprinzip auf Lösung von Dirichlet-
          Problemen bei Verwendung von Teilsraumen, die keinen Randbedin-
          gungen unteworfen sind. *Math. Sem. Univ. Hamburg*, 36:9–15, 1971.

[OBB98]   J. T. Oden, I. Babuška, and C. E. Baumann. A discontinuous hp finite
          element method for diffusion problems. *J. Comput. Phys.*, 146:491–516,
          1998.

[OW75]   J. T. Oden and L. C. Wellford, Jr. Discontinuous finite element approx-
          imations for the analysis of shock waves in nonlinearly elastic materials.
          J. Comput. Phys., 19(2):179–210, 1975.

[Pir89]   O. Pironneau. *Finite Element Methods for Fluids.* Wiley, Chichester,
          1989.

[RH73]   W. H. Reed and T. R. Hill. Triangular mesh methods for the neutron transport equation. Los Alamos Scientific Laboratory Report LA-UR-73-479, 1973.

[Riv00]   B. Rivière. *Discontinuous Galerkin finite element methods for solving the miscible displacement problem in porous media.* PhD thesis, University of Texas at Austin, Austin, TX, 2000.

[RT75]   P. A. Raviart and J. M. Thomas. A mixed finite element method for second order elliptic problems. In *Mathematical Aspects of Finite Element Methods*, number 606 in Lecture Notes in Mathematics. Springer-Verlag, Berlin, 1975.

[RW74]   H. Rachford and M. F. Wheeler. An H1-Galerkin procedure for the two-point boundary value problem. In C. deBoor, editor, *Mathematical Aspects of Finite Elements in Partial Differential Equations*, pages 353–382. Academic Press, 1974.

[RWG99]   B. Riviere, M. F. Wheeler, and V. Girault. Part I: Improved energy estimates for interior penalty, constrained and discontinuous Galerkin methods for elliptic problems. *Comput. Geosci.*, 3:337–360, 1999.

[RWG01]   B. Riviere, M. F. Wheeler, and V. Girault. A priori error estimates for finite element methods based on discontinuous approximation spaces for elliptic problems. *SIAM J. Numer. Anal.*, 39(3):902–931, 2001.

[SST03]   D. Shotzau, C. Schwab, and A. Toselli. Mixed *hp*-DGFEM for incompressible flows. *SIAM J. Numer. Anal.*, 40(319):2171–2194, 2003.

[SW]   S. Sun and M. F. Wheeler. Discontinuous Galerkin methods for multiphase compressible flows. In preparation.

[SW06a]   S. Sun and M. F. Wheeler. Anisotropic and dynamic mesh adaptation for discontinuous Galerkin methods applied to reactive transport. *Comput. Methods Appl. Mech. Engrg.*, 195(25–28):3382–3405, 2006.

[SW06b]   S. Sun and M. F. Wheeler. A posteriori error estimation and dynamic adaptivity for symmetric discontinuous Galerkin approximations of reactive transport problems. *Comput. Methods Appl. Mech. Engrg.*, 195:632–652, 2006.

[Tem79]   R. Temam. *Navier–Stokes equations. Theory and numerical analysis.* North-Holland, Amsterdam, 1979.

[WD80]   M. F. Wheeler and B. L. Darlow. Interior penalty Galerkin procedures for miscible displacement problems in porous media. In *Computational methods in nonlinear mechanics (Proc. Second Internat. Conf., Univ. Texas, Austin, Tex., 1979)*, pages 485–506, Amsterdam, 1980. North-Holland.

[WESR03] M. F. Wheeler, O. Eslinger, S. Sun, and B. Rivière. Discontinuous Galerkin method for modeling flow and reactive transport porous media. In *Analysis and Simulation of Multifield Problems*, pages 37–58. Springer-Verlag, Berlin, 2003.

[Whe78]   M. F. Wheeler. An elliptic collocation-finite element method with interior penalties. *SIAM J. Numer. Anal.*, 15(1):152–161, 1978.

Mixed Finite Element Methods on Polyhedral Meshes for Diffusion Equations

Yuri A. Kuznetsov

Department of Mathematics, University of Houston, 651 Philip G. Hoffman Hall, Houston, TX 77204–3008, USA `kuz@math.uh.edu`

Summary. In this paper, a new mixed finite element method for the diffusion equation on polyhedral meshes is proposed. The method is applied to the diffusion equation on meshes with mixed cells when all the coefficients and the source function may have discontinuities inside polyhedral mesh cells. The resulting discrete equations operate only with the degrees of freedom for normal fluxes on the boundaries of cells and one degree of freedom per cell for the solution function.

Key words: Diffusion equation, mixed finite element method, polyhedral meshes, mixed cells

1 Introduction

In this paper, we propose a new mixed finite element method for the diffusion equation on general polyhedral meshes in the case when the coefficients of the equation and the source function may have strong discontinuities inside mesh cells. Such mesh cells are called mixed ones. The major idea of the method is reported in [Kuz05]. This work is a natural extension of the method in [Kuz06] to 3D diffusion equations.

The discretization method consists of several steps. At the first step, we partition each polyhedral cell into polyhedral subcells assuming that inside each subcell the coefficients and the source function are relatively smooth. Then, in each subcell we impose a local conforming tetrahedral mesh subject to a structure of the neighboring subcells. The subcell tetrahedral meshes are not required to be conforming on the interfaces between subcells. A special finite element subspace of $H_{\mathrm{div}}(\Omega)$ is invented, and the classical mixed finite element method [BF91, RT91] is used for discretization of the diffusion equation with the Neumann boundary condition. At the final step, the interior (with respect to the boundaries of polyhedral mesh cells) degrees of freedom for the normal fluxes and for the solution function are eliminated, and a new

degree of freedom per mesh cell for the solution function is defined. The final system of discrete equations has the same structure as for the classical mixed FE method.

The paper is organized as follows. In Section 2, we formulate the problem and requirements for the discretization. In Section 3, we describe partitionings of mesh cells into subcells and polyhedral meshes to be used for the discretization. We also propose a special finite element subspace of $H_{\mathrm{div}}(\Omega)$ for the mixed finite element method. Finally, in Section 4, we describe a condensation procedure for the underlying algebraic system and transform the condensed system into the standard form which is typical for the classical finite element method on simplicial meshes. In the final part of Section 3, we propose an alternative discretization method. In Remark 2 of Section 4, we prove that this discretization method is equivalent to the "div-const" mixed finite element method invented and investigated in [KR03, KR05].

2 Problem Formulation

We consider the diffusion equation

$$- \operatorname{div}(a \operatorname{grad} p) + cp = f \quad \text{in } \Omega \tag{1}$$

with the Neumann boundary condition

$$(a \operatorname{grad} p) \cdot \mathbf{n} = 0 \quad \text{on } \partial\Omega \tag{2}$$

where Ω is a polyhedral domain in $\mathbb{R}^3$ with the boundary $\partial\Omega$, $a = a(x)$ is a symmetric positive definite 3×3 matrix (diffusion tensor) for any $x = (x_1, x_2, x_3) \in \Omega$, c is a nonnegative function, f is a given source function, and $\mathbf{n}$ is the outward unit normal to $\partial\Omega$. The domain Ω is partitioned into m open non-overlapping simply connected polyhedral subdomains Ω_k with the boundaries $\partial\Omega_k$, $k = \overline{1, m}$, i.e. $\overline{\Omega} = \bigcup_{k=1}^{m} \overline{\Omega}_k$. For the sake of simplicity, we assume that in each of the subdomains Ω_k the matrix a has constant entries and the coefficient c is a nonnegative constant, $k = \overline{1, m}$. We naturally assume that in the case $c \equiv 0$ in Ω the compatibility condition

$$\int_{\Omega} f \, \mathrm{dx} = 0 \tag{3}$$

holds.

In this paper, we consider problem (1), (2) in the form of the first order system

$$
\begin{aligned}
a^{-1}\mathbf{u} + \operatorname{grad} p &= 0 && \text{in } \Omega, \\
- \operatorname{div} \mathbf{u} - \quad cp &= -f && \text{in } \Omega, \\
\mathbf{u} \cdot \mathbf{n} \qquad\quad &= 0 && \text{on } \partial\Omega,
\end{aligned}
\tag{4}
$$

where $\mathbf{u}$ is said to be the flux vector function.

Let Ω_H be a polyhedral mesh in Ω with polyhedral mesh cells $E_k = \overline{E}_k \setminus \partial E_k$ where ∂E_k are the boundaries of E_k, $k = \overline{1, n}$. Here, n is a positive integer. We assume that $E_k \cap E_l = \emptyset$, $l \neq k$, $k, l = \overline{1, n}$, and $\overline{\Omega} = \bigcup_{k=1}^{n} \overline{E}_k$. We do not assume that the mesh Ω_H is geometrically conforming, i.e. the interfaces $\partial E_k \cap \partial E_l$ between two neighboring cells E_k and E_l are not obliged to be either a face, or an edge, or a vertex of these cells, $l \neq k$, $k, l = \overline{1, n}$. An example of two nonconforming neighboring prismatic cells is given in Figure 1.

The intersection of E_k with $\bigcup_{l=1}^{m} \partial \Omega_l$ defines the partitioning of E_k into n_k polyhedral subcells $E_{k,s}$, $s = \overline{1, n_k}$, $k = \overline{1, n}$. An example of a partitioning of a mesh cell into three subcells is given in Figure 2.
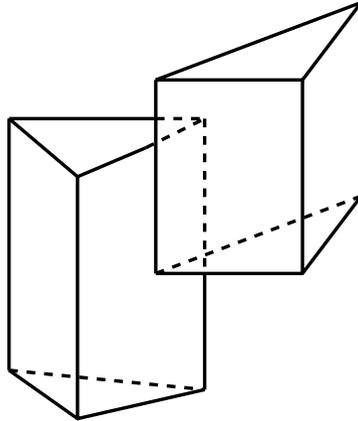

Fig. 1. An example of two neighboring prismatic mesh cells with nonconforming intersecting faces
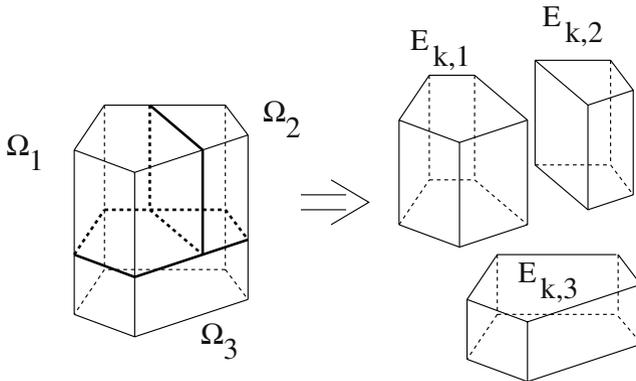

Fig. 2. An example of a partitioning of a polyhedral cell into three polyhedral subcells

A mesh cell E with discontinuities either of the entries of the matrix a, or the coefficient c, or both is said to be a mixed cell.

On the boundary ∂E_k of a polyhedral cell E_k we define a set of s_k non-overlapping flat polygons $\Gamma_{k,i}$, $i = \overline{1, s_k}$, which satisfies the following three conditions:

1. $\partial E_k = \bigcup_{i=1}^{s_k} \overline{\Gamma}_{k,i}$;
2. each $\Gamma_{k,i}$ belongs to $\partial E_{k,s}$ for some $s \leq n_k$;
3. each $\Gamma_{k,i}$ belongs either to $\partial \Omega$ or to $\partial E_{k',s'}$ for some $k' \neq k$, $s' \leq n_{k'}$, $k' \leq n$,

where s_k is a positive integer, $k = \overline{1, n}$. A 2D example of the partitioning of ∂E_k into $\Gamma_{k,i}$, $i = \overline{1, s_k}$, with $s_k = 8$ is given in Figure 3.

The goal of this paper is to develop a mixed finite element method for the diffusion problem (4) on the above described polyhedral meshes under special conditions on the degrees of freedom (DOF) which can be used for discretization. Namely, the final discretization can use only one DOF representing the normal component of the solution flux vector function $\mathbf{u}$ in (4) on each $\Gamma_{k,i}$, $i = \overline{1, s_k}$, and only one DOF representing the solution function p in (4) in each E_k, $k = \overline{1, n}$.

To predict the final discretization scheme to be derived in Section 4, we define the required discrete equation in E_k for the second equation in (4) by integrating this equation over the mesh cell E_k:
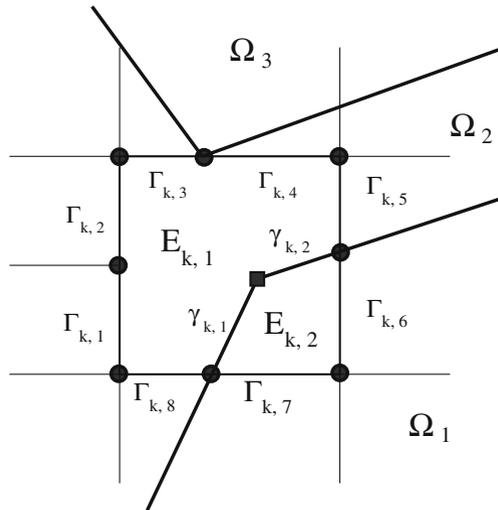

Fig. 3. A 2D example of the partitionings ∂E_k into $\Gamma_{k,i}$, $i = \overline{1, 8}$, and $\partial E_{k,1} \bigcap \partial E_{k,2}$ into $\gamma_{k,j}$, $j = 1, 2$

$$\int_{E_k} [-\operatorname{div} \mathbf{u} - cp]\, dx = -\int_{E_k} f\, dx, \quad k = \overline{1,n}. \tag{5}$$

The latter equality results in the discrete equation

$$-\sum_{i=1}^{s_k} u_{k,i}|\Gamma_{k,i}| - c_k|E_k|\hat{p}_k = -|E_k|f_k, \tag{6}$$

where

$$u_{k,i} = \frac{1}{|\Gamma_{k,i}|}\int_{\Gamma_{k,i}} \mathbf{u}\cdot\mathbf{n}_k\, ds \tag{7}$$

is the mean value of the normal flux $\mathbf{u}\cdot\mathbf{n}_k$ on $\Gamma_{k,i}$,

$$c_k = \frac{1}{|E_k|}\int_{E_k} c\, dx \quad \text{and} \quad f_k = \frac{1}{|E_k|}\int_{E_k} f\, dx \tag{8}$$

are the mean values of c and f in E_k, respectively,

$$\hat{p}_k = \frac{\displaystyle\int_{E_k} cp\, dx}{\displaystyle\int_{E_k} c\, dx} \tag{9}$$

is the c-weighted mean value of p in E_k. Here, $|\Gamma_{k,i}|$ and $|E_k|$ denote the length of $\Gamma_{k,i}$ and the area of E_k, respectively, $i = \overline{1, s_k}$, and $\mathbf{n}_k$ is the outward unit normal to ∂E_k, $k = \overline{1,n}$.

The equation (6) can be written in the matrix form by

$$B_H^{0,(k)}\bar{u}^{(k)} - c_k|E_k|\hat{p}_k = -|E_k|f_k, \tag{10}$$

where

$$B_H^{0,(k)} = -\left[|\Gamma_{k,1}| \cdots |\Gamma_{k,s_k}|\right] \in \mathbb{R}^{1\times s_k} \tag{11}$$

and $\bar{u}^{(k)} = \left[u_{k,1}, \ldots, u_{k,s_k}\right]^T \in \mathbb{R}^{s_k}$, $k = \overline{1,n}$. The matrix $B_H^{0,(k)}$ will be used later to derive the final discretization for the problem (4).

The formula (9) assumes that the coefficient c is not equal identically to zero in E_k. In the case $c \equiv 0$ in E_k the discrete equation (6) is replaced by the equation

$$-\sum_{i=1}^{s_k} u_{k,i}|\Gamma_{k,i}| = -|E_k|f_k, \tag{12}$$

and (10) is replaced by the equation

$$B_H^{0,(k)}\bar{u}^{(k)} = -|E_k|f_k. \tag{13}$$

3 Mixed Finite Element Method

Let $\partial_0 E_{k,s}$ be the part of the boundary $\partial E_{k,s}$ of a polyhedral subcell $E_{k,s}$ belonging to the interior of E_k, i.e. $\partial_0 E_{k,s} = \partial E_{k,s} \bigcap E_k$, $s = \overline{1, n_k}$, $k = \overline{1, n}$. On $\partial_0 E_{k,s}$ we define a set of $t_{k,s}$ non-overlapping flat polygons $\gamma_{k,s,j}$ which satisfies the following two conditions:

1. $\overline{\partial_0 E_{k,s}} = \bigcup_{j=1}^{t_{k,s}} \bar{\gamma}_{k,s,j}$,
2. each $\gamma_{k,s,j}$ belongs to $\partial_0 E_{k,s'}$ for some $s' \neq s$, $s' \leq n_k$,

where $t_{k,s}$ is a positive integer, $s = \overline{1, n_k}$, $k = \overline{1, n}$.

Examples of the partitionings of $\partial_0 E_{k,s}$ into polygons $\gamma_{k,s,j}$ are given in Figures 3 and 4. In Figure 3, the interface $\partial_0 E_{k,1} = \partial_0 E_{k,2}$ between $E_{k,1}$ and $E_{k,2}$ consists of $\gamma_{k,1}$ and $\gamma_{k,2}$. In Figure 4, $\partial_0 E_{k,1}$ consists of $\gamma_{k,1,1} = \gamma_1$, $\gamma_{k,1,2} = \gamma_2$, and $\gamma_{k,1,3} = \gamma_3$, and $\partial_0 E_{k,2}$ consists of $\gamma_{k,2,1} = \gamma_3$, $\gamma_{k,2,2} = \gamma_4$, and $\gamma_{k,2,3} = \gamma_5$. Finally, $\partial_0 E_{k,3}$ consists of $\gamma_{k,3,1} = \gamma_1$, $\gamma_{k,3,2} = \gamma_2$, $\gamma_{k,3,3} = \gamma_4$, and $\gamma_{k,3,4} = \gamma_5$.

Let $\mathcal{T}_{h,k,s} = \{e_{k,s,i}\}$ be conforming tetrahedral partitionings of $E_{k,s}$, $s = \overline{1, n_k}$, $k = \overline{1, n}$. The conformity of a tetrahedral partitioning (tetrahedral mesh) means that any two different intersecting closed tetrahedrons in $\mathcal{T}_{h,k,s}$ have either a common vertex, or a common edge, or a common face.

The boundaries $\partial E_{k,s}$ of $E_{k,s}$ are unions of polygons in $\{\Gamma_{k,i}\}$ and in $\{\gamma_{k,s,j}\}$, $s = \overline{1, n_k}$, $k = \overline{1, n}$. We assume that each of the tetrahedral meshes $\mathcal{T}_{h,k,s}$ is also conforming with respect to the boundaries of polygons in $\{\Gamma_{k,i}\}$ and in $\{\gamma_{k,s,j}\}$ belonging to $\partial E_{k,s}$, i.e. these boundaries belong to the union of
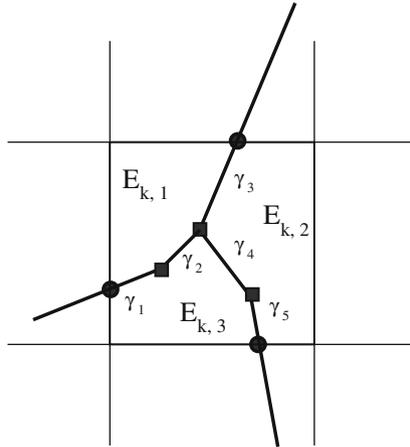

Fig. 4. An example of partitionings $\partial_0 E_{k,s}$ into segments $\gamma_{k,j,s}$, $j = \overline{1, t_k}$, $s = \overline{1, 3}$

edges of tetrahedrons in $\mathcal{T}_{h,k,s}$, $s = \overline{1, n_k}$, $k = \overline{1, n}$. We do not assume that the tetrahedral meshes $\mathcal{T}_{h,k,s}$ and $\mathcal{T}_{h,k',s'}$ are conforming on the interfaces between neighboring cells E_k and $E_{k'}$ when $k' \neq k$ as well as on the interfaces between neighboring subcells $E_{k,s}$ and $E_{k,s'}$ when $k' = k$.

Let $\mathcal{T}_h$ be a tetrahedral partitioning of Ω such that its restrictions onto $E_{k,s}$ coincide with the tetrahedral meshes $\mathcal{T}_{h,k,s}$, and let $RT_0(E_{k,s})$ be the lowest order Raviart–Thomas finite element spaces on $\mathcal{T}_{h,k,s}$, $s = \overline{1, n_k}$, $k = \overline{1, n}$. We define the finite element spaces $\mathbf{V}_{h,k,s}$ consisting of vector functions $\mathbf{w} \in RT_0(E_{k,s})$ which have constant normal fluxes $\mathbf{w} \cdot \mathbf{n}_{k,s}$ on each of the flat polygons $\Gamma_{k,i}$ and $\gamma_{k,j}$ belonging to $\partial E_{k,s}$, where $\mathbf{n}_{k,s}$ are the outward unit normals to $\partial E_{k,s}$, $s = \overline{1, n_k}$, $k = \overline{1, n}$. Then, we define the spaces $\mathbf{V}_{h,k}$ on E_k assuming that the restrictions $\mathbf{w}_{k,s}$ of any vector function $\mathbf{w}_k \in \mathbf{V}_{h,k}$ onto $E_{k,s}$ belong to the spaces $\mathbf{V}_{h,k,s}$, $s = \overline{1, n_k}$, and the normal components of $\mathbf{w}_k$ are continuous through $\gamma_{k,s,j}$, $j = \overline{1, t_k}$. To satisfy the latter condition we assume that on each polygon $\gamma_{k,s,j}$ belonging to $\partial E_{k,s} \cap \partial E_{k,s'}$, $s' \neq s$, the outward normal components of vector functions $\mathbf{w}_{k,s}$ and $\mathbf{w}_{k,s'}$ satisfy the equalities $\mathbf{w}_{k,s} \cdot \mathbf{n}_{k,s} + \mathbf{w}_{k,s'} \cdot \mathbf{n}_{k,s'} = 0$ (we recall that $\mathbf{n}_{k,s} + \mathbf{n}_{k,s'} = 0$), $j = \overline{1, t_{k,s}}$, $k = \overline{1, n}$.

Finally, we define the finite element space $\mathbf{V}_h$ assuming that the restrictions $\mathbf{w}_k$ of any vector function $\mathbf{w} \in \mathbf{V}_h$ onto E_k belong to the spaces $\mathbf{V}_{h,k}$ and the normal components of $\mathbf{w}$ are continuous on the interfaces $\partial E_k \cap \partial E_l$ between E_k and E_l. To satisfy the latter condition we assume that on each polygon $\Gamma_{k,i}$ belonging to $\partial E_k \cap \partial E_l$ the outward normal components of vector functions $\mathbf{w}_k$ and $\mathbf{w}_l$ satisfy the condition $\mathbf{w}_k \cdot \mathbf{n}_k + \mathbf{w}_l \cdot \mathbf{n}_l = 0$, $1 \leq i \leq s_k$, $l \neq k$, $k, l = \overline{1, n}$.

We define the finite element space Q_h for the solution function p by setting that functions in Q_h are constant in each of the tetrahedrons in the partitionings $\mathcal{T}_{h,k,s}$, $s = \overline{1, n_k}$, $k = \overline{1, n}$. With the defined FE spaces $\mathbf{V}_h$ and Q_h, the mixed finite element discretization to (4) is as follows: Find $\mathbf{u}_h \in \mathbf{V}_h$, $\mathbf{u}_h \cdot \mathbf{n} = 0$ on $\partial \Omega$, and $p_h \in Q_h$, such that

$$\int_{\Omega} (a^{-1} \mathbf{u}_h) \cdot \mathbf{v} \, dx - \int_{\Omega} p_h \operatorname{div} \mathbf{v} \, dx = 0,$$
$$-\int_{\Omega} \operatorname{div} \mathbf{u}_h q \, dx \quad - \int_{\Omega} c p_h q \, dx \quad = -\int_{\Omega} f q \, dx$$

(14)

for all $\mathbf{v} \in \mathbf{V}_h$, $\mathbf{v} \cdot \mathbf{n} = 0$ on $\partial \Omega$, and $q \in Q_h$.

Finite element problem (14) results in the system of linear algebraic equations

$$\begin{aligned} M\bar{u} + B^T \bar{p} + C^T \bar{\lambda} &= 0, \\ B\bar{u} - \Sigma \bar{p} \quad\quad &= \bar{F}, \\ C\bar{u} \quad\quad\quad &= 0. \end{aligned}$$

(15)

Here, $M \in \mathbb{R}^{\hat{n} \times \hat{n}}$ is a symmetric positive definite matrix, $\Sigma \in \mathbb{R}^{N \times N}$ is either a symmetric positive definite or a symmetric positive semidefinite matrix, $B \in \mathbb{R}^{N \times \hat{n}}$, and $C \in \mathbb{R}^{\tilde{n} \times \hat{n}}$, where $\hat{n} = \dim \mathbf{V}_h$, N is the total number of

tetrahedrons in $\mathcal{T}_h$, and $\tilde{n}$ is the total number of polygons $\Gamma_{k,i}$, $i = \overline{1, s_k}$, $k = \overline{1, n}$, belonging to $\partial\Omega$. The components of the Lagrange multiplier vector $\bar{\lambda} \in \mathbb{R}^{\tilde{n}}$ represent the mean values of the solution function p on the polygons $\Gamma_{k,i} \subset \partial\Omega$, $i = \overline{1, s_k}$, $k = \overline{1, n}$. The third matrix equation in (15) takes care of the Neumann boundary condition on $\partial\Omega$.

We also consider another discretization to (4): Find $\mathbf{u}_h \in \mathbf{V}_h$, $\mathbf{u}_h \cdot \mathbf{n} = 0$ on $\partial\Omega$, and $p_h \in Q_h$ such that

$$
\begin{aligned}
\int_\Omega \left(a^{-1}\mathbf{u}_h\right) \cdot \mathbf{v}\,\mathrm{dx} - \int_\Omega \tilde{p}_h \operatorname{div} \mathbf{v}\,\mathrm{dx} &= 0, \\
-\int_\Omega \operatorname{div} \mathbf{u}_h q\,\mathrm{dx} \quad - \int_\Omega c\tilde{p}_h q\,\mathrm{dx} \quad &= -\int_\Omega \tilde{f}_h q\,\mathrm{dx}
\end{aligned}
\tag{16}
$$

for all $\mathbf{v} \in \mathbf{V}_h$, $\mathbf{v} \cdot \mathbf{n} = 0$ on $\partial\Omega$, and $q \in Q_h$. Here,

$$
\tilde{p}_h(x) = \frac{1}{|E_{k,s}|} \int_{E_{k,s}} p_h(x')\,\mathrm{dx}', \quad x \in E_{k,s},
\tag{17}
$$

and

$$
\tilde{f}_h(x) = \frac{1}{|E_{k,s}|} \int_{E_{k,s}} f(x')\,\mathrm{dx}', \quad x \in E_{k,s},
\tag{18}
$$

where $|E_{k,s}|$ is the volume of $E_{k,s}$, $s = \overline{1, n_k}$, $k = \overline{1, n}$.

The finite element problem (16) results in the system of linear algebraic equations

$$
\begin{aligned}
M\bar{u} + B^T\bar{p} + C^T\bar{\lambda} &= 0, \\
B\bar{u} - \widetilde{\Sigma}\bar{p} \quad\quad &= \overline{F}_1, \\
C\bar{u} \quad\quad\quad &= 0,
\end{aligned}
\tag{19}
$$

where the matrices M, B, and C are the same as in the system (15). The matrix $\widetilde{\Sigma} \in \mathbb{R}^{N \times N}$ is a block diagonal matrix with $\widetilde{N} = \sum_{k=1}^n n_k$ diagonal submatrices

$$
\widetilde{\Sigma}_{k,s} = \frac{1}{|E_{k,s}|} c_{k,s} D_{k,s} \bar{e}_{k,s} \bar{e}_{k,s}^T D_{k,s} \in \mathbb{R}^{N_{k,s} \times N_{k,s}}
\tag{20}
$$

and the vector $\overline{F}_1 \in \mathbb{R}^N$ consists of $\widetilde{N}$ subvectors

$$
\overline{F}_{k,s} = -f_{k,s} D_{k,s} \bar{e}_{k,s} \in \mathbb{R}^{N_{k,s}}
\tag{21}
$$

(one matrix $\widetilde{\Sigma}_{k,s}$ and one vector $\overline{F}_{k,s}$ per subcell $E_{k,s}$), where $c_{k,s}$ is the value of the coefficient c in $E_{k,s}$, $f_{k,s}$ is the value of the function $\tilde{f}_h$ in $E_{k,s}$, $\bar{e}_{k,s} = (1, \ldots, 1)^T \in \mathbb{R}^{N_{k,s}}$, and $N_{k,s}$ is the total number of tetrahedrons in $\mathcal{T}_{h,k,s}$, $s = \overline{1, n_k}$, $k = \overline{1, n}$. Here, $D_{k,s}$ are diagonal $N_{k,s} \times N_{k,s}$ matrices with the volumes of tetrahedrons $\{e_{k,s,i}\}$ in $\mathcal{T}_{h,k,s}$ on the diagonals, $s = \overline{1, n_k}$, $k = \overline{1, n}$.

In Section 4, we shall prove that the method (16)–(18) is equivalent to the "div-const" mixed finite element method [KR03, KR05] on the polyhedral mesh consisting of the polyhedral mesh cells $E_{k,s}$, $s = \overline{1, n_k}$, $k = \overline{1, n}$.

4 Hybridization and Condensation

The underlying system of algebraic equations for the problem (14) can be written in the macro-hybrid form as follows:

$$
\begin{aligned}
M_k \bar{u}_k + B_k^T \bar{p}_k + C_k^T \bar{\lambda}_k &= 0, \\
B_k \bar{u}_k - \Sigma_k \bar{p}_k &= \overline{F}_k,
\end{aligned}
\tag{22}
$$

$k = \overline{1,n}$, complemented by the continuity conditions for the normal fluxes on the interfaces $\partial E_k \cap \partial E_l$ between neighboring cells E_k and E_l, $k,l = \overline{1,n}$, and by the Neumann boundary condition for the normal fluxes on $\partial \Omega$. The vector $\bar{\lambda}_k \in \mathbb{R}^{s_k}$ represents the mean values of the solution function p on polygons $\Gamma_{k,i}$, $i = \overline{1,s_k}$, $k = \overline{1,n}$. The matrices Σ_k are diagonal blocks of the matrix Σ and the vectors $\overline{F}_k$ are subvectors of the vector $\overline{F}$ in (15). The matrices M and B in (15) can be defined by assembling of the matrices M_k and B_k in (22), respectively.

We partition the components of the vector $\bar{u}_k$ in (22) into two groups. In the first group, denoted by subindex H, we include the DOF assigned for the polygons $\Gamma_{k,i}$, $i = \overline{1,s_k}$, on the boundary of E_k, and to the second group, denoted by subindex h, we include the rest of the DOF which are interior for the cell E_k, $k = \overline{1,n}$. Then, the equations (22) can be written in the equivalent block form (the subindex k is omitted) as follows:

$$
\begin{aligned}
M_H \bar{u}_H + M_{Hh} \bar{u}_h + B_H^T \bar{p} + C^T \bar{\lambda} &= 0, \\
M_{hH} \bar{u}_H + M_h \bar{u}_h + B_h^T \bar{p} &= 0, \\
B_H \bar{u}_H + B_h \bar{u}_h - \Sigma \bar{p} &= \overline{F}.
\end{aligned}
\tag{23}
$$

At first, we consider the case when the coefficient c is a positive function in E_k, i.e. the matrix Σ_k in (22) is symmetric and positive definite, $1 \leq k \leq n$. We eliminate the vectors $\bar{u}_h$ and $\bar{p}$ from (23) in two steps. At the first step, we eliminate the vector $\bar{u}_h$ and get the system

$$
\begin{aligned}
\widetilde{M}_H \bar{u}_H + \widetilde{B}_H^T \bar{p} + C^T \bar{\lambda} &= 0, \\
\widetilde{B}_H \bar{u}_H - S_h \bar{p} &= \overline{F},
\end{aligned}
\tag{24}
$$

where

$$
\widetilde{M}_H = M_H - M_{Hh} M_h^{-1} M_{hH}, \quad \widetilde{B}_H = B_H - B_h M_h^{-1} M_{hH}, \tag{25}
$$

and

$$
S_h = B_h M_h^{-1} B_h^T + \Sigma. \tag{26}
$$

It is obvious that the matrices $\widetilde{M}_H$ and S_h are symmetric and positive definite. Moreover, the dimension of the null space of the matrix $B_h M_h^{-1} B_h^T$ equals to one, and the vector $\bar{e} = (1, \ldots, 1)^T$ belongs to the null space of this matrix ($\bar{e} \in \ker B_h^T$).

At the second step, we eliminate the vector $\bar{p}$ in (24). Then, we get the system

$$\widehat{M}\bar{u}_H + C^T\bar{\lambda} = \bar{g} \tag{27}$$

complemented by the interface and boundary conditions for the components of $\bar{u}_H$. Here,

$$\widehat{M}_H = \widetilde{M}_H + \widetilde{B}_H^T S_h^{-1} \widetilde{B}_H \tag{28}$$

and

$$\bar{g} = \widetilde{B}_H^T S_h^{-1} \overline{F}. \tag{29}$$

To analyze the matrix $\widehat{M}_H$ in (28), we consider the eigenvalue problem

$$S_h\bar{w} = \mu\Sigma\bar{w}. \tag{30}$$

Let ν be the dimension of S_h. Then problem (30) has ν positive eigenvalues

$$1 = \mu_1 < \mu_2 \leq \cdots \leq \mu_\nu \tag{31}$$

and ν corresponding Σ-orthonormal eigenvectors

$$\bar{w}_1 = \frac{1}{\sigma}\bar{e}, \quad \bar{w}_2, \ldots, \bar{w}_\nu, \tag{32}$$

where the vector $\bar{e} = \left(1, \ldots, 1\right)^T \in \mathbb{R}^\nu$ and

$$\sigma \equiv \sigma_k = \left(\int_{E_k} c\,dx\right)^{1/2}. \tag{33}$$

Thus, we get

$$S_h^{-1} = \frac{1}{\sigma^2}\bar{e}\bar{e}^T + \sum_{j=2}^{\nu} \frac{1}{\mu_j}\bar{w}_j\bar{w}_j^T \equiv \frac{1}{\sigma^2}\bar{e}\bar{e}^T + Q_h \tag{34}$$

and

$$\widehat{M}_H = M_H^0 + \frac{1}{\sigma^2}\widetilde{B}_H^T\bar{e}\bar{e}^T\widetilde{B}_H, \tag{35}$$

where the matrix

$$M_H^0 = \widetilde{M}_H + \widetilde{B}_H^T Q_h \widetilde{B}_H \tag{36}$$

is symmetric and positive definite.

Statement 1 *The equality*

$$\bar{e}^T\widetilde{B}_H = B_H^0 \tag{37}$$

holds where the matrix $B_H^0 \equiv B_H^{0,(k)}$ is defined in (11), $1 \leq k \leq n$.

To derive the required final discretization for the problem (4) we introduce the new variable $\hat{p}$ by the formula

$$\hat{p} = \frac{1}{\sigma^2}\left[\bar{e}^T\widetilde{B}_H\bar{u}_H - \bar{e}^T\overline{F}\right] \equiv \frac{1}{\sigma^2}\left[B_H^0\bar{u}_H + |E|f\right], \tag{38}$$

where

$$f = -\frac{1}{|E|}\bar{e}^T\overline{F}. \tag{39}$$

Then, we get the system in terms of $\bar{u}_H^{(k)}$ and $\hat{p}_k$ (we return the index k):

$$\begin{aligned} M_H^{0,(k)}\bar{u}_H^{(k)} + \left[B_H^{0,(k)}\right]^T\hat{p}_k + C_k^T\bar{\lambda} &= \hat{g}_k, \\ B_H^{0,(k)}\bar{u}_H^{(k)} - c_k|E_k|\hat{p}_k &= -|E_k|f_k, \end{aligned} \tag{40}$$

$k = \overline{1,n}$, complemented by the equations of continuity of normal fluxes on the interfaces between neighboring polyhedral cells and by the equations for the normal fluxes on $\partial\Omega$. Here,

$$\hat{g}_k = \bar{g}_k - \frac{1}{\sigma_k^2}\left[\widetilde{B}_H^{(k)}\right]^T\bar{e}_k\bar{e}_k^T\overline{F}_k \tag{41}$$

and the values of c_k and f_k are defined in (8). Recall that $\sigma_k^2 = c_k|E_k|$.

Now, we return to the system (23) and consider the case when the coefficient $c \equiv 0$ in E_k, i.e. Σ_k is the zero matrix. In this case, the matrix

$$S_h = B_h M_h^{-1} B_h^T \tag{42}$$

in (26) is singular.

Let us consider the eigenvalue problem

$$S_h\bar{w} = \mu D\bar{w}, \tag{43}$$

where the subindex k staying for the number of the cell $E = E_k$ is again omitted. This eigenvalue problem has one zero eigenvalue $\mu_1 = 0$ and $\nu - 1$ positive eigenvalues $\mu_2 \leq \mu_3 \leq \cdots \leq \mu_\nu$ where ν is the dimension of S_h. We denote the system of D-orthonormal eigenvectors of problem (43) by

$$\bar{w}_1, \bar{w}_2, \ldots, \bar{w}_\nu, \tag{44}$$

where

$$\bar{w}_1 = \frac{1}{|E|^{1/2}}\bar{e}. \tag{45}$$

The spectral decomposition of the matrix S_h with respect to eigenvalue problem (43) is defined by the following formula:

$$S_h = DW\Lambda W^T D, \tag{46}$$

where

$$\Lambda = \mathrm{diag}\left\{\mu_1, \mu_2, \ldots, \mu_\nu\right\} \tag{47}$$

and

$$W = \begin{bmatrix} \bar{w}_1 \ \bar{w}_2 \ \cdots \ \bar{w}_\nu \end{bmatrix}. \tag{48}$$

Consider the second equation in (24) in the form

$$S_h \bar{p} = \widetilde{B}_H \bar{u}_H - \overline{F}. \tag{49}$$

A solution vector $\bar{p}$ of this system can be presented by the formula

$$\bar{p} = S_h^+ \left[\widetilde{B}_H \bar{u}_H - \overline{F}\right] + \alpha \bar{e} \tag{50}$$

with an arbitrary coefficient $\alpha \in \mathbb{R}$ in the right-hand side and

$$S_h^+ = W \Lambda^+ W^T. \tag{51}$$

Here,

$$\Lambda^+ = \mathrm{diag}\left\{0, \mu_2^{-1}, \ldots, \mu_\nu^{-1}\right\} \tag{52}$$

is a diagonal matrix.

Substituting vector $\bar{p}$ in (50) to the second equation in (23), we get the equation

$$\left[M_{hH} + B_h^T S_h^+ \widetilde{B}_H\right]\bar{u}_H + M_h \bar{u}_h = B_h^T S_h^+ \overline{F}. \tag{53}$$

Thus,

$$\bar{u}_h = R_1 \bar{u}_H + R_2 \overline{F}, \tag{54}$$

where

$$R_1 = -M_h^{-1}\left[M_{hH} + B_h^T S_h^+ \widetilde{B}_H\right] \tag{55}$$

and

$$R_2 = M_h^{-1} B_h^T S_h^+. \tag{56}$$

Now, we replace the first two equations in (23) by a single equation. To derive this equation, we multiply the first two equations in (23) by the matrix

$$\begin{bmatrix} I_H \ R_1^T, \end{bmatrix}$$

where I_H is the identity $s_k \times s_k$ matrix, and then substitute the vector $\bar{u}_h$ defined by formula (54) into the new equation. We get the resulting equation in terms of vectors $\bar{u}_H$, $\bar{p}$, and $\bar{\lambda}$ in the following form:

$$M_H^0 \bar{u}_H + \widehat{B}_H^T \bar{p} + C^T \bar{\lambda} = \bar{g}, \tag{57}$$

where the matrix

$$M_H^0 = \begin{bmatrix} I_H \ R_1^T \end{bmatrix} \begin{bmatrix} M_H & M_{Hh} \\ M_{hH} & M_h \end{bmatrix} \begin{bmatrix} I_H \\ R_1 \end{bmatrix} \tag{58}$$

is symmetric and positive definite,

$$\widehat{B}_H^T = B_H^T + R_1^T B_h^T, \tag{59}$$

and

$$\bar{g} = -\left(M_{Hh} + R_1^T M_h\right) R_2 \overline{F}. \tag{60}$$

Let us analyze the matrix $\widehat{B}_H^T$ in (59):

$$
\begin{aligned}
\widehat{B}_H^T &= B_H^T - \left[M_{hH}^T + \widetilde{B}_H^T S_h^+ B_h\right] M_h^{-1} B_h^T = \\
&= \left(B_H^T - M_{hH}^T M_h^{-1} B_h^T\right)\left(I - S_h^+ S_h\right) = \\
&= \frac{1}{|E|} B_H^T \bar{e}\bar{e}^T D.
\end{aligned}
\tag{61}
$$

To derive the latter formula we used the identity

$$I - S_h^+ S_h = \frac{1}{|E|} \bar{e}\bar{e}^T D \tag{62}$$

and the fact that $\bar{e} \in \ker B_h^T$.

Thus, the equation (57) is equivalent to the equation

$$M_H^0 \bar{u}_H + \left[B_H^0\right]^T \hat{p} + C^T \bar{\lambda} = \bar{g} \tag{63}$$

where the matrix B_H^0 is defined in (11), i.e.

$$B_H^0 = \bar{e}^T B_H, \tag{64}$$

and

$$\hat{p} = \frac{1}{|E|} \bar{e}^T D \bar{p} \equiv \frac{1}{|E|} \int_E p_h \, \mathrm{dx} \tag{65}$$

is the mean value of p_h in the polyhedral cell E.

Complementing the equation (63) in $E \equiv E_k$ by the equation (10) with $c_k = 0$, we get the system in terms of $\bar{u}_H^{(k)}$ and $\hat{p}_k$ (we again return the index k):

$$
\begin{aligned}
M_H^{0,(k)} \bar{u}_H^{(k)} + \left[B_H^{0,(k)}\right]^T \hat{p}_k + C_k^T \bar{\lambda} &= \hat{g}_k, \\
B_H^{0,(k)} \bar{u}_H^{(k)} &= -|E_k| f_k,
\end{aligned}
\tag{66}
$$

where $M_H^{0,(k)} = M_H^0$ and M_H^0 is defined in (58). Recall that the equations (66) are derived for the case $c \equiv 0$ in E_k, $1 \le k \le n$.

Using the assembling procedure we get the system in terms of $\bar{u}_H$, $\bar{p}_H$, and the boundary Lagrange multipliers $\bar{\lambda}$:

$$
\begin{aligned}
M^0 \bar{u}_H + \left[B_H^0\right]^T \bar{p}_H + \left[C^0\right]^T \bar{\lambda} &= \bar{g}^0, \\
B_H^0 \bar{u}_H - \Sigma^0 \bar{p}_H &= \overline{F}^0, \\
C^0 \bar{u}_H &= 0.
\end{aligned}
\tag{67}
$$

The matrix M^0 in (67) is obtained by the assembling of matrices $M_H^{0,(k)}$ defined in (36) if the coefficient c is a positive function in E_k or in (58) if $c \equiv 0$ in E_k, $k = \overline{1,n}$. Respectively, the components $\hat{p}_k$ of the vector $\bar{p}_H$ in (67) are defined either in (9) if the coefficient c is a positive function in E_k or in (65) if $c \equiv 0$ in E_k, $k = \overline{1,n}$.

The elimination of $\bar{u}_H$ (condensation of the system (67)) results in the algebraic system in terms of vector $\bar{p}_H$ and the interface and boundary Lagrange multiplier vector $\bar{\lambda}$:

$$\mathcal{A} \begin{bmatrix} \bar{p}_H \\ \bar{\lambda} \end{bmatrix} = \bar{q}. \tag{68}$$

Here,

$$\mathcal{A} = \sum_{k=1}^{n} \mathcal{N}_k \mathcal{A}_k \mathcal{N}_k^T, \tag{69}$$

where

$$\mathcal{A}_k = \begin{bmatrix} c_k|E_k| & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} B_H^{0,(k)} \\ C_k \end{bmatrix} \left[M_H^{0,(k)} \right]^{-1} \left(B_H^{0,(k)} \right)^T C_k^T \tag{70}$$

are symmetric and positive definite matrices, and $\mathcal{N}_k$ are the underlying assembling matrices, $k = \overline{1,n}$. The formula for the vector $\bar{q}$ in (68) can be easily derived.

Remark 1. If the function f is constant in $E \equiv E_k$ then the vector $\overline{F}$ in (23) is defined by the formula

$$\overline{F} = -f_E D\bar{e}, \tag{71}$$

where f_E is the value of f in E, and belongs to the null space of the matrix S^+ in (51). To this end, instead of (54) we have

$$\bar{u}_h = R_1 \bar{u}_H, \tag{72}$$

and $\bar{g}$ in (57) is the zero vector. Simple analysis shows that the resulting discretization (66) is equivalent to the "div-const" discretization proposed in [KR03] (see also [KLS04, KR05]).

Remark 2. The previous remark is concerned the case when $c \equiv 0$ in $E \equiv E_k$, $1 \leq k \leq n$. Consider the case when c is a positive function in E, the diffusion equation is discretized by the method (16)–(18) and the value n_k for this cell is equal to one. Under the assumptions made, the equation (index k is omitted)

$$B_H \bar{u}_H + B_h \bar{u}_h - \widetilde{\Sigma} \bar{p} = \overline{F}_1, \tag{73}$$

where the matrix $\widetilde{\Sigma}$ and the vector $\overline{F}_1$ are defined in (20) and (21), respectively, is the underlying counterpart of the third equation in (23). Similar to (50), we can consider the following formula for the solution subvector $\bar{p}$:

$$\bar{p} = S_h^+ \left[\widetilde{B}_H \bar{u}_H - \widetilde{\Sigma} \bar{p} - \overline{F}_1 \right] + \alpha \bar{e} \tag{74}$$

with some coefficient $\alpha \in \mathbb{R}$ where

$$S_h = B_h M_h^{-1} B_h^T \tag{75}$$

and S_h^+ is defined in (51). The vectors $\widetilde{\Sigma}\bar{p}$ and $\overline{F}_1$ belong to $\ker S_h^+$. Therefore, instead of (74) we get

$$\bar{p} = S_h^+ \widetilde{B}_H \bar{u}_H + \alpha \bar{e}. \tag{76}$$

It proves that for the discretization method (16)–(18) the formula (72) is still valid, and the final discretization (66) is equivalent to the "div-const" discretization in [KR03].

References

[BF91]   F. Brezzi and M. Fortin. *Mixed and hybrid finite element methods.* Springer-Verlag, Berlin 1991

[Kuz05]   Yu. Kuznetsov. Mixed finite element method in domains of complex geometry. In *Abstract Book – 1st International Seminar of SCOMA*, number A4/2005 in Reports of the Department of Mathematical Information Technology, Series A, Collections, University of Jyväskylä, Jyväskylä, 2005.

[Kuz06]   Yu. Kuznetsov. Mixed finite element method for diffusion equations on polygonal meshes with mixed cells. *J. Numer. Math.*, 14(4):305–315, 2006

[KLS04]   Yu. Kuznetsov, K. Lipnikov, and M. Shashkov. The mimetic finite difference method on polygonal meshes for diffusion-type equations. *Comput. Geosci.*, 8:301–324, 2004

[KR03]   Yu. Kuznetsov and S. Repin. New mixed finite element method on polygonal and polyhedral meshes. *Russian J. Numer. Anal. Math. Modelling*, 18(3):261–278, 2003

[KR05]   Yu. Kuznetsov and S. Repin. Convergence analysis and error estimates for mixed finite element method on distorted meshes. *J. Numer. Math.*, **13**(1):33–51, 2005

[RT91]   J. E. Roberts and J.-M. Thomas. Mixed and hybrid methods. In P.-G. Ciarlet and J.-L. Lions, editors, *Handbook of Numerical Analysis, Vol. II*, pages 523–639. North-Holland, Amsterdam, 1991.

On the Numerical Solution of the Elliptic Monge–Ampère Equation in Dimension Two: A Least-Squares Approach

Edward J. Dean and Roland Glowinski

University of Houston, Department of Mathematics, 651 P. G. Hoffman Hall, Houston, TX 77204-3008, USA `roland@math.uh.edu, dean@math.uh.edu`

1 Introduction

During his outstanding career, *Olivier Pironneau* has addressed the solution of a large variety of problems from the Natural Sciences, Engineering and Finance to name a few, an evidence of his activity being the many articles and books he has written. It is the opinion of these authors, and former collaborators of O. Pironneau (cf. [DGP91]), that this chapter is well-suited to a volume honoring him. Indeed, the two pillars of the solution methodology that we are going to describe are: (1) a nonlinear least squares formulation in an appropriate Hilbert space, and (2) a mixed finite element approximation, reminiscent of the one used in [DGP91] and [GP79] for solving the Stokes and Navier–Stokes equations in their stream function-vorticity formulation; the contributions of O. Pironneau on the two above topics are well-known world wide. Last but not least, we will show that the solution method discussed here can be viewed as a solution method for a non-standard variant of the incompressible Navier–Stokes equations, an area where O. Pironneau has many outstanding and celebrated contributions (cf. [Pir89], for example).

The main goal of this article is to discuss the *numerical solution* of the *Dirichlet problem* for the prototypical *two-dimensional elliptic Monge–Ampère equation*, namely

$$\det \mathbf{D}^2\psi = f \text{ in } \Omega, \quad \psi = g \text{ on } \Gamma. \qquad \text{(E-MA-D)}$$

In (E-MA-D): (1) Ω is a bounded domain of $\mathbb{R}^2$ and Γ is its boundary; (2) f and g are given functions with $f > 0$; $\mathbf{D}^2\psi = (\partial^2\psi/\partial x_i\partial x_j)_{1\le i,j\le 2}$ is the *Hessian* of the unknown function ψ. The partial differential equation in (E-MA-D) is a *fully nonlinear elliptic* one (in the sense of, e.g., Gilbarg and Trudinger [GT01] and Caffarelli and Cabré [CC95]). The *mathematical analysis* of problems such as (E-MA-D) has produced a quite abundant literature; let us mention, among many others, [GT01, CC95, Aub82, Aub98, Cab02] and the references therein. On the other hand, and to the best of our knowledge, the *numerical analysis* community has largely ignored these problems,

so far, some notable exceptions being provided by [BB00, OP88, CKO99] (see also [DG03, DG04]). Indeed we can not resist quoting [BB00] (an article dedicated to the numerical solution of the celebrated *Monge–Kantorovitch optimal transportation problem*):

> "It follows from this theoretical result that a natural computational solution of the L2 MKP is the numerical resolution of the Monge–Ampère equation (6). Unfortunately, this fully nonlinear second-order elliptic equation has not received much attention from numerical analysts and, to the best of our knowledge, there is no efficient finite-difference or finite-element methods, comparable to those developed for linear second-order elliptic equations (such as fast Poisson solvers, multigrid methods, preconditioned conjugate gradient methods, . . .)."

We will show in this article that, actually, fully nonlinear elliptic problems such as (E-MA-D) can be solved by appropriate combinations of fast Poisson solvers and preconditioned conjugate gradient methods. However, unlike the (*closely related*) Dirichlet problem for the Laplace operator, the problem (E-MA-D) may have *multiple solutions* (actually, two at most; cf., e.g., [CH89, Chapter 4]), and the *smoothness* of the data does not imply the existence of a smooth solution. Concerning the last property, suppose that $\Omega = (0, 1) \times (0, 1)$ and consider the special case where (E-MA-D) is defined by

$$\frac{\partial^2 \psi}{\partial x_1^2} \frac{\partial^2 \psi}{\partial x_2^2} - \left| \frac{\partial^2 \psi}{\partial x_1 \partial x_2} \right|^2 = 1 \quad \text{in } \Omega, \qquad \psi = 0 \text{ on } \Gamma. \tag{1}$$

The problem (1) can not have smooth solutions since, for those solutions, the boundary condition $\psi = 0$ on Γ implies that the product $(\partial^2 \psi / \partial x_1^2)(\partial^2 \psi / \partial x_2^2)$ and the cross-derivative $\partial^2 \psi / \partial x_1 \partial x_2$ vanish at the boundary, implying in turn that $\det \mathbf{D}^2 \psi$ is strictly less than one in some neighborhood of Γ. The above (non-existence) result is not a consequence of the non-smoothness of Γ, since a similar non-existence property holds if in (1) one replaces the above Ω by the ovoïd-shaped domain whose C^∞-boundary is defined by

$$\Gamma = \bigcup_{i=1}^4 \Gamma_i,$$

with

$$\Gamma_1 = \{x \mid x = \{x_1, x_2\}, \ x_2 = 0, \ 0 \le x_1 \le 1\},$$
$$\Gamma_3 = \{x \mid x = \{x_1, x_2\}, \ x_2 = 1, \ 0 \le x_1 \le 1\},$$
$$\Gamma_2 = \{x \mid x = \{x_1, x_2\}, \ x_1 = 1 - \ln 4/(\ln x_2(1 - x_2)), \ 0 \le x_2 \le 1\},$$
$$\Gamma_4 = \{x \mid x = \{x_1, x_2\}, \ x_1 = \ln 4/(\ln x_2(1 - x_2)), \ 0 \le x_2 \le 1\}.$$

Actually, for the above two Ωs the non-existence of solutions for the problem (1) follows from the *non-strict convexity* of these domains. Albeit the problem

(1) has no classical solution it has *viscosity solutions* in the sense of Crandall–Lions, as shown in, e.g., [CC95, Cab02, Jan88, Urb88, CIL92]. The Crandall–Lions viscosity approach relies heavily on the *maximum principle*, unlike the *variational methods* used to solve, for example, the second order linear elliptic equations in divergence form in some appropriate subspace of the Hilbert space $H^1(\Omega)$. The *least-squares* approach discussed in this article operates in the space $H^2(\Omega) \times \mathbf{Q}$ where $\mathbf{Q}$ is the Hilbert space of the 2×2 symmetric tensor-valued functions with component in $L^2(\Omega)$. Combined with *mixed finite element approximations* and *operator-splitting methods* it will have the ability, if g has the $H^{3/2}(\Gamma)$-regularity, to capture classical solutions, if such solutions exist, and to compute *generalized solutions* to problems like (1) which have no classical solution. Actually, we will show that these generalized solutions are also *viscosity solutions*, but in a sense different from Crandall–Lions'.

Remark 1. Suppose that Ω is simply connected. Let us define a vector-valued function $\mathbf{u}$ by $\mathbf{u} = \{\frac{\partial \psi}{\partial x_2}, -\frac{\partial \psi}{\partial x_1}\}$ ($= \{u_1, u_2\}$). The problem (E-MA-D) takes then the equivalent formulation

$$\begin{cases} \det \boldsymbol{\nabla}\mathbf{u} = f & \text{in } \Omega, \quad \boldsymbol{\nabla} \cdot \mathbf{u} = 0 \text{ in } \Omega, \\ \quad \mathbf{u} \cdot \mathbf{n} = \dfrac{dg}{ds} & \text{on } \Gamma, \end{cases} \tag{2}$$

where $\mathbf{n}$ denotes the outward unit vector normal at Γ, and s is a counter-clockwise curvilinear abscissa. Once $\mathbf{u}$ is known, one obtains ψ via the solution of the following Poisson–Dirichlet problem:

$$-\triangle \psi = \frac{\partial u_2}{\partial x_1} - \frac{\partial u_1}{\partial x_2} \quad \text{in } \Omega, \quad \psi = g \text{ on } \Gamma.$$

The problem (2) has clearly an *incompressible fluid flow flavor*, ψ playing here the role of a *stream function*. The relations (2) can be used to solve the problem (E-MA-D) but this approach will not be further investigated here.

Remark 2. As shown in [DG05], the methodology discussed in this article applies also (among other problems) to the *Pucci–Dirichlet problem*

$$\alpha \lambda^+ + \lambda^- = 0 \quad \text{in } \Omega, \qquad \psi = g \text{ on } \Gamma, \tag{PUC-D}$$

with λ^+ (resp., λ^-) the *largest* (resp., the *smallest*) *eigenvalue* of $\mathbf{D}^2\psi$ and $\alpha \in (1, +\infty)$. (If $\alpha = 1$, one recovers the linear Poisson–Dirichlet problem.)

Remark 3. A shortened version of this article can be found in [DG04].

Remark 4. The solution of (E-MA-D) by *augmented Lagrangian methods* is discussed in [DG03, DG06a, DG06b].

2 A Least Squares Formulation of the Problem (E-MA-D)

From now on, we suppose that $f > 0$ and that $\{f, g\} \in \{L^1(\Omega), H^{3/2}(\Gamma)\}$, implying that the following space and set are *non-empty*:

$$V_g = \{\varphi \mid \varphi \in H^2(\Omega),\ \varphi = g \text{ on } \partial\Omega\},$$
$$\mathbf{Q}_f = \{\mathbf{q} \mid \mathbf{q} \in \mathbf{Q},\ \det \mathbf{q} = f\},$$

with

$$\mathbf{Q} = \{\mathbf{q} \mid \mathbf{q} \in (L^2(\Omega))^{2\times 2},\ \mathbf{q} = \mathbf{q}^t\}.$$

Solving the Monge–Ampère equation in $H^2(\Omega)$ is equivalent to looking for the intersection in $\mathbf{Q}$ of the two sets $\mathbf{D}^2 V_g$ and $\mathbf{Q}_f$, an infinite dimensional geometry problem "visualized" in Figures 1 and 2.

If $\mathbf{D}^2 V_g \cap \mathbf{Q}_f \neq \emptyset$ as "shown" in Figure 1, then the problem (E-MA-D) has a solution in $H^2(\Omega)$. If, on the other hand, it is the situation of Figure 2 which prevails, namely $\mathbf{D}^2 V_g \cap \mathbf{Q}_f = \emptyset$, (E-MA-D) has no solution in $H^2(\Omega)$. However, Figure 2 is *constructive* in the sense that it suggests looking for a pair $\{\psi, \mathbf{p}\}$ which *minimizes*, globally or locally, some distance between $\mathbf{D}^2\varphi$ and $\mathbf{q}$ when $\{\varphi, \mathbf{q}\}$ describes the set $V_g \times \mathbf{Q}_f$.

According to the above suggestion, and in order to handle those situations where (E-MA-D) has no solution in $H^2(\Omega)$, despite the fact that neither V_g nor $\mathbf{Q}_f$ are empty, we suggest to solve the above problem via the following (nonlinear) *least squares formulation*:

$$\begin{cases} \text{Find } \{\psi, \mathbf{p}\} \in V_g \times \mathbf{Q}_f \text{ such that} \\ j(\psi, \mathbf{p}) \leq j(\varphi, \mathbf{q}),\ \forall \{\varphi, \mathbf{q}\} \in V_g \times \mathbf{Q}_f, \end{cases} \qquad \text{(LSQ)}$$

where, in (LSQ) and below, we have (with $dx = dx_1 dx_2$):

$$j(\varphi, \mathbf{q}) = \tfrac{1}{2} \int_\Omega |\mathbf{D}^2\varphi - \mathbf{q}|^2 \, dx \qquad (3)$$

and

$$|\mathbf{q}| = (q_{11}^2 + q_{22}^2 + 2q_{12}^2)^{1/2}, \quad \forall \mathbf{q}(= (q_{ij})_{1 \leq i,j \leq 2}) \in \mathbf{Q}. \qquad (4)$$
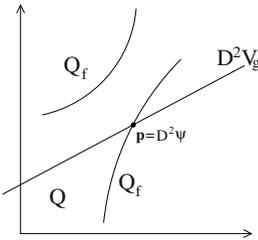

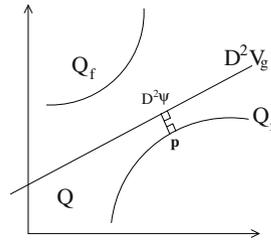**Fig. 1.** Problem (E-MA-D) has a solution in $H^2(\Omega)$.

Fig. 2. Problem (E-MA-D) has no solution in $H^2(\Omega)$.

Remark 5. The results (described in [DG05]), concerning the numerical solution of the Pucci's problem (PUC-D) (see Remark 2), suggest that defining $|\mathbf{q}|$ by

$$|\mathbf{q}| = (q_{11}^2 + q_{22}^2 + q_{12}^2)^{1/2}, \quad \forall \mathbf{q}(= (q_{ij})_{1 \le i,j \le 2}) \in \mathbf{Q}, \tag{5}$$

instead of (4), may improve the convergence of the algorithms to be described in the following sections. We intend to check this conjecture in a near future.

In order to solve (LSQ) by *operator-splitting techniques* it is convenient to observe that (LSQ) is *equivalent* to

$$\begin{cases} \{\psi, \mathbf{p}\} \in V_g \times \mathbf{Q}, \\ j_f(\psi, \mathbf{p}) \le j_f(\varphi, \mathbf{q}), \quad \forall \{\varphi, \mathbf{q}\} \in V_g \times \mathbf{Q}, \end{cases} \tag{LSQ-P}$$

where

$$j_f(\varphi, \mathbf{q}) = j(\varphi, \mathbf{q}) + I_f(\mathbf{q}), \quad \forall \{\varphi, \mathbf{q}\} \in V_g \times \mathbf{Q}, \tag{6}$$

with

$$I_f(\mathbf{q}) = \begin{cases} 0, & \text{if } \mathbf{q} \in \mathbf{Q}_f, \\ +\infty, & \text{if } \mathbf{q} \in \mathbf{Q} \setminus \mathbf{Q}_f, \end{cases}$$

i.e., $I_f(\cdot)$ is the *indicator functional* of the set $\mathbf{Q}_f$.

3 An Operator-Splitting Based Method for the Solution of (E-MA-D) via (LSQ-P)

We can solve the least-squares problem (LSQ) by a *block relaxation method* operating *alternatively* between V_g and Q_f. Such relaxation algorithms are discussed in, e.g., [Glo84]. Closely related algorithms are obtained as follows:

Step 1. Derive the *Euler-Lagrange equation* of (LSQ-P).
Step 2. Associate to the above *Euler-Lagrange* equation an *initial value problem (flow* in the Dynamical System terminology) in $V_g \times \mathbf{Q}$.
Step 3. Use *operator-splitting* to time discretize the above flow problem.

Applying the above program, Step 1 provides us with the *Euler–Lagrange equation* of the problem (LSQ-P). A *variational formulation* of this equation reads as follows:

$$\begin{cases} \{\psi, \mathbf{p}\} \in V_g \times \mathbf{Q}, \\ \int_\Omega (\mathbf{D}^2 \psi - \mathbf{p}) : (\mathbf{D}^2 \varphi - \mathbf{q}) \, dx + \langle \partial I_f(\mathbf{p}), \mathbf{q} \rangle = 0, \quad \forall \{\varphi, \mathbf{q}\} \in V_0 \times \mathbf{Q}, \end{cases} \tag{7}$$

where $\partial I_f(\mathbf{p})$ denotes a *generalized differential* of the functional $I_f(\cdot)$ at $\mathbf{p}$. Next, we have denoted by $\mathbf{S} : \mathbf{T}$ the *Fröbenius scalar product* of the two 2×2 symmetric tensors $\mathbf{S} \ (= (s_{ij}))$ and $\mathbf{T} \ (= (t_{ij}))$, namely

$$\mathbf{S} : \mathbf{T} = s_{11}t_{11} + s_{22}t_{22} + 2s_{12}t_{12}$$

and, finally,

$$V_0 = H^2(\Omega) \cap H_0^1(\Omega).$$

Next, we achieve Step 2 by associating with (7) the following *initial value problem* (flow), written in *semi-variational form*:

$$
\begin{cases}
\text{Find } \{\psi(t), \mathbf{p}(t)\} \in V_g \times \mathbf{Q} \text{ for all } t > 0 \text{ such that} \\
\displaystyle\int_\Omega [\partial(\triangle\psi)/\partial t]\triangle\varphi \, dx + \int_\Omega \mathbf{D}^2\psi : \mathbf{D}^2\varphi \, dx = \int_\Omega \mathbf{p} : \mathbf{D}^2\varphi \, dx, \quad \forall \varphi \in V_0, \\
\partial\mathbf{p}/\partial t + \mathbf{p} + \partial I_f(\mathbf{p}) = \mathbf{D}^2\psi, \\
\{\psi(0), \mathbf{p}(0)\} = \{\psi_0, \mathbf{p}_0\},
\end{cases}
$$

(8)

and we look at the limit of $\{\psi(t), \mathbf{p}(t)\}$ as $t \to +\infty$. The choice of ψ^0 and $\mathbf{p}^0$ will be discussed in Remark 6.

Finally, concerning Step 3 we advocate the following *operator-splitting scheme* (à la *Marchuk–Yanenko*, see, e.g., [Glo03, Chapter 6] and the references therein), but we acknowledge that other splitting schemes are possible:

$$\{\psi^0, \mathbf{p}^0\} = \{\psi_0, \mathbf{p}_0\}.$$

(9)

Then, for $n \geq 0$, $\{\psi^n, \mathbf{p}^n\}$ being known, we obtain $\{\psi^{n+1}, \mathbf{p}^{n+1}\}$ from the solution of

$$(\mathbf{p}^{n+1} - \mathbf{p}^n)/\tau + \mathbf{p}^{n+1} + \partial I_f(\mathbf{p}^{n+1}) = \mathbf{D}^2\psi^n,$$

(10)

$$
\begin{cases}
\psi^{n+1} \in V_g; \\
\displaystyle\int_\Omega \triangle\left[\left(\psi^{n+1} - \psi^n\right)/\tau\right] \triangle\varphi \, dx + \int_\Omega \mathbf{D}^2\psi^{n+1} : \mathbf{D}^2\varphi \, dx = \\
= \displaystyle\int_\Omega \mathbf{p}^{n+1} : \mathbf{D}^2\varphi \, dx, \quad \forall \varphi \in V_0;
\end{cases}
$$

(11)

above, $\tau \, (> 0)$ is a *time-discretization step*.

The solution of the sub-problems (10) and (11) will be discussed in Sections 4 and 5, respectively.

Remark 6. The initialization of the flow defined by (8) and of its time-discrete variant defined by (9)–(11) are clearly important issues. Let us denote by λ_1 and λ_2 the *eigenvalues* of the Hessian $\mathbf{D}^2\psi$. It follows from (E-MA-D) that $\lambda_1\lambda_2 = f$, implying in turn that

$$\sqrt{\lambda_1\lambda_2} = \sqrt{f}.$$

(12)

We have, on the other hand,

$$|\triangle\psi| = |\lambda_1 + \lambda_2|.$$

(13)

Suppose that we look for a *convex solution* of (E-MA-D). We have then λ_1 and λ_2 positive. Comparing (12) (geometric mean) and (13) (arithmetic mean) suggests to define ψ_0 as the solution of

$$\triangle\psi_0 = 2\sqrt{f} \quad \text{in } \Omega, \quad \psi_0 = g \text{ on } \Gamma. \tag{14}$$

If we look for a *concave solution* we suggest to define ψ_0 as the solution of

$$-\triangle\psi_0 = 2\sqrt{f} \quad \text{in } \Omega, \quad \psi_0 = g \text{ on } \Gamma. \tag{15}$$

If $\{f, g\} \in L^1(\Omega) \times H^{3/2}(\Gamma)$, then $\{\sqrt{f}, g\} \in L^2(\Omega) \times H^{3/2}(\Gamma)$, implying that each of the problems (14) and (15) has a unique solution in V_g (assuming of course that Ω is convex and/or that Γ is sufficiently smooth). Concerning $\mathbf{p}^0$ an obvious choice is provided by

$$\mathbf{p}_0 = \mathbf{D}^2\psi_0, \tag{16}$$

another possibility being

$$\mathbf{p}_0 = \begin{pmatrix} \sqrt{f} & 0 \\ 0 & \sqrt{f} \end{pmatrix}. \tag{17}$$

The symmetric tensor defined by (17) belongs clearly to $\mathbf{Q}_f$.

4 On the Solution of the Nonlinear Sub-Problems (10)

Concerning the solution of the sub-problems of type (10), we interpret (10) as the *Euler–Lagrange* equation of the following minimization problem:

$$\begin{cases} \mathbf{p}^{n+1} \in \mathbf{Q}_f, \\ J_n(\mathbf{p}^{n+1}) \le J_n(\mathbf{q}), \quad \forall \mathbf{q} \in \mathbf{Q}_f, \end{cases} \tag{18}$$

with

$$J_n(\mathbf{q}) = \frac{1}{2}(1+\tau) \int_\Omega |\mathbf{q}|^2 \, dx - \int_\Omega (\mathbf{p}^n + \tau\mathbf{D}^2\psi^n) : \mathbf{q} \, dx. \tag{19}$$

It follows from (19) that the problem (18) can be solved point-wise on Ω (in practice, at the grid points of a finite element or finite difference mesh). To be more precise, we have to solve, *a.e.* on Ω, a minimization problem of the following type:

$$\begin{cases} \min_{\mathbf{z}} \left[\frac{1}{2}(z_1^2 + z_2^2 + 2z_3^2) - b_1(x)z_1 - b_2(x)z_2 - 2b_3(x)z_3 \right] \\ \text{with } \mathbf{z} \left(= \{z_i\}_{i=1}^3 \right) \in \left\{ \mathbf{z} \mid \mathbf{z} \in \mathbb{R}^3, \; z_1z_2 - z_3^2 = f(x) \right\}. \end{cases} \tag{20}$$

Actually, if one looks for *convex* (resp., *concave*) *solutions* of (E-MA-D), we should prescribe the following additional constraints: $z_1 \ge 0, z_2 \ge 0$ (resp., $z_1 \le 0, z_2 \le 0$). For the solution of the problem (20) (a constrained

minimization problem in R^3) we advocate those methods discussed in, e.g., [DS96] (after introduction of a Lagrange multiplier to handle the constraint $z_1 z_2 - z_3^2 = f(x)$). Other methods are possible, including the reduction of (20) to a two-dimensional problem via the elimination of z_3. Indeed, we observe that (20) is equivalent to

$$\begin{cases} \min_{\mathbf{z}} \left[\frac{1}{2}(z_1 + z_2)^2 - b_1(x)z_1 - b_2(x)z_2 - 2|b_3(x)|(z_1 z_2 - f(x))^{\frac{1}{2}} \right] \\ \text{with } \mathbf{z}(= \{z_i\}_{i=1}^3) \in \left\{ \mathbf{z} \mid \mathbf{z} \in \mathbb{R}^3,\ z_1 z_2 - f(x) \geq 0, \right. \\ \qquad z_3 = \mathrm{sgn}(b_3(x))(z_1 z_2 - f(x))^{\frac{1}{2}} \Big\}, \end{cases} \qquad (21)$$

which leads to the above mentioned reduction; then we make "almost" trivial the solution of the problem (21) by using the following change of variables (reminiscent of the polar coordinate based technique used in [DG05] for the solution of the Pucci's equation (PUC-D), introduced in Remark 2):

$$z_1 = \rho\sqrt{f}e^{\theta}, \quad z_2 = \rho\sqrt{f}e^{-\theta},$$

with $\theta \in R$ and $\rho \geq 1$ (resp., $\rho \leq -1$) if one looks for a convex (resp., concave) solution of (E-MA-D).

5 On the Conjugate Gradient Solution of the Linear Sub-Problems (11)

The sub-problems (11) are all members of the following family of *linear variational problems*:

$$\begin{cases} u \in V_g, \\ \displaystyle\int_\Omega \triangle u \triangle v\, dx + \tau \int_\Omega \mathbf{D}^2 u : \mathbf{D}^2 v\, dx = L(v), \quad \forall v \in V_0, \end{cases} \qquad (22)$$

with the functional L linear and continuous from $H^2(\Omega)$ into $\mathbb{R}$; the problems in (22) are clearly of the *biharmonic* type. The *conjugate gradient* solution of linear variational problems in Hilbert spaces, such as (22), has been addressed in, e.g., [Glo03, Chapter 3]. Following the above reference, we are going to solve (22) by a conjugate gradient algorithm operating in the spaces V_0 and V_g, both spaces being equipped with the scalar product defined by

$$\{v, w\} \rightarrow \int_\Omega \triangle v \triangle w\, dx,$$

and the corresponding norm. This conjugate gradient algorithm reads as follows:

Algorithm 1

Step 1. u^0 is given in V_g.

Step 2. Solve then

$$
\begin{cases}
g^0 \in V_0, \\
\displaystyle\int_\Omega \triangle g^0 \triangle v \, dx = \int_\Omega \triangle u^0 \triangle v \, dx + \tau \int_\Omega \mathbf{D}^2 u^0 : \mathbf{D}^2 v \, dx - L(v), \\
\forall v \in V_0,
\end{cases}
$$
(23)

and set $w^0 = g^0$.

Step 3. Then, for $k \geq 0$, u^k, g^k, w^k being known, the last two different from 0, we compute u^{k+1}, g^{k+1}, and if necessary w^{k+1}, as follows: Solve

$$
\begin{cases}
\bar{g}^k \in V_0, \\
\displaystyle\int_\Omega \triangle \bar{g}^k \triangle v \, dx = \int_\Omega \triangle w^k \triangle v \, dx + \tau \int_\Omega \mathbf{D}^2 w^k : \mathbf{D}^2 v \, dx, \\
\forall v \in V_0,
\end{cases}
$$
(24)

and compute

$$
\rho_k = \frac{\int_\Omega |\triangle g^k|^2 \, dx}{\int_\Omega \triangle \bar{g}^k \triangle w^k \, dx},
$$
(25)

$$
u^{k+1} = u^k - \rho_k w^k,
$$
(26)

$$
g^{k+1} = g^k - \rho_k \bar{g}^k.
$$
(27)

Step 4. If $\int_\Omega |\triangle g^{k+1}|^2 \, dx / \int_\Omega |\triangle g^0|^2 \, dx \leq \mathrm{tol}$ take $u = u^{k+1}$; else, compute

$$
\gamma_k = \frac{\int_\Omega |\triangle g^{k+1}|^2 \, dx}{\int_\Omega |\triangle g^k|^2 \, dx}
$$
(28)

and

$$
w^{k+1} = g^{k+1} + \gamma_k w^k.
$$
(29)

Step 5. Do $k = k + 1$ and return to Step 3.

Numerical experiments have shown that Algorithm 1 (in fact, its discrete variants) has excellent convergence properties when applied to the solution of (E-MA-D). Combined with an appropriate mixed finite element approximation of (E-MA-D) it requires the solution of two discrete Poisson problems at each iteration.

6 On a Mixed Finite Element Approximation of the Problem (E-MA-D)

6.1 Generalities

Considering the highly *variational* flavor of the methodology discussed in Sections 2 to 5, it makes sense to look for *finite element* based methods for the approximation of (E-MA-D). In order to avoid the complications associated to the construction of finite element subspaces of $H^2(\Omega)$, we will employ a *mixed finite element approximation* (closely related to those discussed in, e.g., [DGP91, GP79] for the solution of linear and nonlinear *biharmonic* problems). Following this approach, it will be possible to solve (E-MA-D) employing approximations commonly used for the solution of the second order elliptic problems (piecewise linear and globally continuous over a triangulation of Ω, for example).

6.2 A Mixed Finite Element Approximation

For simplicity, we suppose that Ω is a bounded polygonal domain of $\mathbb{R}^2$. Let us denote by $\mathcal{T}_h$ a *finite element triangulation* of Ω (like those discussed in, e.g., [Glo84, Appendix 1]). From $\mathcal{T}_h$ we approximate spaces $L^2(\Omega)$, $H^1(\Omega)$ and $H^2(\Omega)$ by the finite dimensional space V_h defined by

$$V_h = \{v \mid v \in C^0(\bar{\Omega}),\ v|_T \in P_1,\ \forall T \in \mathcal{T}_h\}, \tag{30}$$

with P_1 the space of the two-variable polynomials of degree ≤ 1. A function φ being given in $H^2(\Omega)$ we denote $\frac{\partial^2 \varphi}{\partial x_i \partial x_j}$ by $D_{ij}^2(\varphi)$. It follows from *Green's formula* that

$$\int_\Omega \frac{\partial^2 \varphi}{\partial x_i^2} v\, dx = -\int_\Omega \frac{\partial \varphi}{\partial x_i} \frac{\partial v}{\partial x_i}\, dx, \qquad \forall v \in H_0^1(\Omega), \quad \forall i = 1, 2, \tag{31}$$

$$\int_\Omega \frac{\partial^2 \varphi}{\partial x_1 \partial x_2} v\, dx = -\frac{1}{2} \int_\Omega \left[\frac{\partial \varphi}{\partial x_1} \frac{\partial v}{\partial x_2} + \frac{\partial \varphi}{\partial x_2} \frac{\partial v}{\partial x_1} \right] dx, \quad \forall v \in H_0^1(\Omega). \tag{32}$$

Consider now $\varphi \in V_h$. Taking advantage of the relations (31) and (32), we define the discrete analogues of the differential operators D_{ij}^2 by

$$\begin{cases} \forall i = 1, 2, \quad D_{hii}^2(\varphi) \in V_{0h}, \\ \displaystyle\int_\Omega D_{hii}^2(\varphi) v\, dx = -\int_\Omega \frac{\partial \varphi}{\partial x_i} \frac{\partial v}{\partial x_i}\, dx, \quad \forall v \in V_{0h}, \end{cases} \tag{33}$$

$$\begin{cases} D_{h12}^2(\varphi) \in V_{0h}, \\ \displaystyle\int_\Omega D_{h12}^2(\varphi) v\, dx = -\frac{1}{2} \int_\Omega \left[\frac{\partial \varphi}{\partial x_1} \frac{\partial v}{\partial x_2} + \frac{\partial \varphi}{\partial x_2} \frac{\partial v}{\partial x_1} \right] dx, \quad \forall v \in V_{0h}, \end{cases} \tag{34}$$

where the space V_{0h} is defined by

$$V_{0h} = V_h \cap H_0^1(\Omega) \ (= \{v \mid v \in V_h, \ v = 0 \text{ on } \Gamma\}). \tag{35}$$

The functions $D_{hij}^2(\Omega)$ are *uniquely* defined by the relations (33) and (34). However, in order to simplify the computation of the above discrete second order partial derivatives we will use the *trapezoidal rule* to evaluate the integrals in the left hand sides of (33) and (34). Owing to their practical importance, let us detail these calculations:

1. First we introduce the set Σ_h of the vertices of $\mathcal{T}_h$ and then $\Sigma_{0h} = \{P \mid P \in \Sigma_h, \ P \notin \Gamma\}$. Next, we define the integers N_h and N_{0h} by $N_h = \text{Card}(\Sigma_h)$ and $N_{0h} = \text{Card}(\Sigma_{0h})$. We have then $\dim V_h = N_h$ and $\dim V_{0h} = N_{0h}$. We suppose that $\Sigma_{0h} = \{P_k\}_{k=1}^{N_{0h}}$ and $\Sigma_h = \Sigma_{0h} \cup \{P_k\}_{k=N_{0h}+1}^{N_h}$.

2. To $P_k \in \Sigma_h$ we associate the function w_k *uniquely* defined by

$$w_k \in V_h, \ w_k(P_k) = 1, \ w_k(P_l) = 0, \ \text{if } l = 1, \cdots N_h, \ l \neq k. \tag{36}$$

It is well known (see, e.g., [Glo84, Appendix 1]) that the sets $\mathcal{B}_h = \{w_k\}_{k=1}^{N_h}$ and $\mathcal{B}_{0h} = \{w_k\}_{k=1}^{N_{0h}}$ are *vector bases* of V_h and V_{0h}, respectively.

3. Let us denote by A_k the area of the polygonal which is the union of those triangles of $\mathcal{T}_h$ which have P_k as a common vertex. Applying the trapezoidal rule to the integrals in the left hand side of the relations (33) and (34), we obtain:

$$\begin{cases} \forall i = 1, 2, \ D_{hii}^2(\varphi) \in V_{0h}, \\ D_{hii}^2(\varphi)(P_k) = -\dfrac{3}{A_k} \displaystyle\int_\Omega \dfrac{\partial \varphi}{\partial x_i} \dfrac{\partial w_k}{\partial x_i} \, dx, \quad \forall k = 1, 2, \cdots, N_{0h}, \end{cases} \tag{37}$$

$$\begin{cases} D_{h12}^2(\varphi)(= D_{h21}^2(\varphi)) \in V_{0h}, \\ D_{h12}^2(\varphi)(P_k) = -\dfrac{3}{2A_k} \displaystyle\int_\Omega \left[\dfrac{\partial \varphi}{\partial x_1} \dfrac{\partial w_k}{\partial x_2} + \dfrac{\partial \varphi}{\partial x_2} \dfrac{\partial w_k}{\partial x_1} \right] dx, \\ \forall k = 1, 2, \cdots, N_{0h}. \end{cases} \tag{38}$$

Computing the integrals in the right hand sides of (37) and (38) is quite simple since the first order derivatives of φ and w_k are *piecewise constant*.

Taking the above relations into account, approximating (E-MA-D) is now a fairly simple issue. Assuming that the boundary function g is *continuous* over Γ, we approximate the affine space V_g by

$$V_{gh} = \{\varphi \mid \varphi \in V_h, \ \varphi(P) = g(P), \ \forall P \in \Sigma_h \cap \Gamma\}, \tag{39}$$

and then (E-MA-D) by

$$\begin{cases} \text{Find } \psi_h \in V_{gh} \text{ such that for all } k = 1, 2, \ldots, N_{0h}, \\ D_{h11}^2(\psi_h)(P_k) D_{h22}^2(\psi_h)(P_k) - |D_{h12}^2(\psi_h)(P_k)|^2 = f_h(P_k). \end{cases} \tag{E-MA-D}_h$$

The *iterative solution* of the problem (E-MA-D)$_h$ will be discussed in the following paragraph.
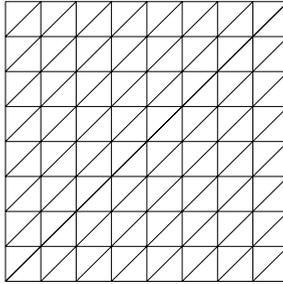
Fig. 3. A uniform triangulation of $\Omega = (0,1)^2$ $(h = 1/8)$

Remark 7. Suppose that $\Omega = (0,1)^2$ and that triangulation $\mathcal{T}_h$ is like the one shown in Figure 3.

Suppose that $h = \frac{1}{I+1}$, I being a positive integer greater than 1. In this particular case, the sets Σ_h and Σ_{0h} are given by

$$\begin{cases} \Sigma_h = \{P_{ij} \mid P_{ij} = \{ih, jh\},\ 0 \leq i, j \leq I+1\}, \\ \Sigma_{0h} = \{P_{ij} \mid P_{ij} = \{ih, jh\},\ 1 \leq i, j \leq I\}, \end{cases} \tag{40}$$

implying that $N_h = (I+2)^2$ and $N_{0h} = I^2$. It follows then from the relations (37) and (38) that (with obvious notation):

$$D^2_{h11}(\varphi)(P_{ij}) = \frac{\varphi_{i+1,j} + \varphi_{i-1,j} - 2\varphi_{ij}}{h^2}, \quad 1 \leq i, j \leq I, \tag{41}$$

$$D^2_{h22}(\varphi)(P_{ij}) = \frac{\varphi_{i,j+1} + \varphi_{i,j-1} - 2\varphi_{ij}}{h^2}, \quad 1 \leq i, j \leq I, \tag{42}$$

and

$$D^2_{h12}(\varphi)(P_{ij}) = \frac{(\varphi_{i+1,j+1} + \varphi_{i-1,j-1} + 2\varphi_{ij})}{2h^2}$$
$$- (\varphi_{i+1,j} + \varphi_{i-1,j} + \varphi_{i,j+1} + \varphi_{i,j-1})/(2h^2), \quad 1 \leq i, j \leq I. \tag{43}$$

The finite difference formulas (41)–(43) are exact for the polynomials of degree ≤ 2. Also, as expected,

$$D^2_{h11}(\varphi)(P_{ij}) + D^2_{h22}(\varphi)(P_{ij}) = \frac{\varphi_{i+1,j} + \varphi_{i-1,j} + \varphi_{i,j+1} + \varphi_{i,j-1} - 4\varphi_{ij}}{h^2}; \tag{44}$$

we have recovered, thus, the well-known 5-point discretization formula for the finite difference approximation of the Laplace operator.

6.3 On the Least-squares Formulation of (E-MA-D)$_h$

Inspired by Sections 3 to 5, we will discuss now the solution of (E-MA-D)$_h$ by a discrete variant of the solution methods discussed there. The first step in

this direction is to approximate the least-squares problem (LSQ). To achieve this goal, we approximate the sets $\mathbf{Q}$ and $\mathbf{Q}_f$ by

$$\mathbf{Q}_h = \{\mathbf{q} \mid \mathbf{q} = (q_{ij})_{1 \leq i,j \leq 2}, \ q_{21} = q_{12}, \ q_{ij} \in V_{0h}\} \tag{45}$$

and

$$\mathbf{Q}_{fh} = \{\mathbf{q} \mid \mathbf{q} \in \mathbf{Q}_h, \ q_{11}(P_k)q_{22}(P_k) - |q_{12}(P_k)|^2 = f_h(P_k),$$
$$\forall k = 1, 2, \ldots, N_{0h}\}, \tag{46}$$

respectively, the function f_h in (46) (and in $(\text{E-MA-D})_h$) being a continuous approximation of f. Next, we approximate the *least-squares functional* $j(\cdot, \cdot)$ (defined by (3) in Section 2) by $j_h(\cdot, \cdot)$ defined as follows:

$$j_h(\varphi, \mathbf{q}) = \tfrac{1}{2}\|\mathbf{D}_h^2\varphi - \mathbf{q}\|_h^2, \quad \forall \varphi \in V_h, \ \mathbf{q} \in \mathbf{Q}_h, \tag{47}$$

with

$$\mathbf{D}_h^2\varphi = (D_{hij}^2(\varphi))_{1 \leq i,j \leq 2}, \tag{48}$$

$$((\mathbf{S}, \mathbf{T}))_h = \frac{1}{3} \sum_{k=1}^{N_{0h}} A_k \mathbf{S}(P_k) : \mathbf{T}(P_k)$$

$$\left(= \frac{1}{3} \sum_{k=1}^{N_{0h}} A_k (s_{11}t_{11} + s_{22}t_{22} + 2s_{12}t_{12})(P_k) \right), \quad \forall \mathbf{S}, \mathbf{T} \in \mathbf{Q}_h, \tag{49}$$

and then

$$\|\mathbf{S}\|_h = ((\mathbf{S}, \mathbf{S}))_h^{1/2}, \quad \forall \mathbf{S} \in \mathbf{Q}_h. \tag{50}$$

From the above relations, we approximate the problem (LSQ) by the following *discrete least-squares problem*:

$$\begin{cases} \{\psi_h, \mathbf{p}_h\} \in V_{gh} \times \mathbf{Q}_{fh}, \\ j_h(\psi_h, \mathbf{p}_h) \leq j_h(\varphi, \mathbf{q}), \quad \forall \{\varphi, \mathbf{q}\} \in V_{gh} \times \mathbf{Q}_{fh}. \end{cases} \tag{51}$$

6.4 On the Solution of the Problem (51)

To solve the minimization problem (51), we shall use the following discrete variant of the algorithm (9)–(11):

$$\{\psi^0, \mathbf{p}^0\} = \{\psi_0, \mathbf{p}_0\}. \tag{52}$$

Then, for $n \geq 0$, $\{\psi^n, \mathbf{p}^n\}$ being known, compute $\{\psi^{n+1}, \mathbf{p}^{n+1}\}$ via the solution of

$$\mathbf{p}^{n+1} = \operatorname*{arg\,min}_{\mathbf{q} \in \mathbf{Q}_{fh}} \left[\tfrac{1}{2}(1+\tau)\|\mathbf{q}\|_h^2 - ((\mathbf{p}^n + \tau\mathbf{D}_h^2\psi^n, \mathbf{q}))_h \right], \tag{53}$$

and

$$\begin{cases} \psi^{n+1} \in V_{gh}, \\ (\triangle_h[(\psi^{n+1} - \psi^n)/\tau], \triangle_h\varphi)_h + ((\mathbf{D}_h^2\psi^{n+1}, \mathbf{D}_h^2\varphi))_h \\ \quad = ((\mathbf{p}^{n+1}, \mathbf{D}_h^2\varphi))_h, \quad \forall\varphi \in V_{0h}, \end{cases} \tag{54}$$

where we have

$$(1) \qquad \triangle_h\varphi = D_{h11}^2(\varphi) + D_{h22}^2(\varphi), \qquad \forall\varphi \in V_h, \tag{55}$$

$$(2) \quad (\varphi_1, \varphi_2)_h = \frac{1}{3}\sum_{k=1}^{N_{0h}} A_k\varphi_1(P_k)\varphi_2(P_k), \quad \forall\varphi_1, \varphi_2 \in V_{0h}, \tag{56}$$

the associated norm being still denoted by $\|\cdot\|_h$.

The constrained minimization sub-problems (53) decompose into N_{0h} three-dimensional minimization problems (one per internal vertex of $\mathcal{T}_h$) similar to those encountered in Section 4, concerning the solution of the problem (10). The various solution methods (briefly) discussed in Section 4 still apply here. For the solution of the *linear* sub-problems (54), we advocate the following discrete variant of the *conjugate gradient* algorithm (23)–(29) (Algorithm 1):

Algorithm 2

Step 1. u^0 is given in V_{gh}.
Step 2. Solve

$$\begin{cases} g_h^0 \in V_{0h}, \\ (\triangle_h g^0, \triangle_h\varphi)_h = (\triangle_h u^0, \triangle_h\varphi)_h + \tau((\mathbf{D}_h^2 u^0, \mathbf{D}_h^2\varphi))_h - L_h(\varphi), \\ \quad \forall\varphi \in V_{0h}, \end{cases} \tag{57}$$

and set

$$w^0 = g^0. \tag{58}$$

Step 3. Then, for $k \geq 0$, assuming that u^k, g^k and w^k are known with the last two different from 0, solve

$$\begin{cases} \bar{g}^k \in V_{0h}, \\ (\triangle_h\bar{g}^k, \triangle_h\varphi)_h = (\triangle_h w^k, \triangle_h\varphi)_h + \tau((\mathbf{D}_h^2 w^k, \mathbf{D}_h^2\varphi))_h, \\ \quad \forall\varphi \in V_{0h}, \end{cases} \tag{59}$$

and compute

$$\rho_k = (\triangle_h g^k, \triangle_h g^k)_h / (\triangle_h\bar{g}^k, \triangle_h w^k)_h, \tag{60}$$

$$u^{k+1} = u^k - \rho_k w^k, \tag{61}$$

$$g^{k+1} = g^k - \rho_k\bar{g}^k. \tag{62}$$

Step 4. If $(\triangle_h g^k, \triangle_h g^k)_h / (\triangle_h g^0, \triangle_h g^0)_h \leq$ tol. take $u = u^{k+1}$; else, compute

$$\gamma_k = (\triangle_h g^{k+1}, \triangle_h g^{k+1})_h / (\triangle_h g^k, \triangle_h g^k)_h \qquad (63)$$

and update w^k via

$$w^{k+1} = g^{k+1} + \gamma_k w^k. \qquad (64)$$

Step 5. Do $k + 1 \to k$ and return to Step 3.

When solving the sub-problems (54), the linear functional $L_h(\cdot)$ encountered in (57) reads as follows:

$$L_h(\varphi) = (\triangle_h \psi^n, \triangle_h \varphi)_h + \tau((\mathbf{p}^{n+1}, \mathbf{D}_h^2 \varphi))_h.$$

Concerning the solution of the discrete *bi-harmonic problems* (57) and (59), let us observe that both problems are of the following type:

$$\begin{cases} \text{Find } u_h \in V_{0h} \text{ (or } V_{gh}) \text{ such that} \\ (\triangle_h u_h, \triangle_h v)_h = L_h(v), \quad \forall v \in V_{0h}, \end{cases} \qquad (65)$$

the functional $L_h(\cdot)$ being linear. Let us denote $-\triangle_h u_h$ by ω_h. It follows then from (37), (55) and (56) that the problem (65) is equivalent to the following system of two coupled discrete Poisson–Dirichlet problems:

$$\begin{cases} \omega_h \in V_{0h}, \\ \displaystyle\int_\Omega \nabla \omega_h \cdot \nabla v \, dx = L_h(v), \quad \forall v \in V_{0h}, \end{cases} \qquad (66)$$

$$\begin{cases} u_h \in V_{0h} (or\ V_{gh}), \\ \displaystyle\int_\Omega \nabla u_h \cdot \nabla v \, dx = (\omega_h, v)_h, \quad \forall v \in V_{0h}. \end{cases} \qquad (67)$$

Both problems are well-posed. Actually, the solution (by direct or iterative methods) of discrete Poisson problems, such as (66) and (67), has motivated an important literature; some related references can be found in [Glo03, Chapter 5].

We shall conclude this section by observing that via the algorithm (52)–(54) we have thus reduced the solution of (E-MA-D)$_h$ to the solution of

1. a sequence of discrete (linear) Poisson–Dirichlet problems.
2. a sequence of minimization problems in $\mathbb{R}^3$ (or $\mathbb{R}^2$).

7 Numerical Experiments

The least-squares based methodology discussed in the above sections has been applied to the solution of three particular (E-MA-D) problems, with $\Omega = (0,1)^2$. The *first test problem* can be formulated as follows (with $|x| = (x_1^2 + x_2^2)^{1/2}$ and $R \geq \sqrt{2}$):

$$\det D^2\psi = \frac{R^2}{(R^2 - |x|^2)^{\frac{1}{2}}} \quad \text{in } \Omega, \quad \psi = (R^2 - |x|^2)^{\frac{1}{2}} \text{ on } \Gamma. \tag{68}$$

The function ψ defined by $\psi(x) = (R^2 - |x|^2)^{1/2}$ is a solution to the problem (68). Its graph is a piece of the sphere of center $\mathbf{0}$ and radius R. We have discretized the problem (68) relying on the mixed finite element approximation discussed in Section 6, associated to a uniform triangulation of Ω (like the one shown on Figure 3, but finer). The uniformity of the mesh allows us to solve the various elliptic problems encountered at each iteration of the algorithm (57)–(64) (Algorithm 2) by fast Poisson solvers taking advantage of the decomposition properties of the discrete analogues of the biharmonic problems (23) and (24). To initialize the algorithm (52)–(54), we followed Remark 6 (see Section 3) and defined ψ_0 as the solution of the discrete Poisson problem

$$\begin{cases} \psi_0 \in V_{gh}, \\ \displaystyle\int_\Omega \boldsymbol{\nabla}\psi_0 \cdot \boldsymbol{\nabla}v \, dx = 2(\sqrt{f_h}, v)_h, \quad \forall v \in V_{0h} \end{cases}$$

and $\mathbf{p}_0$ by $\mathbf{p}_0 = \mathbf{D}_h^2\psi_0$. The algorithm (52)–(54) diverges if $R = \sqrt{2}$ (which is not surprising since the corresponding $\psi \notin H^2(\Omega)$). On the other hand, for $R = 2$ we have a quite fast convergence as soon as τ is large enough, the corresponding results being reported in Table 1. (We stopped iterating as soon as $\|D_h^2\psi_h^n - \mathbf{p}_h^n\|_{0,\Omega} \leq 10^{-6}$.)

Above, $\{\psi_h^c, \mathbf{p}_h^c\}$ is the computed approximate solution, h the space discretization step, n_{it} the number of iterations necessary to achieve convergence, and $\|D_h^2\psi_h^c - \mathbf{p}_h^c\|_{0,\Omega}$ is a trapezoidal rule based approximation of

Table 1. First test problem: convergence results

h	τ	n_{it}	$\|D_h^2\psi_h^c - \mathbf{p}_h^c\|_{\mathbf{Q}}$	$\|\psi_h^c - \psi\|_{L^2(\Omega)}$
1/32	0.1	517	0.9813×10^{-6}	0.450×10^{-5}
1/32	1	73	0.9618×10^{-6}	0.449×10^{-5}
1/32	10	28	0.7045×10^{-6}	0.450×10^{-5}
1/32	100	21	0.6773×10^{-6}	0.449×10^{-5}
1/32	1,000	22	0.8508×10^{-6}	0.449×10^{-5}
1/32	10,000	22	0.8301×10^{-6}	0.449×10^{-5}
1/64	1	76	0.9624×10^{-6}	0.113×10^{-5}
1/64	10	29	0.8547×10^{-6}	0.113×10^{-5}
1/64	100	24	0.8094×10^{-6}	0.113×10^{-5}

$(\int_{\Omega} |\mathbf{D}_h^2 \psi_h^c - \mathbf{p}_h^c|^2 \, dx)^{1/2}$. Table 1 clearly suggests that: (1) For τ large enough the speed of convergence is essentially independent of τ; (2) The speed of convergence is essentially independent of h; (3) The $L^2(\Omega)$-approximation error is $O(h^2)$.

The *second test problem* is defined by

$$\det \mathbf{D}^2 \psi = \frac{1}{|x|} \quad \text{in } \Omega, \quad \psi = \frac{2\sqrt{2}}{3}|x|^{\frac{3}{2}} \quad \text{on } \Gamma. \tag{69}$$

With these data, the function ψ defined by $\psi(x) = \frac{2\sqrt{2}}{3}|x|^{\frac{3}{2}}$ is a solution of the problem (69). It is easily shown that $\psi \in W^{2,p}(\bar{\Omega})$ for all $p \in [1,4)$, but does not have the $C^2(\bar{\Omega})$-regularity. Using the same approximation and algorithms than for the first test problem, we obtain the results reported in Table 2.

The various comments we have done concerning the solution of the first test problem still apply here. The graphs of f and ψ_h^c (for $h = 1/64$) have been visualized in Figures 4 and 5, respectively.

The *third test problem*, namely

$$\det D^2 \psi = 1 \quad \text{in } \Omega, \quad \psi = 0 \quad \text{on } \Gamma, \tag{70}$$

has no solution in $H^2(\Omega)$, despite the smoothness of the data, making it, by far, the more interesting (in some sense) of our test problems, from a computational point of view. We have reported in Table 3 the results produced by the algorithm (52)–(54) using $\|\psi_h^{n+1} - \psi_h^n\|_{L^2(\Omega)} \le 10^{-7}$ as the stopping criterion.

It is clear from Table 3 that the convergence is slower than for the first two test problems, however, some important features remain such as: the number of iterations necessary to achieve convergence is essentially independent of τ, as soon as this parameter is large enough, and increases slowly with $1/h$ (actually like $h^{-1/2}$). In Figures 6, 7 and 8 we have shown, respectively, the graph of ψ_h^c (for $h = 1/64$), the graph of the function $x_1 \to \psi_h^c(x_1, 1/2)$ when $x_1 \in [0,1]$, and the graph of the restriction of ψ_h^c to the line $x_1 = x_2$ (i.e., the

Table 2. Second test problem: convergence results

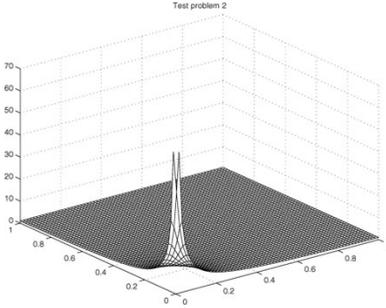| h	τ	n_{it}	$\|D_h^2 \psi_h^c - \mathbf{p}_h^c\|_{\mathbf{Q}}$	$\|\psi_h^c - \psi\|_{L^2(\Omega)}$
1/32	1	145	0.9381×10^{-6}	0.556×10^{-4}
1/32	10	56	0.9290×10^{-6}	0.556×10^{-4}
1/32	100	46	0.9285×10^{-6}	0.556×10^{-4}
1/32	1,000	45	0.9405×10^{-6}	0.556×10^{-4}
1/64	1	151	0.9500×10^{-6}	0.145×10^{-4}
1/64	10	58	0.9974×10^{-6}	0.145×10^{-4}
1/64	100	49	0.9531×10^{-6}	0.145×10^{-4}
1/64	1,000	48	0.9884×10^{-6}	0.145×10^{-4}

Fig. 4. Second test problem: graph of f.
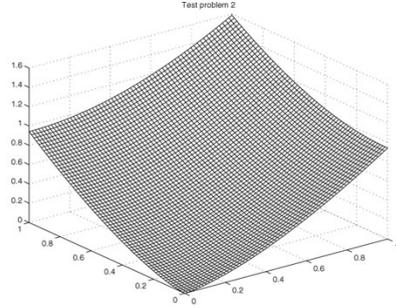
Fig. 5. Second test problem: graph of ψ_h^c ($h = 1/64$)

Table 3. Third test problem: convergence results

h	τ	n_{it}	$\|D_h^2\psi_h^c - \mathbf{p}_h^c\|_\mathbf{Q}$
1/32	1	4,977	0.1054×10^{-1}
1/32	100	3,297	0.4980×10^{-2}
1/32	1,000	3,275	0.4904×10^{-2}
1/32	10,000	3,273	0.4896×10^{-2}
1/64	1	6,575	0.1993×10^{-1}
1/64	100	4,553	0.1321×10^{-1}
1/64	1,000	4,527	0.1312×10^{-1}
1/128	100	5,401	0.1841×10^{-1}
1/128	1,000	5,372	0.1830×10^{-1}
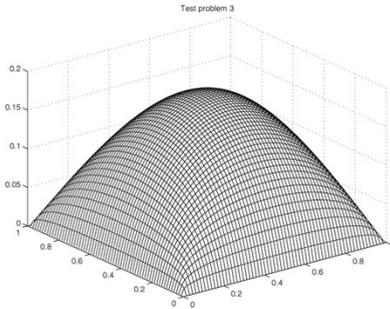
Fig. 6. Third test problem: graph of ψ_h^c ($h = 1/64$)

graph of the function $\xi \to \psi_h^c(\xi,\xi)$ when $\xi \in [0,1]$). In Figures 7 and 8, we used $- \cdot -\cdot$ (resp., $- - -$ and $—$) to represent the results corresponding to $h = 1/32$ (resp., $h = 1/64$ and $h = 1/128$).

The results in Figures 7 and 8 suggest strongly that ψ_h converges to a limit as $h \to 0$. They suggest also that the convergence is *superlinear* with respect to h. The above limit can be viewed as a *generalized solution* of (E-MA-D)
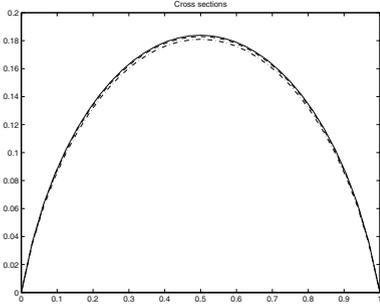
Fig. 7. Third test problem: graph of ψ_h^c restricted to the line $x_2 = 1/2$
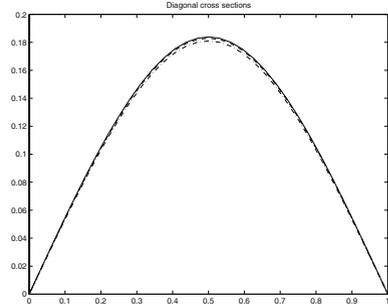
Fig. 8. Third test problem: graph of ψ_h^c restricted to the line $x_1 = x_2$

(in a *least-squares sense*). Actually, a closer inspection of the numerical results shows that the curvature of the graph is negative close to the corners, implying that the Monge–Ampère equation (70) is violated there (since the curvature is given by $\det \mathbf{D}^2 \psi / (1 + |\boldsymbol{\nabla} \psi|^2)^2$). Indeed, as expected, it is also violated along the boundary, since $\|\mathbf{D}_h^2 \psi_h^c\|_{0,\Omega} \approx 10^{-2}$, while $\|\mathbf{D}_h^2 \psi_h^c\|_{0,\Omega_1} \approx 10^{-4}$ and $\|\mathbf{D}_h^2 \psi_h^c\|_{0,\Omega_2} \approx 10^{-5}$, where $\Omega_1 = (1/8, 7/8)^2$ and $\Omega_2 = (1/4, 3/4)^2$. These results show that in that particular case, at least, the Monge–Ampère equation $\det \mathbf{D}^2 \psi = 1$ is verified with a good accuracy, sufficiently far away from Γ.

8 Further Comments

A natural question arising from the material discussed in the above sections is the following one: *Does our least-squares methodology provide viscosity solutions?*

We claim that indeed the solutions obtained by the least-squares methodology discussed in the preceding sections are (kind of) *viscosity solutions*. To show this property, let us consider (as in Section 3) the *flow* associated with the least-squares optimality conditions (7). We have then

$$
\begin{cases}
\text{Find } \{\psi(t), \mathbf{p}(t)\} \in V_g \times \mathbf{Q} \text{ for all } t > 0 \text{ such that} \\[4pt]
\displaystyle \int_{\Omega} \partial(\triangle \psi)/\partial t \; \triangle \varphi \, dx + \int_{\Omega} \mathbf{D}^2 \psi : \mathbf{D}^2 \varphi \, dx \\[4pt]
\qquad\qquad = \displaystyle \int_{\Omega} \mathbf{p} : \mathbf{D}^2 \varphi \, dx, \quad \forall \varphi \in V_0, \\[4pt]
\displaystyle \int_{\Omega} \partial \mathbf{p}/\partial t : \mathbf{q} \, dx + \int_{\Omega} \mathbf{p} : \mathbf{q} \, dx + \langle \partial I_{\mathbf{Q}_f}(\mathbf{p}), \mathbf{q} \rangle \\[4pt]
\qquad\qquad = \displaystyle \int_{\Omega} \mathbf{D}^2 \psi : \mathbf{q} \, dx, \quad \forall \mathbf{q} \in \mathbf{Q}, \\[4pt]
\{\psi(0), \mathbf{p}(0)\} = \{\psi_0, \mathbf{p}_0\}.
\end{cases}
\tag{71}
$$

Assuming that Ω is *simply connected*, we introduce:

$$\mathbf{u} = \{u_1, u_2\} = \{\partial\psi/\partial x_2, -\partial\psi/\partial x_1\},$$
$$\mathbf{v} = \{v_1, v_2\} = \{\partial\varphi/\partial x_2, -\partial\varphi/\partial x_1\},$$
$$\omega = \partial u_2/\partial x_1 - \partial u_1/\partial x_2,$$
$$\theta = \partial v_2/\partial x_1 - \partial v_1/\partial x_2,$$
$$\mathbf{V}_g = \{\mathbf{v} \mid \mathbf{v} \in (H^1(\Omega))^2,\ \boldsymbol{\nabla}\cdot\mathbf{v} = 0,\ \mathbf{v}\cdot\mathbf{n} = dg/ds \text{ on } \Gamma\},$$
$$\mathbf{V}_0 = \{\mathbf{v} \mid \mathbf{v} \in (H^1(\Omega))^2,\ \boldsymbol{\nabla}\cdot\mathbf{v} = 0,\ \mathbf{v}\cdot\mathbf{n} = 0 \text{ on } \Gamma\},$$
$$\mathbf{L} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}.$$

Above, $\mathbf{n}$ is the unit vector of the outward normal at Γ and s is a counterclockwise curvilinear abscissa on Γ. The formulation (71) is equivalent to

$$\begin{cases} \text{Find } \mathbf{u}(t) \in \mathbf{V}_g \text{ for all } t > 0 \text{ such that} \\ \displaystyle\int_\Omega \partial\omega/\partial t\,\theta\,dx + \int_\Omega \boldsymbol{\nabla}\mathbf{u} : \boldsymbol{\nabla}\mathbf{v}\,dx = \int_\Omega \mathbf{L}\mathbf{p} : \boldsymbol{\nabla}\mathbf{v}\,dx, \quad \forall \mathbf{v} \in \mathbf{V}_0, \\ \partial\mathbf{p}/\partial t + \mathbf{p} + \partial I_{\mathbf{Q}_f}(\mathbf{p}) + \mathbf{L}\boldsymbol{\nabla}\mathbf{u} = 0, \\ \{\mathbf{u}(0), \mathbf{p}(0), \omega(0)\} = \{\mathbf{u}_0, \mathbf{p}_0, \omega_0\}. \end{cases} \tag{72}$$

The problem (72) has a *visco-elasticity* flavor, $-\mathbf{L}\mathbf{p}$ playing here the role of the so-called *extra-stress tensor*. As $t \to +\infty$, we obtain thus at the limit a (kind of) *viscosity solution*.

References

[Aub82]   Th. Aubin. *Nonlinear Analysis on Manifolds, Monge–Ampère Equations.* Springer-Verlag, Berlin, 1982.

[Aub98]   Th. Aubin. *Some Nonlinear Problems in Riemanian Geometry.* Springer-Verlag, Berlin, 1998.

[BB00]   J.-D. Benamou and Y. Brenier. A computational fluid mechanics solution to the Monge–Kantorovich mass transfer problem. *Numer. Math.*, 84(3):375–393, 2000.

[Cab02]   X. Cabré. Topics in regularity and qualitative properties of solutions of nonlinear elliptic equations. *Discrete Contin. Dyn. Syst.*, 8(2):331–359, 2002.

[CC95]   L. A. Caffarelli and X. Cabré. *Fully Nonlinear Elliptic Equations.* American Mathematical Society, Providence, RI, 1995.

[CH89]   R. Courant and D. Hilbert. *Methods of Mathematical Physics, Vol. II.* Wiley Interscience, New York, 1989.

[CIL92]   M. G. Crandall, H. Ishii, and P.-L. Lions. User's guide to viscosity so-
           lutions of second order partial differential equations. *Bull. Amer. Math.
           Soc. (N.S.)*, 27(1):1–67, 1992.
[CKO99]   L. A. Caffarelli, S. A. Kochenkgin, and V. I. Oliker. On the numerical
           solution of reflector design with given far field scattering data. In
           L. A. Caffarelli and M. Milman, editors, *Monge-Ampère Equation: Ap-
           plication to Geometry and Optimization*, pages 13–32. American Mathe-
           matical Society, Providence, RI, 1999.
[DG03]   E. J. Dean and R. Glowinski. Numerical solution of the two-dimensional
           elliptic Monge–Ampère equation with Dirichlet boundary conditions:
           an augmented Lagrangian approach. *C. R. Math. Acad. Sci. Paris*,
           336(9):779–784, 2003.
[DG04]   E. J. Dean and R. Glowinski. Numerical solution of the two-dimensional
           elliptic Monge–Ampère equation with Dirichlet boundary conditions: a
           least-squares approach. *C. R. Math. Acad. Sci. Paris*, 339(12):887–892, 2004.
[DG05]   E. J. Dean and R. Glowinski. On the numerical solution of a two-
           dimensional Pucci's equations with Dirichlet boundary conditions: a least-
           squares approach. *C. R. Math. Acad. Sci. Paris*, 341(6):375–380, 2005.
[DG06a]   E. J. Dean and R. Glowinski. An augmented Lagrangian approach to the
           numerical solution of the Dirichlet problem for the elliptic Monge–Ampère
           equation in two dimensions. *Electron. Trans. Numer. Anal.*, 22:71–96, 2006.
[DG06b]   E. J. Dean and R. Glowinski. Numerical methods for fully nonlinear
           elliptic equations of the Monge–Ampère type. *Comput. Methods Appl.
           Mech. Engrg.*, 195(13–16):1344–1386, 2006.
[DGP91]   E. J. Dean, R. Glowinski, and O. Pironneau. Iterative solution of the
           stream function-vorticity formulation of the Stokes problem. Applications
           to the numerical simulation of incompressible viscous flow. *Comput. Meth-
           ods Appl. Mech. Engrg.*, 87(2–3):117–155, 1991.
[DS96]   J. E. Dennis and R. Schnabel. *Numerical Methods for Unconstrained
           Optimization and Nonlinear Equations*. SIAM, Philadelphia, PA, 1996.
[Glo84]   R. Glowinski. *Numerical Methods for Nonlinear Variational Problems*.
           Springer-Verlag, New York, 1984.
[Glo03]   R. Glowinski. Finite element methods for incompressible viscous flow. In
           P. G. Ciarlet and J.-L. Lions, editors, *Handbook of Numerical Analysis,
           Vol. IX*, pages 3–1176. North-Holland, Amsterdam, 2003.
[GP79]   R. Glowinski and O. Pironneau. Numerical methods for the first bihar-
           monic equation and for the two-dimensional Stokes problem. *SIAM Rev.*,
           17(2):167–212, 1979.
[GT01]   D. Gilbarg and N. Trudinger. *Elliptic Partial Differential Equations of
           Second Order*. Springer-Verlag, Berlin, 2001.
[Jan88]   R. Jansen. The maximum principle for viscosity solutions of fully non-
           linear second order partial differential equations. *Arch. Rational Mech.
           Anal.*, 101:1–27, 1988.
[OP88]   V. I. Oliker and L. D. Prussner. On the numerical solution of the equation
           $(\partial^2 z/\partial x^2)(\partial^2 z/\partial y^2) - ((\partial^2 z/\partial x\partial y))^2 = f$ and its discretization, I. *Numer.
           Math.*, 54(3):271–293, 1988.
[Pir89]   O. Pironneau. *Finite Element Methods for Fluids*. Wiley, Chichester,
           1989.
[Urb88]   J. I. E. Urbas. Regularity of generalized solutions of Monge–Ampère
           equations. *Math. Z.*, 197(3):365–393, 1988.

Part II

Linear and Nonlinear Hyperbolic Problems

Higher Order Time Stepping for Second Order Hyperbolic Problems and Optimal CFL Conditions

J. Charles Gilbert and Patrick Joly

INRIA Rocquencourt, BP 105, 78153 Le Chesnay, France
`Jean-Charles.Gilbert@inria.fr`
`Patrick.Joly@inria.fr`

Summary. We investigate explicit higher order time discretizations of linear second order hyperbolic problems. We study the even order $(2m)$ schemes obtained by the modified equation method. We show that the corresponding CFL upper bound for the time step remains bounded when the order of the scheme increases. We propose variants of these schemes constructed to optimize the CFL condition. The corresponding optimization problem is analyzed in detail and the analysis results in a specific numerical algorithm. The corresponding results are quite promising and suggest various conjectures.

1 Introduction

We are concerned here with a very classical problem, namely the numerical approximation of second order hyperbolic problems, more precisely problems of the form

$$\frac{d^2 u}{dt^2} + \mathcal{A}u = 0, \tag{1}$$

where $\mathcal{A}$ is a linear unbounded positive self-adjoint operator in some Hilbert space V. This appears to be the generic abstract form for a large class of partial differential equations in which u denotes a function $u(x,t)$ from $\Omega \subset \mathbb{R}^d \times \mathbb{R}^+$ in $\mathbb{R}^N$ and $\mathcal{A}$ is a second order differential operator in space, of elliptic nature. Such models are used for wave propagation in various domains of application, in particular, in acoustics, electromagnetism, and elasticity [Jol03].

During the past four decades, a considerable literature has been devoted to the construction of numerical methods for the approximation of (1). The most recent research deals with the construction of higher order in space and conservative methods for the space semi-discretization of (1) (see, for instance, [Coh02] and the references therein). These methods lead us to consider a family (indexed by $h > 0$, the approximation parameter which

tends to 0 – typically the step size of the computational mesh) of problems of
the form:

$$\frac{d^2 u_h}{dt^2} + \mathcal{A}_h u_h = 0, \tag{2}$$

where the unknown u_h is a function of time with value in some Hilbert space V_h
(whose norm will be denoted $\|\cdot\|$, even if it does depend on h) and $\mathcal{A}_h$ denotes
a bounded self-adjoint and positive operator in V_h (namely an approximation
of the second order differential operator $\mathcal{A}$). Several approaches lead naturally
to problems of the form (2), among which

- variational finite differences [CJ96, Dab86, AKM74],
- finite element methods [CJRT01, CJKMVV99],
- mixed finite element methods [CF05, PFC05],
- conservative discontinuous Galerkin methods [HW02, FLLP05].

Of course, the norm of $\mathcal{A}_h$ blows up when h goes to 0, as

$$\|\mathcal{A}_h\| = O(h^{-2}).$$

It is well known that one has conservation of the discrete energy:

$$E_h(t) = \frac{1}{2} \left\| \frac{du_h}{dt} \right\|^2 + \frac{1}{2} a_h(u_h, u_h),$$

where $a_h(\cdot, \cdot)$ is the continuous symmetric bilinear form associated with $\mathcal{A}_h$.
From the energy conservation result and the positivity of $\mathcal{A}_h$, one deduces a
stability result: the norm of the solution $u_h(t)$ can be estimated in function
of the norm of the Cauchy data:

$$u_{0,h} = u_h(0), \quad u_{1,h} = \frac{du_h}{dt}(0),$$

with constants independent of h. This is also a direct consequence of the
formula:

$$u_h(t) = \left[\cos \mathcal{A}_h^{\frac{1}{2}} t \right] u_{0,h} + \left[\mathcal{A}_h^{-\frac{1}{2}} \sin \mathcal{A}_h^{\frac{1}{2}} t \right] u_{1,h},$$

which yields

$$\|u_h(t)\| \leq \|u_{0,h}\| + t\|u_{1,h}\|. \tag{3}$$

In what follows, we are interested in the time discretization of (2) by explicit
finite difference schemes. More specifically, we are interested in the stability
analysis of such schemes, i.e., in obtaining a priori estimates of the form
(3) after time discretization. The conservative nature (i.e., the conservation
of energy) of the continuous problem can be seen as a consequence of the
time reversibility of this equation. That is why we shall favor centered finite
difference schemes which preserve such a property at the discrete level.

The most well known scheme is the classical second order leap-frog scheme.
Let us consider a time step $\Delta t > 0$ and denote by $u_h^n \in V_h$ an approximation
of $u_h(t^n)$, $t^n = n\Delta t$. This scheme is

$$\frac{u_h^{n+1} - 2u_h^n + u_h^{n-1}}{\Delta t^2} + \mathcal{A}_h u_h^n = 0. \tag{4}$$

Of course, (4) must be completed by a start-up procedure using the initial conditions to compute u_h^0 and u_h^1. We omit this here for simplicity.

By construction, this scheme is second order accurate in time. Its stability analysis is well known and we have (see, for instance, [Jol03]):

Theorem 1. *A necessary and sufficient condition for the stability of* (4) *is*

$$\frac{\Delta t^2}{4} \|\mathcal{A}_h\| \leq 1. \tag{5}$$

Remark 1. The condition (5) appears as an abstract CFL condition. In the applications to concrete wave equations, it is possible to get a bound for $\|\mathcal{A}_h\|$ of the form

$$\|\mathcal{A}_h\| \leq \frac{4c_+^2}{h^2},$$

where c_+ is a positive constant. This one has the dimension of a propagation velocity and only depends on the continuous problem: it is typically related to the maximum wave velocity for the continuous problem. Therefore, a (weaker) sufficient stability condition takes the form

$$\frac{c_+ \Delta t}{h} \leq 1.$$

In many situations, it is also possible to get a lower bound of the form (where $c_- \leq c_+$ also has the dimension of velocity)

$$\|\mathcal{A}_h\| \geq \frac{4c_-^2}{h^2},$$

so that a necessary stability condition is

$$\frac{c_- \Delta t}{h} \leq 1.$$

$\square$

Next we investigate one way to construct more accurate (in time) discretization schemes for (2). This is particularly relevant when the operator $\mathcal{A}_h$ represents a space approximation of the continuous operator $\mathcal{A}$ in $O(h^k)$ with $k > 2$: if one thinks about taking a time step proportional to the space step h (a usual choice which is in conformity with a CFL condition), one would like to adapt the time accuracy to the space accuracy. In comparison to what has been done on the space discretization side, we found very few work in this direction, even though it is very likely that a lot of interesting solutions could probably be found in the literature on ordinary differential equations

[HW96]. Most of the existing work is in the context of finite difference methods, compact schemes, etc., see, for instance, [Dab86, SB87, CJ96, AJT00] or [DPJ06, TT05] in the context of the first order hyperbolic problems.

The content of the rest of this paper is as follows. In Section 2, we investigate a class of methods for the time discretization of (2), based on the so-called modified equation approach. These schemes can be seen as even higher order variations around the leap-frog scheme of which they preserve the main properties: explicit nature, time reversibility, energy conservation. It appears that the computational cost of one time step of the scheme of order $2m$ is m times larger than for one step of the second order scheme. This can be counterbalanced if one can use larger time steps than for the second order scheme. This is where the stability analysis plays a major role (Section 2). This one shows that even though the maximum allowed time step increases with m (particularly for small even values of m), it remains uniformly bounded with m (Theorem 3). In Section 3, we investigate the question of constructing other schemes, conceived as modifications of the previous one, that should satisfy:

- the good properties of the schemes (explicitness, conservativity, etc.) and the order of approximation are preserved,
- the maximal time step authorized by the CFL condition is larger.

We formulate this as a family of optimization problems that we analyze in detail. We are able to prove the existence and the uniqueness of the solution of these problems (Corollary 2) and to give necessary and sufficient conditions of optimality (Theorems 4 and 5) that we use to construct an algorithm for the effective computation of the solutions of these optimization problems. This algorithm, as well as the corresponding numerical results, are presented and discussed in Section 4. Our first results are quite promising and show that the optimization procedure does allow us to improve significantly the CFL condition. However, the corresponding numerical schemes still have to be tested numerically. This will be the object of a forthcoming work.

2 Higher Order Schemes by the Modified Equation Approach

2.1 The modified Equation Approach

It is possible to construct higher order schemes which remain explicit and centered. In particular, all the machinery of Runge–Kutta methods for ordinary differential equations [HW96] is available. Let us concentrate here on a classical approach, the so-called modified equation approach [SB87, CdLBL97, Dab86]. For instance, to construct a fourth order scheme, we start by looking at the truncation error of (4)

$$\frac{u_h(t^{n+1}) - 2u_h(t^n) + u_h(t^{n-1})}{\Delta t^2} = \frac{d^2 u_h}{dt^2}(t^n) + \frac{\Delta t^2}{12}\frac{d^4 u_h}{dt^4}(t^n) + O(\Delta t^4).$$

Using the equation satisfied by u_h, we get the identity

$$\frac{u_h(t^{n+1}) - 2u_h(t^n) + u_h(t^{n-1})}{\Delta t^2} = -\mathcal{A}_h u_h(t^n) + \frac{\Delta t^2}{12}\mathcal{A}_h^2 u_h(t^n) + O(\Delta t^4),$$

which leads to the following fourth order scheme:

$$\frac{u_h^{n+1} - 2u_h^n + u_h^{n-1}}{\Delta t^2} + \mathcal{A}_h u_h^n - \frac{\Delta t^2}{12}\mathcal{A}_h^2 u_h^n = 0. \tag{6}$$

This one can be implemented in such a way that each time step involves only two applications of the operator $\mathcal{A}_h$, using Horner's rule,

$$u_h^{n+1} = 2u_h^n - u_h^{n-1} - \Delta t^2 \mathcal{A}_h \left(I - \frac{\Delta t^2}{12}\mathcal{A}_h \right) u_h^n.$$

More generally, an explicit centered scheme of order $2m$ is given by

$$\frac{u_h^{n+1} - 2u_h^n + u_h^{n-1}}{\Delta t^2} + \mathcal{A}_h^{(m)}(\Delta t)u_h^n = 0, \quad \mathcal{A}_h^{(m)}(\Delta t) = \mathcal{A}_h P_m(\Delta t^2 \mathcal{A}_h), \tag{7}$$

where the polynomial $P_m(x)$ is defined by

$$P_m(x) = 1 + 2\sum_{l=1}^{m-1}(-1)^l \frac{x^l}{(2l+2)!}. \tag{8}$$

Indeed, a Taylor expansion gives

$$u_h(t^{n\pm 1}) = u_h(t^n) + \sum_{k=1}^{2m+1}(\pm 1)^k \frac{\Delta t^k}{k!}\frac{d^k u_h}{dt^k}(t^n) + O(\Delta t^{2m+2})$$

so that

$$\frac{u_h(t^{n+1}) - 2u_h(t^n) + u_h(t^{n-1})}{\Delta t^2} = 2\sum_{k=1}^{m}\frac{\Delta t^{2k-2}}{2k!}\frac{d^{2k} u_h}{dt^{2k}}(t^n) + O(\Delta t^{2m}).$$

Since $\frac{d^{2k} u_h}{dt^{2k}}(t^n) = (-1)^k \mathcal{A}_h^k u_h(t^n)$, we also have

$$\frac{u_h(t^{n+1}) - 2u_h(t^n) + u_h(t^{n-1})}{\Delta t^2} =$$

$$= -\mathcal{A}_h u_h(t^n) + 2\sum_{k=2}^{m}(-1)^k \frac{\Delta t^{2k-2}}{2k!}\mathcal{A}_h^k u_h(t^n) + O(\Delta t^{2m}),$$

or equivalently

$$\frac{u_h(t^{n+1}) - 2u_h(t^n) + u_h(t^{n-1})}{\Delta t^2} +$$

$$+ \mathcal{A}_h \left[u_h(t^n) + 2 \sum_{k=1}^{m-1} (-1)^k \frac{\Delta t^{2k}}{(2k+2)!} \mathcal{A}_h^k u_h(t^n) \right] = O(\Delta t^{2m}).$$

This identity leads to the scheme (7)–(8).

Using again Horner's rule for the representation of the polynomial P_m, reduces the calculation of u_h^{n+1} to m successive applications of the operator $\mathcal{A}_h(\Delta t)$, according to the following algorithm:

Step 1. Set $u_h^{n,0} = u_h^n$.

Step 2. Compute

$$u_h^{n,k} = u_h^{n,k-1} - 2 \frac{\Delta t^2 \mathcal{A}_h u_h^{n,k-1}}{(2k+1)(2k+2)}, \quad k = 1, \cdots, m.$$

Step 3. Set $u_h^{n+1} = u_h^{n,m}$.

In other words, since the most expensive step of the algorithm is the application of the operator $\mathcal{A}_h$ (a matrix-vector multiplication in practice), the computational cost for one time step of the scheme of order $2m$ is only m times larger than the computational cost for one time step of the scheme of order 2.

2.2 Stability Analysis

The stability analysis of the higher order scheme (7) is similar to the one of the second order scheme but it is complicated by the fact that one must verify that the operator $\mathcal{A}_h(\Delta t)$ is positive, which already imposes an upper bound on Δt.

Theorem 2. *A sufficient stability condition for scheme* (7) *is given by*

$$\Delta t^2 \|\mathcal{A}_h\| \le \alpha_m, \tag{9}$$

where we have defined

$$\alpha_m = \sup\{\alpha \mid \forall x \in [0, \alpha], \ 0 \le Q_m(x) \le 4\}, \tag{10}$$

with

$$Q_m(x) = x P_m(x) = x + 2 \sum_{l=1}^{m-1} (-1)^l \frac{x^{l+1}}{(2l+2)!}. \tag{11}$$

This condition is necessary as soon as the spectrum of $\mathcal{A}_h$ is the whole interval $[0, \|\mathcal{A}_h\|]$.

Proof. Using Von Neumann analysis [RM67] and spectral theory of self-adjoint operators (namely the spectral theorem [RS78]), it is sufficient to look at the (λ-parameterized) family of difference equations (u^n is now a sequence of complex numbers):

$$\frac{u^{n+1} - 2u^n + u^{n-1}}{\Delta t^2} + \lambda P_m(\lambda \Delta t^2)u^n = 0, \quad \lambda \in \sigma(\mathcal{A}_h), \qquad (12)$$

where $\sigma(\mathcal{A}_h)$ is the spectrum of $\mathcal{A}_h$. The characteristic equation of this recurrence is

$$r^2 - \left[2 - Q_m(\lambda \Delta t^2)\right] r + 1 = 0.$$

This is a second degree equation with real coefficients. The product of the roots being 1, the two solutions have modulus less than 1 – which is equivalent to the boundedness of u^n – if and only if the discriminant of this equation is non-positive, in which case the roots belong to the unit circle. This leads to $Q_m(\lambda \Delta t^2)[4 - Q_m(\lambda \Delta t^2)] \geq 0$ or

$$0 \leq Q_m(\lambda \Delta t^2) \leq 4.$$

If (9) holds, since $\sigma(\mathcal{A}_h) \subset [0, \|\mathcal{A}_h\|]$, $\lambda \Delta t^2 \in [0, 4]$ which proves that (9) is a sufficient stability condition. The second part of the proof is left to the reader. $\qquad\square$

Remark 2. The equality $\sigma(\mathcal{A}_h) = [0, \|\mathcal{A}_h\|]$ holds, for instance, when one uses a finite difference scheme of the wave equation with constant coefficients in the whole space. The Fourier analysis proves that the spectrum of $\mathcal{A}_h$ is, in this case, purely continuous. $\qquad\square$

The finiteness of α_m for each m is quite obvious. However, its value is difficult to compute explicitly, except for the first values of m. One has, in particular,

$$\alpha_1 = 4, \quad \alpha_2 = 12, \quad \alpha_3 = 2(5 + 5^{\frac{1}{3}} - 5^{\frac{2}{3}}) \simeq 7.572, \quad \alpha_4 \simeq 21.4812, \dots \quad (13)$$

For the exact – but very complicated – expression of α_4, we refer to [CJRT01] or [Jol03]; other values of α_m are given in the column "$k = 0$" of Table 1 on page 88. It is particularly interesting to note that for the fourth order scheme, one is allowed to take a time step which is $\sqrt{\alpha_2/\alpha_1}$ ($\simeq 1.732$) times larger than for the second order scheme, which almost balances the fact that the cost of one time step is twice larger. In the same way, with the scheme of order 8, one can take a time step $\sqrt{\alpha_4/\alpha_1}$ ($\simeq 2.317$) times larger (while each time step costs four times more). Surprisingly, the scheme of order 6 seems less interesting: the stability condition is more constraining that for the fourth order scheme.

From the theoretical point of view, it would be interesting to know the behaviour of α_m for large m. For this we first identify the limit behaviour of the polynomials $Q_m(x)$. One easily checks that

$$\lim_{m \to +\infty} Q_m(x) = Q_\infty(x) \equiv x + 2 \sum_{l=1}^{+\infty} (-1)^l \frac{x^{l+1}}{(2l+2)!} = 2(1 - \cos \sqrt{x}). \quad (14)$$

Remark 3. Setting $P_\infty(x) = 2\frac{1-\cos\sqrt{x}}{x}$ and taking (formally) the limit of (7) when $m \to +\infty$, we obtain the scheme

$$\frac{u_h^{n+1} - 2u_h^n + u_h^{n-1}}{\Delta t^2} + A_h P_\infty(\Delta t^2 A_h) = 0. \quad (15)$$

This scheme is, in fact, an *exact scheme* for the differential equation (2). It suffices to remark that

$$\left|
\begin{aligned}
\sin \left(A_h^{\frac{1}{2}} t^{n+1}\right) &- 2 \sin \left(A_h^{\frac{1}{2}} t^n\right) + \sin \left(A_h^{\frac{1}{2}} t^{n-1}\right) \\
&= - \left[2 - \cos \left(A_h^{\frac{1}{2}} \Delta t\right)\right] \sin \left(A_h^{\frac{1}{2}} t^n\right) \\
\cos \left(A_h^{\frac{1}{2}} t^{n+1}\right) &- 2 \cos \left(A_h^{\frac{1}{2}} t^n\right) + \cos \left(A_h^{\frac{1}{2}} t^{n-1}\right) \\
&= - \left[2 - \cos \left(A_h^{\frac{1}{2}} \Delta t\right)\right] \cos \left(A_h^{\frac{1}{2}} t^n\right),
\end{aligned}
\right.$$

so that any solution of (2), of the form (for some a and b in V_h)

$$u_h(t) = \cos \left(A_h^{\frac{1}{2}} t\right) a + \sin \left(A_h^{\frac{1}{2}} t\right) b$$

satisfies

$$\frac{u_h(t^{n+1}) - 2u_h(t^n) + u_h(t^{n-1})}{\Delta t^2} = -\left(A_h \Delta t^2\right)^{-1} \left[2 - \cos \left(A_h^{\frac{1}{2}} \Delta t\right)\right] A_h u_h(t^n),$$

that is to say

$$\frac{u_h(t^{n+1}) - 2u_h(t^n) + u_h(t^{n-1})}{\Delta t^2} = -A_h P_\infty(\Delta t^2 A_h).$$

$\square$

Since $0 \le Q_\infty(x) \le 4$, if we define α_∞ by (19) for $m = +\infty$ we have $\alpha_\infty = +\infty$. Unfortunately, this does not mean, as we are going to see, that $\alpha_m \to +\infty$ when $m \to +\infty$. In fact, to describe the behaviour of α_m, we have to distinguish between the even and odd sequences α_{2m} and α_{2m+1}. Our first observation is that the convergence of the sequences $Q_{2m}(x)$ and $Q_{2m+1}(x)$ is monotone. Indeed, for $m \ge 1$

$$Q_{2m-1}(x) - Q_{2m+1}(x) = 2\frac{x^{2m}}{4m!} \left[1 - \frac{x}{(4m+1)(4m+2)}\right]$$

which shows that $Q_{2m+1}(x)$ is a strictly decreasing sequence for large m:

$$Q_{2m+1}(x) < Q_{2m-1}(x) \quad \text{as soon as} \quad (4m+1)(4m+2) > x.$$

In particular, since $(4m+1)(4m+2) > \pi^2$ for $m \geq 1$:

$$Q_\infty(\pi^2) = \lim_{m \to +\infty} Q_{2m+1}(\pi^2) = 4 \quad \Longrightarrow \quad Q_{2m+1}(\pi^2) > 4,$$

which shows, using the definition (10), that

$$\alpha_{2m+1} \leq \pi^2, \quad \text{for } m \geq 1.$$

Moreover, by the definition of α_m, we know that $Q_m(\alpha_m) = 0$ or 4. On the other hand, since the sequence $Q_{2m+1}(x)$ is decreasing, for any $x \in [0, \pi^2]$, we have

$$Q_{2m+1}(x) > Q_\infty(x) = 2(1 - \cos \sqrt{x}) \quad \text{in } [0, \pi^2].$$

This makes impossible $Q_{2m+1}(\alpha_{2m+1}) = 0$, which implies that

$$Q_{2m+1}(\alpha_{2m+1}) = 4.$$

Finally, the inequality

$$Q_{2m+1}(x) < Q_1(x) = x$$

implies

$$Q_{2m+1}(x) < 4, \quad \forall x \in [0, 4],$$

which implies, in particular,

$$\alpha_{2m+1} > 4.$$

Let $\alpha_{\text{odd}} \in [4, 4\pi^2]$ be any accumulation point of α_{2m+1}, since the convergence of Q_m to Q_∞ is uniform in any compact set, we get:

$$Q_\infty(\alpha_{\text{odd}}) \quad \Longrightarrow \quad (\text{since } \alpha_{\text{odd}} \in [4, \pi^2]) \quad \alpha_{\text{odd}} = \pi^2.$$

In the same way

$$Q_{2m+2}(x) - Q_{2m}(x) = 2\frac{x^{2m+1}}{(4m+2)!}\left[1 - \frac{x}{(4m+3)(4m+4)}\right]$$

shows that the sequence $Q_{2m}(x)$ is strictly increasing for large m:

$$Q_{2m+2}(x) > Q_{2m}(x) \quad \text{as soon as } (4m+3)(4m+4) > x.$$

In particular, as soon as $m \geq 1$,

$$Q_\infty(4\pi^2) = \lim_{m \to +\infty} Q_{2m}(4\pi^2) = 0 \quad \Longrightarrow \quad Q_{2m}(4\pi^2) < 0,$$

which shows that

$$\alpha_{2m} \leq 4\pi^2, \quad m \geq 1,$$

while the inequality $Q_{2m}(x) < 2(1 - \cos \sqrt{x}) \leq 4$ in $[0, \pi^2]$ for $m \geq 1$ implies that

$$Q_{2m}(\alpha_{2m}) = 0.$$

Finally, the inequality, for $m > 1$,

$$Q_{2m}(x) > Q_2(x) = x(1 - x/12) \quad \text{for } x < 132$$

shows that $Q_{2m}(x) > 0$ for $x < 12$ which implies that

$$\alpha_{2m} \geq 12.$$

Let $\alpha_{\text{even}} \in [12, 4\pi^2]$ be any accumulation point of α_{2m}. We thus get

$$Q_{\infty}(\alpha_{\text{even}}) = 0 \quad \Longrightarrow \quad (\text{since } \alpha_{\text{even}} \in [12, 4\pi^2]) \quad \alpha_{\text{even}} = 4\pi^2.$$

We have shown the following result:

Theorem 3. *Let α_m be defined by* (10). *Then*

$$\lim_{m \to +\infty} \alpha_{2m} = 4\pi^2, \qquad \lim_{m \to +\infty} \alpha_{2m+1} = \pi^2. \tag{16}$$

3 Modified Higher Order Schemes: an Optimization Approach

For an integer k, we denote by $\mathbf{P}_k$ the set of polynomials of degree less or equal to k and define $\mathbf{P} \equiv \bigcup_{k \geq 0} \mathbf{P}_k$.

A general explicit scheme of order $2m$ is given by

$$\frac{u_h^{n+1} - 2u_h^n + u_h^{n-1}}{\Delta t^2} + \left[P_m(\Delta t^2 \mathcal{A}_h) + \Delta t^{2m} \mathcal{A}_h^m R_k(\Delta t^2 \mathcal{A}_h) \right] \mathcal{A}_h u_h^n = 0, \tag{17}$$

where $R_k \in \mathbf{P}_{k-1}$. The cost of this new scheme is a priori $(m+k)/m$ times larger than the cost of the scheme corresponding to $R_k = 0$. As in Theorem 2, the stability condition of this new scheme is

$$\frac{\Delta t^2}{4} \|\mathcal{A}_h\| \leq \alpha_m(R_k), \tag{18}$$

where we have defined

$$\alpha_m(R) = \sup\{\alpha \mid \forall x \in [0, \alpha], \ 0 \leq x[P_m(x) + x^m R(x)] \leq 4\}. \tag{19}$$

The natural idea, in some sense, to get an optimal scheme would be to solve the optimization problem:

$$\text{Find } R_{m,k} \in \mathbf{P}_{k-1} \quad \text{such that } \alpha_m(R_{m,k}) = \sup_{R \in \mathbf{P}_{k-1}} \alpha_m(R). \tag{20}$$

Then, assuming that this problem has a solution $R_{m,k}$, one gets the optimal CFL constant for the schemes in the class, namely

$$\alpha_{m,k} = \alpha_m(R_{m,k}). \tag{21}$$

Clearly, since $\mathbf{P}_{k-1} \subset \mathbf{P}_k$, $\alpha_{m,k}$ increases with k. We have also $\alpha_{m,k} > 0$, since $P_m(0) = 1$ $(m \geq 1)$.

For what follows, it is useful to introduce the following affine map:

$$\left| \begin{aligned} \psi_m : \mathbf{P} &\to \mathbf{P} \\ R &\to \psi_m(R) = Q_m + x^{m+1} R, \end{aligned} \right. \tag{22}$$

where we recall that $Q_m(x) = x P_m(x)$. Note that ψ_m maps $\mathbf{P}_{k-1}$ into $\mathbf{P}_{m+k}$.

Lemma 1. *The function $R \in \mathbf{P}_{k-1} \mapsto \alpha_m(R) \in \mathbb{R}_+^*$ has the following properties:*

(i) It goes to 0 at infinity:

$$\lim_{\|R\| \to +\infty} \alpha_m(R) = 0.$$

(ii) It is upper semi-continuous:

$$R_n \to R \quad \text{in } \mathbf{P}_{k-1} \quad \Longrightarrow \quad \alpha_m(R) \geq \limsup \alpha_m(R_n).$$

Proof. Let $r_j(R)$ denote the coefficient of x^j in $R \in \mathbf{P}_{k-1}$ and consider $R_n \in \mathbf{P}_{k-1}$ such that

$$\|R_n\|_\infty \equiv \sup_{0 \leq j \leq k-1} |r_j(R_n)| \longrightarrow +\infty.$$

Referring to the fact that $\mathbf{P}_{k-1}$ is finite dimensional, one can find a subsequence (still denoted R_n for simplification) and a fixed non-zero polynomial $\varphi \in \mathbf{P}_{k-1}$ such that, as soon as $\varphi(x) \neq 0$,

$$R_n(x) \sim \|R_n\|_\infty \varphi(x) \quad (n \to +\infty).$$

For such positive values of x, $[\psi_m(R_n)](x) \notin [0, 4]$ for sufficiently large n which means that $\alpha_m(R_n) < x \implies \limsup \alpha_m(R_n) < x$. Since φ is a non-zero polynomial, one can find arbitrarily small values of such x so that $\limsup \alpha_m(R_n) \leq 0$. As $\alpha_m(R_n)$ is a sequence of positive real numbers, this means that $\alpha_m(R_n)$ tends to 0.

On the other hand, let $R_n \in \mathbf{P}_{k-1}$ be a sequence converging to R. Let ε be any arbitrarily small positive number. By the uniform convergence of R_n to R in the interval $I_R(\varepsilon) = [0, \alpha(R) + \varepsilon]$ we have:

$$\lim_{n \to +\infty} \|\psi_m(R_n) - 2\|_{L^\infty(I_R(\varepsilon))} = \|\psi_m(R) - 2\|_{L^\infty(I_R(\varepsilon))} > 2.$$

Thus, there exists an integer N_ε such that:

$$n \geq N_\varepsilon \implies \|\psi_m(R_n) - 2\|_{L^\infty(I_R(\varepsilon))} > 2 \implies \alpha_m(R_n) < \alpha_m(R) + \varepsilon.$$

Therefore,

$$\limsup \alpha_m(R_n) \leq \alpha_m(R) + \varepsilon,$$

which yields (ε being arbitrarily small) $\limsup \alpha_m(R_n) \leq \alpha_m(R)$. $\qquad \square$
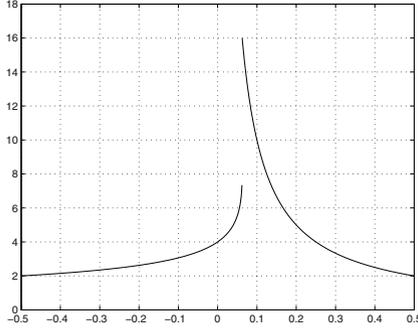
Fig. 1. Graph of the function $\alpha_1(r)$

The classical existence theory in analysis [Sch91, Theorem 2.7.11] leads an existence result.

Corollary 1 (Existence of a solution). *The optimization problem* (20) *has (at least) one solution.*

Clearly, the function $R \to \alpha_m(R)$ is not continuous. Let us consider, for instance, the case when $m = 1$ and $k = 1$. Then, the function $\alpha_1(R)$ can be identified to the function of the real variable r defined by

$$\alpha_1(r) = \sup\{\alpha \mid \forall x \in [0, \alpha], \ 0 \le x - rx^2 \le 4\}. \tag{23}$$

It is straightforward to compute that

$$\alpha_1(r) = \frac{1 - \sqrt{1 - 16r}}{2r} \quad \text{if } r < \frac{1}{16}, \quad \text{and} \quad \alpha_1(r) = \frac{1}{r} \quad \text{if } r \ge \frac{1}{16}.$$

It is clear that α_1 is discontinuous at $r = 1/16$ since (see also Figure 1)

$$\alpha_1(1/16) = 16 \quad \text{and} \quad \lim_{r \uparrow 1/16} \alpha_1(r) = 8.$$

Note that for $r = 1/16$ the graph of the polynomial $x - rx^2$ is tangent to the line $y = 4$ at $x = 8 < \alpha_1(1/16) = 16$. This is an illustration of a more general property.

Lemma 2. *Let D_k be the set of polynomials $R \in \mathbf{P}_{k-1}$ such that*

$$\exists x_* \in \,]0, \alpha_m(R)[\quad | \quad [\psi_m(R)](x_*) = 0 \ or \ 4. \tag{24}$$

The function $R \to \alpha_m(R)$ is discontinuous at every point of D_k and continuous everywhere else.

Proof. Let $R \in D_k$ be such that $[\psi_m(R)](x_*) = 4$ for some $x_* \in]0, \alpha_m(R)[$. (A similar argument works if $[\psi_m(R)](x_*) = 0$.) For any $\varepsilon > 0$, $\psi_m(R + \varepsilon) = \psi_m(R) + \varepsilon x^{m+1} > 4$ in a small neighborhood of x_*. This implies that $\alpha_m(R + \varepsilon) < x_* < \alpha_m(R)$, hence the discontinuity of α_m at R.

On the other hand, let $R \in \mathbf{P}_{k-1} \setminus D_k$ and consider a sequence of polynomials $R_n \in \mathbf{P}_{k-1}$ converging to R. Since

$$\left| \frac{[\psi_m(R_n)](x) - [\psi_m(R)](x)}{x^{m+1}} \right| = |R_n(x) - R(x)| \to 0,$$

uniformly in $x \in [0, \alpha_m(R)]$, there exists an integer N_1 such that $[\psi_m(R)](x) - x^{m+1} \leq [\psi_m(R_n)](x) \leq [\psi_m(R)](x) + x^{m+1}$ for $n \geq N_1$ and $x \in [0, \alpha_m(R)]$. These inequalities and the fact that $[\psi_m(R)](0) = 0$ and $[\psi_m(R)]'(0) = 1$ imply that there is $\varepsilon_1 > 0$ such that $[\psi_m(R_n)](x) \in [0, 4]$ for $n \geq N_1$ and $x \in [0, \varepsilon_1]$. In other words,

$$\text{for } n \geq N_1, \quad \alpha_m(R_n) \geq \varepsilon_1.$$

For any $\varepsilon \in]0, \varepsilon_1]$, small enough, and $J_R(\varepsilon) = [\varepsilon, \alpha_m(R) - \varepsilon]$, there holds

$$\|\psi_m(R) - 2\|_{L^\infty(J_R(\varepsilon))} < 2.$$

Then there exists an integer $N_\varepsilon \geq N_1$ such that for $n \geq N_\varepsilon$

$$\|\psi_m(R_n) - 2\|_{L^\infty(J_R(\varepsilon))} < 2 \quad \text{or} \quad \alpha_m(R_n) > \alpha_m(R) - \varepsilon.$$

Now $\varepsilon > 0$ is arbitrary small, so that $\liminf \alpha_m(R_n) \geq \alpha_m(R)$. The continuity of α_m at R follows, since α_m is upper semi-continuous by Lemma 1. $\qquad \square$

Lemma 3. *The set of solutions of the optimization problem* (20) *is a convex subset of D_k.*

Proof. Let us first prove that any local maximum of α_m belongs to D_k. Indeed, it is easy to see that, if $R \notin D_k$, the function

$$t \in \mathbb{R} \mapsto \alpha_m(R + t)$$

is continuous and strictly monotone in the neighborhood of the origin. This shows that R cannot be a local maximum of α_m.

Let R_1 and R_2 be two solutions of (20):

$$\alpha_m(R_1) = \alpha_m(R_2) = \alpha_{m.k} \equiv \sup_{R \in \mathbf{P}_{k-1}} \alpha_m(R).$$

By definition of α_m

$$\forall x \leq \alpha_{m,k}, \quad 0 \leq [\psi_m(R_1)](x) \leq 4 \quad \text{and} \quad 0 \leq [\psi_m(R_2)](x) \leq 4.$$

Therefore, since ψ_m is an affine function, for any $t \in [0,1]$, there holds

$$\forall x \le \alpha_{m,k}, \quad 0 \le \left[\psi_m\left(tR_1 + (1-t)R_2\right)\right](x) \le 4.$$

Hence

$$\alpha_m\left(tR_1 + (1-t)R_2\right) = \alpha_{m,k}.$$

In other words, any point of the segment $[R_1, R_2]$ is a solution of (20), i.e., the set of solutions of (20) is convex. □

As a consequence of Lemmas 2 and 3, we know that any solution R of (20) is such that

$$\mathcal{T}_R \equiv \{\tau \in \left]0, \alpha_{m,k}\right[\mid [\psi_m(R)](\tau) = 0 \text{ or } 4\}$$

is nonempty. Let us call *tangent point* an element of $\mathcal{T}_R$. Theorem 4 below is more precise, since it claims that there is at least $M \ge k$ tangent points τ_j at which $\psi_m(R)$ takes *alternatively* the values 0 and 4. For any R, it is convenient to construct and enumerate these tangent points in decreasing order:

$$\tau_{M+1} = 0 < \tau_M < \cdots < \tau_1 < \tau_0 = \alpha_{m,k}.$$

The selected subset $\{\tau_1, \tau_2, \ldots, \tau_M\} \subset \mathcal{T}_R$ is built as follows. Let us start by setting

$$\tau_0 = \alpha_{m,k} \quad \text{and} \quad s_0 = \begin{cases} -1 & \text{if } [\psi_m(R)](\tau_0) = 4, \\ +1 & \text{if } [\psi_m(R)](\tau_0) = 0. \end{cases} \tag{25}$$

The points $\tau_j \in \mathcal{T}_R$, $j = 1, \ldots, M$ and their number M are determined by the following recurrence: For $j \ge 1$,

1. set $s_j = -s_{j-1}$;
2. if this is possible, take τ_j as the largest $\tau \in \left]0, \tau_{j-1}\right[$ such that

$$[\psi_m(R)](\tau_j) = \begin{cases} 4 & \text{if } s_j = -1, \\ 0 & \text{if } s_j = +1. \end{cases}$$

The procedure stops when there is no relevant τ_j in the step 2 above (it must stop because of the polynomial nature of $\psi_m(R)$). In the proof of Theorem 4 below, s_j is actually the sign at τ_j of a certain function φ that is added to a potential solution R.

A priori, because of the chosen selection procedure, it may occur that $M = 0$, even though the number of tangent points is nonzero. The next theorem shows that this is not the case for a local maximum.

Theorem 4 (Necessary optimality condition). *Let R be a local maximum of (20). Then the number M of alternate tangent points selected by the procedure (25)+1+2 satisfies $M \ge k$.*

Proof. We proceed by contradiction, assuming that $M \leq k - 1$. For $j = 0, \ldots, M - 1$, one can find a point

$$\tau_{j+\frac{1}{2}} \in \,]\tau_{j+1}, \tau_j[\quad \text{such that } [\psi_m(R)] \left( ]\tau_{j+1}, \tau_{j+\frac{1}{2}}] \right) \subset \,]0, 4[. \qquad (26)$$

Consider the polynomial φ defined at $x \in \mathbb{R}$ by

$$\varphi(x) = s_0 \prod_{j=0}^{M-1} \left(x - \tau_{j+\frac{1}{2}} \right).$$

Hence $\varphi \equiv s_0$ if $M = 0$. This polynomial is of degree $M \leq k - 1$, so that it is a possible increment to R in $\mathbf{P}_{k-1}$. For $t > 0$, consider the polynomial $p_t = \psi_m(R + t\varphi)$, which verifies for all $x \in \mathbb{R}$:

$$p_t(x) = [\psi_m(R)](x) + t x^{m+1} \varphi(x).$$

We shall get a contradiction and conclude the proof if we show that, for any small $t > 0$, $p_t(x) \in \,]0, 4[$ for $x \in \,]0, \alpha_{m,k}]$ (since then $\alpha_m(R + t\varphi) > \alpha_{m,k}$ and R would not be a local maximum).

We shall only consider the case when $[\psi_m(R)] (\alpha_{m,k}) = 4$, since the reasoning is similar when $[\psi_m(R)] (\alpha_{m,k}) = 0$. Then $s_0 = -1$ by (25).

- On the interval $]\tau_{1/2}, \alpha_{m,k}]$, $\psi_m(R)$ is greater than a positive constant (since it is positive on $]\tau_1, \alpha_{m,k}]$ by the definition of τ_1 and $\tau_1 < \tau_{1/2} < \alpha_{m,k}$). On the other hand, on this interval, φ is negative (since $s_0 < 0$) and bounded. Therefore, for $t > 0$ small enough, $p_t \in \,]0, 4[$ on this interval.
- Since $\varphi(\tau_{1/2}) = 0$, $p_t(\tau_{1/2}) = [\psi_m(R)] (\tau_{1/2})$, which is in $]0, 4[$ by the definition of $\tau_{1/2}$ in (26).
- On the interval $]\tau_{3/2}, \tau_{1/2}[$, $\psi_m(R)$ is less than a constant < 4 (since it is < 4 on $]\tau_2, \tau_{1/2}]$ by the definition of τ_2 and $\tau_{1/2}$, see 2 and (26), and $\tau_2 < \tau_{3/2} < \tau_1 < \tau_{1/2}$). On the other hand, φ is positive and bounded on this interval. Therefore, for $t > 0$ small enough, $p_t \in \,]0, 4[$ on this interval.

We proceed similarly for the other points $\tau_{j+1/2}$ $(j = 1, \ldots, M - 1)$ and intervals $]\tau_{j+3/2}, \tau_{j+1/2}[$ $(j = 1, \ldots, M - 2)$. Let us now consider the interval $]0, \tau_{M-1/2}[$, which contains tangent points that are all at $y = 0$ or all at $y = 4$.

- If $s_M > 0$ then, on the considered interval, the tangent points are all at $y = 0$, $\psi_m(R)$ is less than a constant < 4, and φ is positive. It results that, for $t > 0$ small enough, $p_t(\cdot) \in \,]0, 4[$ on the interval.
- If $s_M < 0$ then, on the considered interval, the tangent points are all at $y = 4$, $\psi_m(R)$ is positive, and φ is negative. Since the map $x \mapsto [\psi_m(R)] (x)/x = 1 + c_1 x + \ldots$ is greater than a positive constant on the considered interval, the map $x \mapsto [\psi_m(R)] (x)/x + t x^m \varphi(x) = p_t(x)/x$ is also positive on the interval for $t > 0$ sufficiently small. It results that, for $t > 0$ small enough, $p_t(\cdot) \in \,]0, 4[$ on the considered interval. $\qquad \square$

Our next result shows that the necessary optimality conditions of Theorem 4 are also sufficient. We shall need the following lemma on polynomials.

Lemma 4. *If $P \in \mathbf{P}_{k-1}$ takes alternatively nonnegative and non-positive values at $k+1$ successive distinct points, then $P = 0$.*

Proof. Without loss of generality, we can assume that, for points $x_0 < x_1 < \cdots < x_k$, there hold

$$(-1)^j P(x_j) \geq 0, \quad \text{for } j = 0, 1, \ldots, k. \tag{27}$$

Let us introduce the set of indices

$$\mathbf{I}(P) = \{j \in \{0, 1, \ldots, k\} \mid P(x_j) = 0\}.$$

When $\mathbf{I}(P) = \{0, 1, \ldots, k\}$ (resp. $\mathbf{I}(P) = \emptyset$), the conclusion is straightforward since then P has $k+1$ (resp. k) roots.

Suppose now that $\mathbf{I}(P) \neq \emptyset$ and $\mathbf{I}(P) \neq \{0, 1, \ldots, k\}$. Let us introduce the Lagrange interpolation polynomials associated with the x_j's:

$$P_l(x) = \prod_{\substack{j \in \mathbf{I}(P) \\ j \neq l}} \frac{(x - x_j)}{(x_l - x_j)}.$$

Note that all the P_l's belong to $\mathbf{P}_{k-1}$ since $\mathbf{I}(P)$ contains at most k points. For $\varepsilon > 0$, we introduce

$$P_\varepsilon = P + \varepsilon \sum_{l \in \mathbf{I}(P)} (-1)^l P_l$$

and note that
$$\forall j \in \mathbf{I}(P), \quad (-1)^j P_\varepsilon(x_j) = \varepsilon > 0.$$

On the other hand, since $P_\varepsilon \to P$ uniformly on $[x_0, x_k]$, there exists $\varepsilon_0 > 0$ such that
$$\forall \varepsilon < \varepsilon_0, \quad \forall j \notin \mathbf{I}(P), \quad (-1)^j P_\varepsilon(x_j) > 0.$$

Therefore, for $\varepsilon < \varepsilon_0$, P_ε satisfies (27) with, moreover, $\mathbf{I}(P_\varepsilon) = \emptyset$. This implies that $P_\varepsilon = 0$. By taking the limit when ε tends to 0, we get $P = 0$ (actually this contradicts the fact that $\mathbf{I}(P)$ can be nonempty and different from $\{0, \ldots, k\}$). □

Theorem 5 (Sufficient condition of optimality). *Suppose that $P = \psi_m(R)$, for some $R \in \mathbf{P}_{k-1}$, have k tangent points $\{\tau_j\}_{j=1}^k$ such that $0 < \tau_k < \cdots < \tau_1 < \tau_0 = \alpha_m(R)$ and $P(\tau_j) + P(\tau_{j+1}) = 4$ for $j = 0, \ldots, k-1$. Then R is optimal for problem (20).*

Proof. Let $P_{m,k} = \psi_m(R_{m,k})$ be an optimal polynomial (Corollary 1). The difference $D = R - R_{m,k} \in \mathbf{P}_{k-1}$ takes at $x > 0$ the value

$$D(x) = \frac{P(x) - P_{m,k}(x)}{x^{m+1}}.$$

Since $R_{m,k}$ is optimal, $P_{m,k}(\tau_j) \in [0,4]$ for $j = 0,\ldots,k$. Then $D(\tau_j) \geq 0$ (resp. $D(\tau_j) \leq 0$) when $P(\tau_j) = 4$ (resp. $P(\tau_j) = 0$). Since $P(\tau_j), j = 0,\ldots,k$, alternates in $\{0,4\}$, we have shown that

$$(-1)^j \left(P(\tau_0) - 2 \right) D(\tau_j) \geq 0, \quad \text{for } j = 0,\ldots,k.$$

These inequalities tell us that $D \in \mathbf{P}_{k-1}$ satisfies the conditions of Lemma 4. Therefore, $D = 0$ proving that R is optimal. $\qquad\square$

The necessary and sufficient optimality conditions of Theorems 4 and 5 will be used to determine the optimal polynomials in Section 4. We conclude this section with two corollaries of these optimality conditions. The first one deals with the uniqueness of the solution. The second one provides a full description of the optimal polynomials when $m = 1$, relating them to the Chebyshev polynomials of the first kind [Che66, LT86, Wei06].

Corollary 2 (Uniqueness of the solution). *The maximization problem* (20) *has one and only one solution. It has no other local maximum.*

Proof. Existence has been quoted in Corollary 1. Uniqueness is is actually a by-product of the proof of Theorem 5, where it is shown that if a polynomial $P = \psi_m(R)$, for some $R \in \mathbf{P}_{k-1}$, satisfies the optimality conditions (this is the case for any local maximum, by Theorem 4), then R is equal to an arbitrarily fixed solution. Hence there cannot be more than one solution or local maximum. $\qquad\square$

Corollary 3 (Optimal polynomials when $m = 1$). *For $k \geq 0$,*

$$\alpha_{1,k} = 4(k+1)^2 \tag{28}$$

and the optimal polynomial $\psi_1(R_{1,k})$ takes at $x \in [0, \alpha_{1,k}]$ the value

$$[\psi_1(R_{1,k})](x) = 2 \left[1 - T_{k+1} \left(1 - \frac{2x}{\alpha_{1,k}} \right) \right], \tag{29}$$

where T_k denotes the Chebyshev polynomial of the first kind and degree k, which verifies $T_k(x) = \cos(k \arccos x)$ for $x \in [-1, 1]$.

Proof. Let $\alpha_{1,k}$ be defined by (28) and let φ be the function defined at $x \in [0, \alpha_{1,k}]$ by the right-hand side of (29). The fact that $\varphi \equiv \psi_1(R_{1,k})$ will result from the following observations:

- $\varphi \in \psi_1(\mathbf{P}_{k-1})$. Indeed, $\varphi \in \mathbf{P}_{k+1}$. On the other hand, the above formula of T_k shows that $T_k'(1) = k^2$, so that $\varphi'(0) = 4T_{k+1}'(1)/\alpha_{1,k} = 1$, which indicates that the coefficient of x in φ is the one of Q_1.
- The formula of T_k clearly shows that $\varphi(x) \in [0,4]$ for $x \in [0, \alpha_{1,k}]$. On the other hand, $\varphi(\alpha_{1,k}) = 2[1 + (-1)^k]$ and $\varphi'(\alpha_{1,k}) = 4T_{k+1}'(-1)/\alpha_{1,k} = (-1)^k$, so that φ gets out of $[0,4]$ at $x = \alpha_{1,k}$.
- The formula of T_k shows that

$$\varphi(\tau) = 0 \quad \text{when } \tau = 2(k+1)^2 \left(1 - \cos \frac{2j\pi}{k+1}\right), \qquad 0 < 2j < k+1,$$

$$\varphi(\tau) = 4 \quad \text{when } \tau = 2(k+1)^2 \left(1 - \cos \frac{(2j+1)\pi}{k+1}\right), \quad 0 < 2j+1 < k+1,$$

in which $j \in \mathbf{N}$. Therefore, φ has k tangent points in $]0, \alpha_{1,k}[$, at which φ takes alternatively the value 4 and 0.

Using the last observation and the fact that $\varphi(\alpha_{1,k}) = 2[1 + (-1)^k]$ ($= 0$ if k is odd and $= 4$ if k is even), we show that φ satisfies the sufficient optimality conditions (Theorem 5). Hence $\varphi = \psi_1(R_{1,k})$. $\square$

Remark 4. A natural question is whether the number of tangent points of an optimal polynomial $\psi_m(R_{m,k})$ can be greater than k. The answer to this question depends actually on the coefficients of $x^0, \ldots, x^m$, which are fixed in the optimization process. We do not know the answer when the coefficients are those of the polynomial Q_m, but for other coefficients the number of tangent points can be greater than k. The argument is the following. Let $[\psi_{m-1}(R_{m-1,2})](x) = Q_{m-1}(x) + x^m(r_0 + r_1 x)$ be the optimal polynomial with $m-1$ fixed and two free coefficients. By the previous theorem, it has at least two tangent points. Now, consider the function $\tilde{\psi}_m$ obtained by replacing in ψ_m defined by (22), Q_m by the polynomial $x \mapsto Q_{m-1}(x) + r_0 x^m$. Clearly the optimal polynomial associated with $\tilde{\psi}_m$ on $\mathbf{P}_0$ is $\tilde{\psi}_m(\tilde{R}_{m,1})$ where $\tilde{R}_{m,1}$ is the constant r_1. Therefore, $\tilde{\psi}_m(\tilde{R}_{m,1}) = \psi_{m-1}(R_{m-1,2})$ has two tangent points, although the minimization has been done on $\mathbf{P}_0$. $\square$

Remark 5. When checking optimality by looking at the alternate character of $[\psi_m(R)](\tau_j)$ in $\{0,4\}$, one has to include the point $\tau_0 = \alpha_m(R)$. In particular, when $k = 1$, a polynomial with a single tangent point may not be optimal. An example with $m = 4$ and $k = 1$ is shown in Figure 2. The optimal polynomial, given by

$$[\psi_4(R_{4,1})](x) = x - \frac{x^2}{12} + \frac{x^3}{360} - \frac{x^4}{20160} + rx^5 \quad \text{with } r \simeq 4.28 \times 10^{-7},$$

is represented by the solid curve; the dashed curve is Q_4. The optimal polynomial $[\psi_4(R_{4,1})]$ has only one tangent point $\tau_1 \simeq 33,39$, while $\tau_0 = \alpha_{4,1} \simeq 44.03$. As predicted by Theorem 4, $[\psi_4(R_{4,1})](\tau_1) + [\psi_4(R_{4,1})](\tau_0) = 4$. Now, by increasing r to $r \simeq 5.13 \times 10^{-7}$, one gets the dash-dotted curve, which
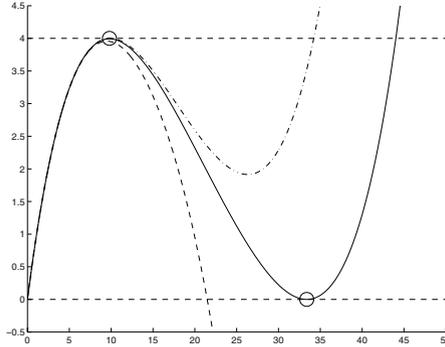
Fig. 2. Checking the sufficient condition of optimality for $m = 4$ and $k = 1$

has a tangent point at $\tau_1 \simeq 9.88$, but is not optimal since the value of the polynomial at this point does not satisfy $[\psi_4(R_{4,1})](\tau_1) + [\psi_4(R_{4,1})](\tau_0) = 4$ (for this polynomial $\tau_0 \simeq 34.22$). □

4 Computational Issues

4.1 Algorithm Based on the Parametrization by the Tangent Points

In the numerical results discussed below, the optimal polynomial is searched by its k alternate tangent points $(\tau_j)_{1 \le j \le k}$, with $\tau_1 > \tau_2 > \cdots > \tau_k$, whose existence is ensured by Theorem 4. These points are determined in the following manner. For $\tau = (\tau_1, \ldots, \tau_k)$, let $R(\tau)$ be the polynomial in $\mathbf{P}_{k-1}$ satisfying

$$\psi_m(R(\tau)) = v \in \mathbb{R}^k,$$

in which the components of v take alternatively the values 0 and 4. Whether one has to impose $v_1 = 0$ or $v_1 = 4$ is further discussed below. The coefficients $r = (r_0 \ \cdots \ r_{k-1})^{\mathrm{T}}$ of $R(\tau)$ are uniquely determined by the equation above, which can also be written

$$\begin{pmatrix} \tau_1^{m+1} & \cdots & \tau_1^{m+k} \\ \vdots & & \vdots \\ \tau_k^{m+1} & \cdots & \tau_k^{m+k} \end{pmatrix} r = v - \begin{pmatrix} [\psi_m(0)](\tau_1) \\ \vdots \\ [\psi_m(0)](\tau_k) \end{pmatrix}. \tag{30}$$

Next, let us introduce the function $F : \tau \in \mathbb{R}^k \mapsto F(\tau) \in \mathbb{R}^k$, where the components of $F(\tau)$ are the derivatives of the polynomial $\psi_m(R(\tau))$ at the τ_j's:

$$F(\tau) = \begin{pmatrix} [\psi_m(R(\tau))]'(\tau_1) \\ \vdots \\ [\psi_m(R(\tau))]'(\tau_k) \end{pmatrix}.$$

Obviously, there holds $F(\tau) = 0$ if τ is the vector of the alternate tangent points of the optimal polynomial. We propose to determine the root(s) τ of F by Newton's method (see [Deu04, BGLS06], for instance). The procedure could have been improved by using a version of Newton's method that exploits inequalities (see, for example, [Kan01, BM05] and the references thereof) to impose $\tau_1 > \tau_2 > \cdots > \tau_k$ as well as the curvature of the solution polynomial at the tangent points: $[\psi_m(R(\tau))]''(\tau_j)(2 - v_j) \geq 0$, for $1 \leq j \leq k$. We have not adopted this additional sophistication, however.

The Newton method requires the computation of $F'(\tau)$. If we denote by $r_l(\tau)$, $1 \leq l \leq k$, the coefficients of $R(\tau)$, by δ_{ij} the Kronecker symbol, and by $V_k(\tau)$ the Vandermonde matrix of order k, there holds

$$\frac{\partial F_i}{\partial \tau_j}(\tau) = \delta_{ij}\,[\psi_m(R(\tau))]''(\tau_i) + \sum_{l=1}^{k} \frac{\partial r_l}{\partial \tau_j}(\tau)(m+l)\tau_i^{m+l-1}$$

$$= \delta_{ij}\,[\psi_m(R(\tau))]''(\tau_i)$$
$$+ [\mathrm{Diag}(\tau_1^m, \ldots, \tau_k^m)V_k(\tau)\,\mathrm{Diag}((m+1), \ldots, (m+k))r'(\tau)]_{ij}.$$

To get an expression of $r'(\tau)$, let us differentiate with respect to τ_j the identity $[\psi_m(R(\tau))](\tau_i) = v_i$. It results

$$\delta_{ij}[\psi_m(R(\tau))]'(\tau_i) + \left(\tau_i^{m+1} \cdots \tau_i^{m+k}\right)\frac{\partial r}{\partial \tau_j}(\tau) = 0.$$

Denoting by $M(\tau)$ the coefficient matrix of the linear system (30), we get

$$r'(\tau) = -M(\tau)^{-1}\,\mathrm{Diag}\left([\psi_m(R(\tau))]'(\tau_1), \ldots, [\psi_m(R(\tau))]'(\tau_k)\right)$$
$$= -M(\tau)^{-1}\,\mathrm{Diag}(F(\tau)).$$

Therefore,

$$F'(\tau) = \mathrm{Diag}\left([\psi_m(R(\tau))]''(\tau_1), ..., [\psi_m(R(\tau))]''(\tau_k)\right)$$
$$- \mathrm{Diag}(\tau_1^m, ..., \tau_k^m)V_k(\tau)\,\mathrm{Diag}((m+1), ..., (m+k))M(\tau)^{-1}\,\mathrm{Diag}(F(\tau)).$$

Observe that at a solution τ^* the second term above vanishes, so that $F'(\tau^*)$ is diagonal. It is also nonsingular if the second derivatives $[\psi_m(R(\tau^*))]''(\tau_j^*)$ are nonzero. Around such a solution, Newton's method is, therefore, well defined.

In the numerical results presented below, we have used the solver of nonlinear equations `fsolve` of Matlab (version 7.2), which does not take into account the inequality constraints. The vector v has been determined by adopting the following heuristics. We have *assumed* that the optimal polynomial is negative for all $x < 0$ (it has unit slope at $x = 0$), which implies that r_k, the

coefficient of x^{m+k} of the optimal polynomial, has the sign $(-1)^{m+k+1}$; if the assumption is correct, the optimal polynomial should get out of the interval at $y = 0$ if $m+k$ is even and at $y = 4$ if $m+k$ is odd; according to Theorem 4, one should, therefore, take $v_1 = 4 - \varepsilon_v$ if $m + k$ is even and $v_1 = \varepsilon_v$ if $m + k$ is odd. The value of ε_v is taken nonnegative and as close as possible to 0. A positive value of ε_v is usually necessary for counterbalancing rounding errors. The other values of v_i alternate in $\{\varepsilon_v, 4 - \varepsilon_v\}$. The initial point τ is chosen by trials and errors, or according to suggestions made in the discussion below.

The proposed approach has the following advantages $(+)$ and disadvantages $(-)$:

$+$ The problem has few variables (just k).
$+$ The problem looks well conditioned, provided the second derivatives at the tangent points are reasonable, which seems to be the case.
$-$ There is no guarantee that the solution found is the optimal one since a zero of F will not be a solution to the original problem if the polynomial gets out of $[0, 4]$ at a point τ_0 less than τ_1. An example of this situation is given in Figure 3. However, if $\tau_0 > \tau_1$ and if $[\psi_m(R)](\tau_0) + [\psi_m(R)](\tau_1) = 4$, the sufficient optimality conditions of Theorem 5 guarantee that R is the solution.
$-$ The solution polynomial may get out of the interval $[0, 4]$ near a tangent point due to the lack of precision of the solution, which has motivated the use of the small $\varepsilon_v > 0$.
$-$ Obtaining the convergence to a zero of F (not only a stationary point τ^* of $\|F\|_2^2$, hence verifying $F'(\tau^*)^{\mathrm{T}} F(\tau^*) = 0$) depends on the initialization of the iterative process.
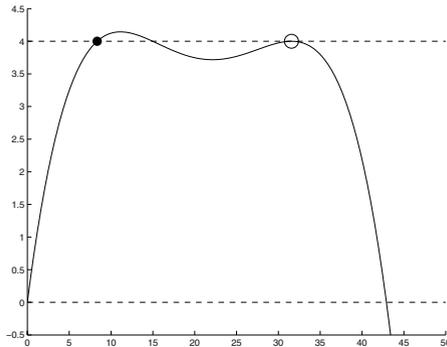

Fig. 3. A zero of F that is not an optimal polynomial ($m = 3$, $k = 1$).

4.2 Numerical Results

Computing $\alpha_{m,k}$

Table 1 shows the computed values of $\alpha_{m,k}$ for $1 \le m \le 8$ and $0 \le k \le 8$. The computed solutions were always satisfying the optimality conditions, so that we are pretty confident in the values of $\alpha_{m,k}$ in the table. In particular, the small $\varepsilon_v > 0$ hardly modifies these values.

The column $k = 0$ of Table 1 corresponds to the polynomials Q_m defined by (11), for which the first values of the $\alpha_{m,0}$'s were already given in (13) (there denoted α_m). We observe that the convergence of $\alpha_{2m+1,0}$ (resp. $\alpha_{2m,0}$) to $\pi^2 \simeq 9.87$ (resp. $4\pi^2 \simeq 39.48$), predicted by Theorem 3, is rather fast. On the other hand, we observe that the values $\alpha_{m,k}$ can be made spectacularly larger than $\alpha_{m,0}$, which was our objective.

We have verified that the optimal polynomials corresponding to $m = 1$ are, indeed, related to the Chebyshev polynomials through formula (29), as claimed by Corollary 3. This fact can be observed in the first row of the table, whose values of $\alpha_{1,k}$ are, indeed, those given by (28).

Another observation is that the oscillating behaviour of α_m with m, highlighted in the analysis leading to Theorem 3, is recovered in the sequences $\{\alpha_{m,k}\}_{m \ge 1}$. The reason is similar. The first positive stationary point of the optimal polynomial, which is close to the one of Q_∞, is (resp. is not) a tangent point when m is odd (resp. even). This observation leads to the following conjecture: if we denote by $\tau_{m,k,j}$ the jth tangent point of the optimal polynomial $\psi_m(R_{m,k})$ ($1 \le j \le k$), then, when m goes to infinity, $\tau_{2m+1,k,j}$ (resp. $\tau_{2m,k,j}$) converges the jth (resp. $(j+1)$th) positive stationary point of Q_∞, the polynomial defined by (14). More specifically,

$$\tau_{2m+1,k,j} \to j^2\pi^2 \quad \text{and} \quad \tau_{2m,k,j} \to (j+1)^2\pi^2, \quad \text{when } m \to \infty. \qquad (31)$$

In practice, these values can be used to choose a good starting point for the algorithm when m is large.

Table 1. Computed values of the first $\alpha_{m,k}$'s

	$k=0$	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$	$k=8$
$m=1$	4.00	16.00	36.00	64.00	100.00	144.00	196.00	256.00	324.00
$m=2$	12.00	32.43	60.56	96.61	140.64	192.66	252.67	320.68	396.69
$m=3$	7.57	23.40	45.72	75.06	111.58	155.38	206.51	265.04	331.00
$m=4$	21.48	44.03	73.45	110.01	153.83	204.98	263.51	329.49	402.92
$m=5$	9.53	31.61	58.23	90.77	129.90	175.84	228.71	288.59	355.23
$m=6$	30.72	57.23	89.78	128.89	174.84	227.71	287.61	354.59	428.71
$m=7$	9.85	37.37	68.93	108.35	151.08	199.56	255.61	317.90	357.95
$m=8$	37.08	70.89	107.67	150.35	199.32	254.89	317.22	386.35	462.27

Diagonal schemes $k = m$

We have found interesting to have a particular look at the case $k = m$. First it gives a computational effort per time step that is twice the one for the original $(2m)$th order scheme, which corresponds to $k = 0$. The second reason is more related to intuition: if one wants to get $\alpha_{m,k}$ roughly proportional to m^2, we have to control the first m maxima or minima of the optimal polynomial $\psi_m(R_{m,k})$, for which we think that we need m parameters, which corresponds to $k = m$. Below, we qualify such a scheme as *diagonal*.

Figure 4 shows the optimal polynomials $\psi_m(R_{m,m})$, for $m = 1, \ldots, 8$. The tangent points are quoted by circles on the graphs, while the $\alpha_{m,m}$'s are quoted by dots.

Table 2 investigates the asymptotic behaviour of the diagonal schemes:

1. Its first column highlights the growth of the ratio between the maximum time step allowed by the stability analysis in a diagonal scheme $\Delta t_{m,m}$ and in the second order scheme $\Delta t_{1,0}$. According to Section 2.2, there holds

$$\frac{\Delta t_{m,m}}{\Delta t_{1,0}} = \left(\frac{\alpha_{m,m}}{\alpha_{1,0}}\right)^{1/2} = \frac{\alpha_{m,m}^{1/2}}{2}. \tag{32}$$

2. The computational cost $C_{m,m}(T)$ of the diagonal scheme of order $2m$ on an integration time T is proportional to the computational cost $C_{m,m}^1$ of one time step multiplied by the number of time steps. Hence, assuming that the largest time step allowed by the stability analysis is taken, one has

$$C_{m,m}(T) \simeq \frac{C_{m,m}^1 T}{\Delta t_{m,m}}.$$

A similar expression holds for the computational cost $C_{1,0}(T)$ of the second order scheme, with $C_{m,m}^1$ and $\Delta t_{m,m}$ replaced by $C_{1,0}^1$ and $\Delta t_{1,0}$, respectively. The second column of Table 2 gives the ratio of these two costs. Using (32) and the fact that $C_{m,m}^1 \simeq 2m C_{1,0}^1$ (each time step of the diagonal scheme requires $2m$ times more operator multiplications than each time step of the second order scheme), the ratio can be estimated by

$$\frac{C_{m,m}(T)}{C_{1,0}(T)} \simeq \frac{4m}{\alpha_{m,m}^{1/2}}.$$

The numbers in the second column of Table 2 suggest that this ratio is bounded. If the conjecture (33) below is correct, it should converge to $4\sqrt{2}/\pi \simeq 1.80$, when m goes to infinity.

3. Taking $k = m$ and $j = \lceil m/2 \rceil$ in (31), and assuming that $\alpha_{m,m} \sim 2\tau_{m,m,\lceil m/2 \rceil}$ (suggested by the approximate symmetry of the optimal polynomials) lead us to the following conjecture:

$$\frac{\alpha_{m,m}}{m^2} \to \frac{\pi^2}{2}, \quad \text{when } m \to \infty. \tag{33}$$
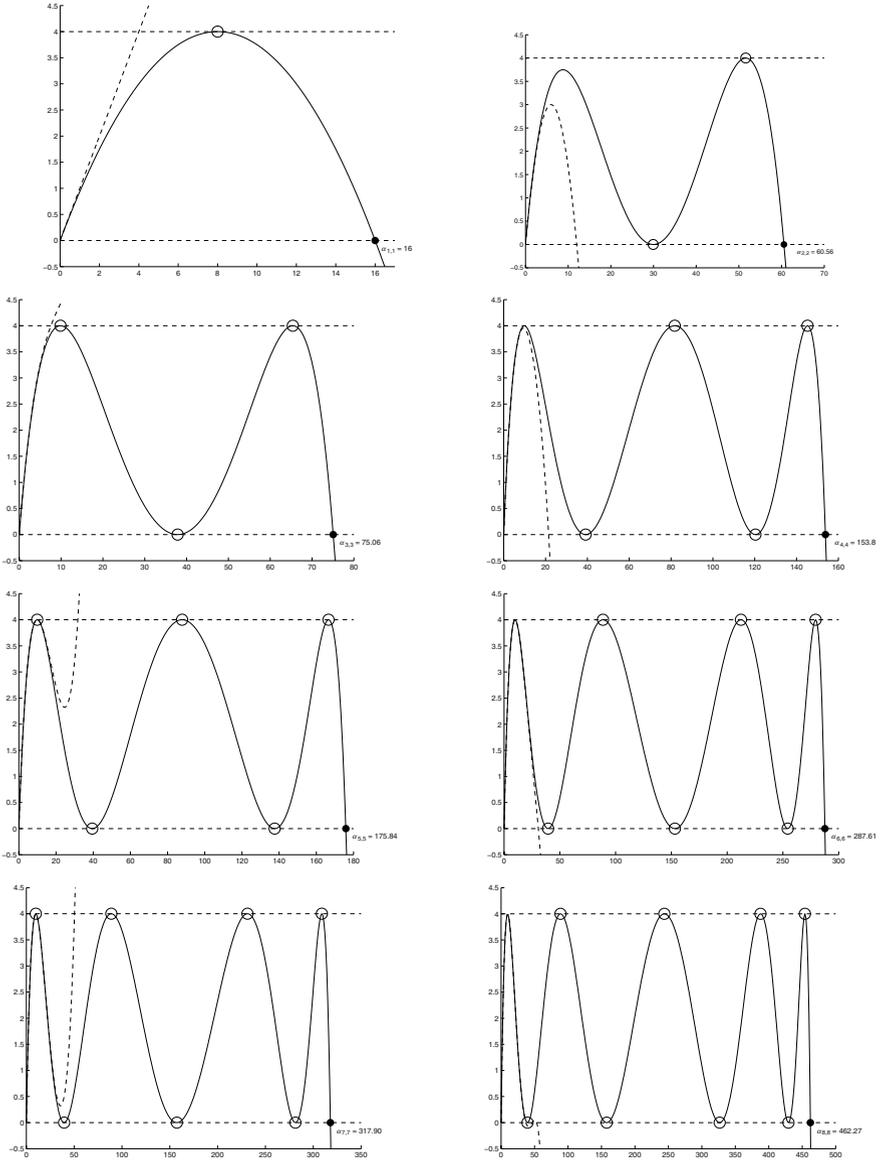
Fig. 4. The polynomials $Q_m = \psi_m(0)$ (dashed curves) and the optimal polynomials $\psi_m(R_{m,m})$ for $m = 1, \ldots, 8$ (solid curves)

Table 2. Asymptotic behaviour of the diagonal schemes

m	$\dfrac{\Delta t_{m,m}}{\Delta t_{m,0}}$	$\dfrac{C_{m,m}(T)}{C_{1,0}(T)}$	$\dfrac{2\alpha_{m,m}}{m^2\pi^2}$
1	2.00	1.00	3.24
2	3.89	1.03	3.07
3	4.33	1.39	1.69
4	6.20	1.29	1.95
5	6.63	1.51	1.43
6	8.48	1.42	1.62
7	8.91	1.57	1.31
8	10.75	1.49	1.46
∞		1.80	1.00

This conjecture is explored numerically in the third column of Table 2. Note that it does not distinguish between even and odd values of m, at least asymptotically. However, looking at the $\alpha_{m,m}$'s on the diagonal of Table 1, it appears that the even values of $k = m$ look more interesting than the odd ones.

5 Conclusion

In this paper, we have analyzed the stability of higher order time discretization schemes for second order hyperbolic problems based on the modified equation approach. We have in particular proven that the upper bound for the time step (the CFL limit) remains uniformly bounded for large m ($2m$ is the order of the scheme). On the basis of this information, we have proposed the construction of new schemes that are seen as modifications of the previous ones and are designed in order to optimize the CFL condition: this is formulated as an optimization problem in a space of polynomials of given degree. Despite some unpleasant properties (the objective function is non-convex and even discontinuous at the solution!), this problem can be fully analyzed. In particular, we prove the existence and uniqueness of the solution and give necessary and sufficient conditions of optimality. These conditions are exploited to design an algorithm for the effective numerical solution of the optimization problem. The obtained results are more than satisfactory with respect to our original objective. They suggest some conjectures that would mean that we would be able to produce schemes of arbitrary high order in time and whose computational cost would be almost independent of the order.

Of course, this is a preliminary work and much has still to be done, including the following items:

- The effective efficiency of the new schemes should be tested on realistic wave propagation problems.
- The impact of the modification of the initial schemes (the ones which are based on the modified equation technique) on the effective accuracy (we are only guaranteed that the order of approximation is preserved) should be analyzed thorough numerical dispersion studies.
- Our various theoretical conjectures should be addressed in a rigorous way.

These will be the subjects of forthcoming works.

References

[AJT00]   L. Anné, P. Joly, and Q. H. Tran. Construction and analysis of higher order finite difference schemes for the 1D wave equation. *Comput. Geosci.*, 4(3):207–249, 2000.

[AKM74]   R. M. Alford, K. R. Kelly, and Boore D. M. Accuracy of finite difference modeling of the acoustic wave equation. *Geophysics*, 39:834–842, 1974.

[BGLS06]   J. F. Bonnans, J. Ch. Gilbert, C. Lemaréchal, and C. Sagastizábal. *Numerical Optimization – Theoretical and Practical Aspects.* Universitext. Springer Verlag, Berlin, 2nd edition, 2006.

[BM05]   S. Bellavia and B. Morini. An interior global method for nonlinear systems with simple bounds. *Optim. Methods Softw.*, 20(4–5):453–474, 2005.

[CdLBL97]   R. Carpentier, A. de La Bourdonnaye, and B. Larrouturou. On the derivation of the modified equation for the analysis of linear numerical methods. *RAIRO Modél. Math. Anal. Numér.*, 31(4):459–470, 1997.

[CF05]   G. Cohen and S. Fauqueux. Mixed spectral finite elements for the linear elasticity system in unbounded domains. *SIAM J. Sci. Comput.*, 26(3):864–884 (electronic), 2005.

[Che66]   E. W. Cheney. *Introduction to Approximation Theory.* McGraw-Hill, 1966.

[CJ96]   G. Cohen and P. Joly. Construction analysis of fourth-order finite difference schemes for the acoustic wave equation in nonhomogeneous media. *SIAM J. Numer. Anal.*, 33(4):1266–1302, 1996.

[CJKMVV99]   M. J. S. Chin-Joe-Kong, W. A. Mulder, and M. Van Veldhuizen. Higher-order triangular and tetrahedral finite elements with mass lumping for solving the wave equation. *J. Engrg. Math.*, 35(4):405–426, 1999.

[CJRT01]   G. Cohen, P. Joly, J. E. Roberts, and N. Tordjman. Higher order triangular finite elements with mass lumping for the wave equation. *SIAM J. Numer. Anal.*, 38(6):2047–2078 (electronic), 2001.

[Coh02]   G. C. Cohen. *Higher-order numerical methods for transient wave equations.* Scientific Computation. Springer-Verlag, Berlin, 2002.

[Dab86]   M. A. Dablain. The application of high order differencing for the scalar wave equation. *Geophysics*, 51:54–56, 1986.

[Deu04]   P. Deuflhard. *Newton Methods for Nonlinear Problems – Affine Invariance and Adaptive Algorithms*. Number 35 in Computational Mathematics. Springer, Berlin, 2004.

[DPJ06]   S. Del Pino and H. Jourdren. Arbitrary high-order schemes for the linear advection and wave equations: application to hydrodynamics and aeroacoustics. *C. R. Math. Acad. Sci. Paris*, 342(6):441–446, 2006.

[FLLP05] L. Fezoui, S. Lanteri, S. Lohrengel, and S. Piperno. Convergence and stability of a discontinuous Galerkin time-domain method for the 3D heterogeneous Maxwell equations on unstructured meshes. *M2AN Math. Model. Numer. Anal.*, 39(6):1149–1176, 2005.

[HW96]   E. Hairer and G. Wanner. *Solving ordinary differential equations. II*, volume 14 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 2nd edition, 1996. Stiff and differential-algebraic problems.

[HW02]   J. S. Hesthaven and T. Warburton. Nodal high-order methods on unstructured grids. I. Time-domain solution of Maxwell's equations. *J. Comput. Phys.*, 181(1):186–221, 2002.

[Jol03]   P. Joly. Variational methods for time-dependent wave propagation problems. In *Topics in computational wave propagation*, volume 31 of *Lect. Notes Comput. Sci. Eng.*, pages 201–264. Springer, Berlin, 2003.

[Kan01]   Ch. Kanzow. An active set-type Newton method for constrained non-linear systems. In M.C. Ferris, O.L. Mangasarian, and J.S. Pang, editors, *Complementarity: applications, algorithms and extensions*, pages 179–200, Dordrecht, 2001. Kluwer Acad. Publ.

[LT86]   P. Lascaux and R. Théodor. *Analyse Numérique Matricielle Appliquée à l'Art de l'Ingénieur*. Masson, Paris, 1986.

[PFC05]   S. Pernet, X. Ferrieres, and G. Cohen. High spatial order finite element method to solve Maxwell's equations in time domain. *IEEE Trans. Antennas and Propagation*, 53(9):2889–2899, 2005.

[RM67]   R. D. Richtmyer and K. W. Morton. *Difference methods for initial-value problems*, volume 4 of *Interscience Tracts in Pure and Applied Mathematics*. John Wiley & Sons, Inc., New York, 2nd edition, 1967.

[RS78]   M. Reed and B. Simon. *Methods of modern mathematical physics. IV. Analysis of operators*. Academic Press [Harcourt Brace Jovanovich Publishers], New York, 1978.

[SB87]   G. R. Shubin and J. B. Bell. A modified equation approach to constructing fourth-order methods for acoustic wave propagation. *SIAM J. Sci. Statist. Comput.*, 8(2):135–151, 1987.

[Sch91]   L. Schwartz. *Analyse I – Théorie des Ensembles et Topologie*. Hermann, Paris, 1991.

[TT05]   E. F. Toro and V. A. Titarev. ADER schemes for scalar non-linear hyperbolic conservation laws with source terms in three-space dimensions. *J. Comput. Phys.*, 202(1):196–215, 2005.

[Wei06]   E. W. Weisstein. Chebyshev polynomial of the first kind. *MathWorld*. http://mathworld.wolfram.com/ChebyshevPolynomialoftheFirst Kind.html, 2006.

Comparison of Two Explicit Time Domain Unstructured Mesh Algorithms for Computational Electromagnetics

Igor Sazonov, Oubay Hassan, Ken Morgan, and Nigel P. Weatherill

Civil and Computational Engineering Centre, School of Engineering, University of Wales, Swansea SA2 8PP, Wales, UK
{i.sazonov,O.Hassan,K.Morgan,N.P.Weatherill}@swansea.ac.uk

Summary. An explicit finite element time domain method and a co-volume approach, based upon a generalization of the well-known finite difference time domain scheme of Yee to unstructured meshes, are employed for the solution of Maxwell's curl equations in the time domain. A stitching method is employed to produce meshes that are suitable for use with a co-volume algorithm. Examples, involving EM wave propagation and scattering, are included and the numerical performance of the two techniques is compared.

Key words: computational electromagnetics, Delaunay triangulation, Voronoï tessellation, co-volume mesh generation, explicit schemes, finite element method, co-volume method, EM wave propagation and scattering

1 Introduction

Computational methods are widely employed for the solution of Maxwell's equations in a variety of different application areas that fall within the general field of electromagnetics. For practical applications, the requirement of modelling complex geometries means that unstructured mesh methods are particularly attractive, as fully automatic unstructured mesh generation procedures are now widely available [Geo91, WH94, PPM99]. Following this philosophy requires the identification of a suitable unstructured mesh-based solution algorithm and several low-order time domain procedures have been proposed [MSH91, PLD92, CFS93, DL97, MWH$^+$99]. These methods are readily implemented, but may require a significant computational resource to undertake accurate simulations involving wave propagation over a large number of wavelengths [DBB99]. On the other hand, the Yee scheme [Yee66] is a covolume solution technique, on a structured Cartesian mesh, that exhibits a high degree of computationally efficiency, in terms of both CPU and memory requirements.

To provide a practically useful computational procedure, it is natural to attempt to develop hybrid solution procedures, employing an unstructured mesh method in the vicinity of a complex geometry and the co-volume method elsewhere [RBT97, MM98, RB00, EL02, EHM+03]. An alternative approach is to employ an unstructured mesh everywhere and to attempt to use an unstructured mesh implementation of the co-volume scheme [Mad95, GL93]. A basic requirement for the successful implementation of the co-volume scheme is the existence of two, high quality, mutually orthogonal meshes. For an unstructured mesh implementation, the obvious dual mesh choice is the Delaunay–Voronoï diagram. Despite the fact that real progress has been achieved in unstructured mesh generation methods over the last two decades, co-volume schemes have not generally proved to be effective for simulations involving domains of complex shape [NW98]. This is due to the difficulties encountered when attempting to generate sufficiently smooth, high quality dual meshes for such problems. Standard mesh generation methods are designed to create high quality Delaunay triangulations, but do not attempt to provide a high quality dual Voronoï mesh. A stitching method was recently proposed [SWH+06] for the generation of meshes for the co-volume scheme in two dimensions. In this approach, the problem of triangulation of a domain of complicated shape is split into a set of relatively simple problems of local triangulation. Each local mesh is constructed with properties which are close to those of an ideal mesh and the local triangulations are combined, to form a consistent mesh, by using a stitching algorithm. The quality of the stitched mesh is improved by the use of standard mesh quality enhancement methods.

In this paper, we will utilise the meshes produced by the stitching method to compare the efficiency and the accuracy of a co-volume scheme on unstructured meshes and an explicit linear finite element procedure for Maxwell's curl equations [MHP94, MHP96, MHPW00]. The layout of the paper is as follows: Section 2 describes the governing equations. A brief description of the finite element time domain algorithm is given in Section 3, while the implementation of the co-volume scheme on unstructured meshes is described in Section 4. Section 5 provides a brief description of the approach used for the generation of the required meshes. In Section 6, a study of the accuracy and the efficiency of both algorithms is presented for wave propagation and wave scattering examples. Finally, conclusions are drawn in Section 7.

2 Governing Equations

The equations governing the propagation of electromagnetic waves through a free space region may be considered in the dimensionless integral form

$$\frac{\partial}{\partial t} \int_{\Omega} \mathbf{E} \, \mathrm{d}\mathbf{\Omega} = \oint_{\Gamma} \mathbf{H} \, \mathrm{d}\mathbf{\Gamma}, \qquad \frac{\partial}{\partial t} \int_{\Omega} \mathbf{H} \, \mathrm{d}\mathbf{\Omega} = - \oint_{\Gamma} \mathbf{E} \, \mathrm{d}\mathbf{\Gamma} \qquad (1)$$

for an arbitrary surface Ω bounded by a closed contour Γ, or in the corresponding differential form

$$\frac{\partial \mathbf{H}}{\partial t} = -\boldsymbol{\nabla} \times \mathbf{E}, \qquad \frac{\partial \mathbf{E}}{\partial t} = \boldsymbol{\nabla} \times \mathbf{H}. \tag{2}$$

Here, $\mathbf{E}$ and $\mathbf{H}$ denote the electric and magnetic field intensity respectively, $d\boldsymbol{\Omega}$ denotes an element of surface area, in the direction normal to the surface, and $d\boldsymbol{\Gamma}$ is an element of contour length, in the tangent direction to the contour. Consideration will be restricted to the solution of two-dimensional problems, involving TE polarized waves. In this case, relative to a Cartesian x, y, z coordinate system, the field intensity vectors $\mathbf{E} = (E_x, E_y, 0)$ and $\mathbf{H} = (0, 0, H_z)$ are functions of t, x and y only.

The scattering simulations that will be undertaken will involve the interaction between a known incident field, generated by a source located in the far field, and a scatterer, surrounded by free space. It will be assumed that the scatterer is a perfect electrical conductor (PEC) and that the incident field is a plane single frequency wave. For such simulations, it is convenient to split the total electric and magnetic fields as

$$\mathbf{E} = \mathbf{E}^{inc} + \mathbf{E}^{scat}, \qquad \mathbf{H} = \mathbf{H}^{inc} + \mathbf{H}^{scat}, \tag{3}$$

where the subscripts *inc* and *scat* refer to the incident and scattered wave components respectively. The problem is then formulated in terms of the scattered fields. The boundary condition at the surface of the scatterer is the requirement that the tangential component of the total electric field should be zero. The infinite solution domain must be truncated to enable a numerical simulation and the condition that must be imposed at the truncated far field boundary is that the scattered field should only consist of outgoing waves. This requirement is imposed by surrounding the computational domain with an artificial perfectly matched layer (PML) [Ber94, BP97].

3 A Finite Element Method

An explicit finite element time domain (FETD) method, for implementation on a general unstructured mesh of triangles, can be developed by initially writing the equations (2) in the form

$$\frac{\partial \mathbf{U}}{\partial t} = -\frac{\partial \mathbf{F}^k}{\partial x_k} = -\mathbf{A}^k \frac{\partial \mathbf{U}}{\partial x_k}, \tag{4}$$

where k takes the values 1 and 2 and the summation convention is employed. Here $x_1 = x$, $x_2 = y$ and

$$\mathbf{U} = \begin{bmatrix} H_z \\ E_x \\ E_y \end{bmatrix}, \qquad \mathbf{A^k} = \begin{bmatrix} 0 & -(k-1) & (2-k) \\ (k-1) & 0 & 0 \\ -(2-k) & 0 & 0 \end{bmatrix}. \tag{5}$$

This equation is discretised using the explicit TG2 algorithm [DH03]. In this method, the solution is advanced over a time step, Δt, in a two-stage process.

In the first stage, the solution is advanced from time level t_n to time level $t_{n+1/2} = t_n + \Delta t/2$ using the forward difference approximation

$$\mathbf{U}^{(n+1/2)} = \mathbf{U}^{(n)} - \frac{\Delta t}{2} \left(\mathbf{A}^k \frac{\partial \mathbf{U}}{\partial x_k} \right)^{(n)}. \tag{6}$$

Here, the superscript (n) denotes an evaluation at time $t = t_n$. In the second stage, the solution at time level $t_{n+1} = t_n + \Delta t$ is obtained from the central difference approximation

$$\mathbf{U}^{(n+1)} = \mathbf{U}^{(n)} - \Delta t \left(\mathbf{A}^k \frac{\partial \mathbf{U}}{\partial x_k} \right)^{(n+1/2)}. \tag{7}$$

At time $t = t_n$, a continuous piecewise linear approximation, on element e, may be expressed as

$$\mathbf{U}_e^{(n)} = N_{(J)} \mathbf{U}_{(J)}^{(n)}, \tag{8}$$

where $N_{(J)}$ is the piecewise linear shape function associated with node J of the mesh, $\mathbf{U}_{(J)}$ represent nodal values and the implied summations extend over each node J of element e. A variational formulation [ZM06] of the equation (6) is employed to obtain the solution at time level $t = t_{n+1/2}$. To obtain the solution at the end of the time step, at each node I, the weak variational formulation [ZM06]

$$\mathbf{M}_{(IJ)} \mathbf{U}_{(J)}^{(n+1)} = \mathbf{M}_{(IJ)} \mathbf{U}_{(J)}^{(n)} + \mathbf{A}^k \int_{e \in \Omega} \mathbf{U}_e^{(n+1/2)} \frac{\partial N_{(I)}}{\partial x_k} \, \mathrm{d}\Omega - \int_\Gamma \tilde{\mathbf{F}}_n^{(n)} N_{(I)} \, \mathrm{d}\Gamma \tag{9}$$

for the equation (7) is employed over the computational domain, Ω. In the equation (9), Γ denotes the boundary of region Ω, $\tilde{\mathbf{F}}_n$ is a normal boundary flux and $\mathbf{M}_{(IJ)}$ is the standard consistent mass matrix for the mesh of linear triangular elements in Ω. The equation (9) is solved by explicit iteration and the resulting algorithm is stable provided that a CFL condition of the form

$$\Delta t \leq \mathcal{C} \min_e h_e \tag{10}$$

is satisfied, where h_e denotes the minimum height of element e and $\mathcal{C}$ is a safety factor.

For scattering simulations, the boundary condition at the surface of the PEC scatterer is weakly imposed through the Galerkin statement. The truncated far field boundary is taken to be rectangular in shape and a structured grid of triangular elements is used to discretise the PML region.

4 A Co-Volume Method

For the co-volume method, the governing equations are considered in the integral, time domain form of the equation (1) and the discretisation is accomplished using two mutually orthogonal meshes [Mad95, GL93]. For this

purpose, we choose to employ the Delaunay–Voronoï dual diagram, with the integrals taken over the edges of the Delaunay and Voronoï cells. To illustrate the process, consider a triangular element m of the Delaunay mesh. This element will share an edge with N_m elements, with numbers m_i, $1 \leq i \leq N_m$, where $N_m = 3$, unless the element has an edge representing the boundary of the domain. Suppose the Delaunay edge mm_i is the common edge between elements m and m_i and let the length of this edge be denoted by ℓ_{mm_i}. Similarly, suppose that the Voronoï edge mm_i is the line segment connecting the circumcentres of element m and element m_i. The length of this Voronoï edge will be denoted by h_{mm_i}. As basic unknowns in the solution algorithm, we consider the value of the z-component of the magnetic field at the Voronoï vertices, and denote this by H_m, and the projection of the electric field at the midpoint of the Delaunay edge mm_i, in the direction of the edge, and denote this by E_{mm_i}. In this case, the laws of Ampère and Faraday can be approximated, using central differencing, as

$$H_m^{(n+1/2)} = H_m^{(n-1/2)} - \frac{\Delta t}{S_m} \sum_{i=1}^{N_m} E_{mm_i}^{(n)} \ell_{mm_i}, \tag{11}$$

$$E_{mm_i}^{(n+1)} = E_{mm_i}^{(n)} + \frac{\Delta t}{h_{mm_i}} \left[H_m^{(n+1/2)} - H_{m_i}^{(n+1/2)} \right], \tag{12}$$

where S_m is the area of element m. This is a staggered explicit scheme, where the time step size for a stable implementation may be determined from the requirement [TH00]

$$\Delta t < \mathcal{C} \min \left\{ \ell_{\min}, h_{\min} \right\}. \tag{13}$$

Here $\ell_{\min}$ and $h_{\min}$ are the minimum Delaunay and Voronoï edge lengths respectively and $\mathcal{C}$ is a safety factor. This implies the use of meshes which do not include either very short Delaunay, or very short Voronoï, edges. However, Voronoï edge lengths may vanish completely, on a general unstructured mesh, when two adjacent triangles have a common circumcentre. When this happens, the simple remedy is to merge these two triangles to form a single quadrilateral element. The discrete formulae of the equations (11) and (12) may be applied directly to this quadrilateral, with appropriate redefinition of N_m. Moreover, the same merging procedure can be adopted when more than two triangles share a common circumcentre and the discrete equations applied again to the polygonal cell that is created by merging the triangles in this manner. This merging process is illustrated in Figure 1. If the mesh contains short non-zero Voronoï sides, the merging process may still be carried out, to overcome the severe restriction on the time step. However, this will reduce the accuracy of the scheme, due to the slight local non-orthogonality introduced by the merging.

The boundary condition on the tangential component of the electric field can be directly imposed at the surface of the PEC. The far field boundary condition is again approximated by the addition of an artificial PML, with the external boundary of the truncated domain taken to be rectangular in shape.
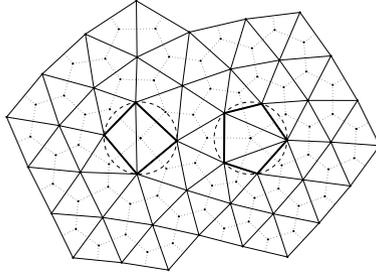
Fig. 1. An example of a Delaunay–Voronoï dual diagram showing two mutually orthogonal meshes suitable for use with a co-volume solution scheme. The dotted lines indicate Voronoï edges and the dots represent Voronoï vertices. Quadrilateral and pentagonal elements, formed by the merging of triangles, are indicated by bold lines.

5 Mesh Generation

With algorithms of the form considered here, wave propagation problems are normally simulated on a mesh, which is as uniform as possible, with a prescribed element size δ which is related to the wavelength. For two-dimensional simulations, in the absence of boundaries, the ideal mesh for the co-volume method is simply a mesh of equilateral triangles, with the Delaunay edge length $l = \delta$. In this case, the Voronoï elements are perfect hexagons, with edge length $h = \delta/\sqrt{3} \approx 0.577\,\delta$. This ideal mesh has the highest quality but, for general scattering simulations, it almost certainly will not be able to represent the geometry of the scatterer. To overcome this problem, a method based on stitching the ideal mesh to a near-boundary unstructured mesh has been developed [SWH$^+$06]. In the vicinity of each boundary, a body fitted local mesh is constructed, with the properties close to those of the ideal mesh. Near-boundary elements are generated by a modified form of the advancing front method. The ideal mesh is employed, away from boundaries, in the major portion of the domain. An additional temporary layer of near-boundary elements is generated to assist the process of connecting the near-boundary mesh to the ideal mesh. The new nodes of this extra layer are marked as potential nodes for connection. For each of these potential nodes, the closest node in the ideal mesh is identified. Joining, consecutively, these identified nodes of the ideal mesh, we obtain a closed polygon, or set of polygons. The gap between the near-boundary elements and the ideal mesh element is triangulated using the Delaunay method. Here, points of the ideal mesh which lie in the gap will also be used during the triangulation. Standard mesh enhancement procedures, such as edge swapping and Laplace smoothing, are used at the end to improve the quality of the generated elements.

6 Numerical Examples

A number of examples will be presented which enable a comparison to be made between the accuracy and the performance of the FETD approach and the co-volume algorithm on unstructured meshes.

6.1 Narrow Waveguide

The first example involves the simulation of the propagation, in the positive x-direction, of a plane harmonic TE wave, of wavelength λ, in a narrow rectangular waveguide. The waveguide occupies the region $0 \leq x \leq 200\lambda$ and its width, 0.4λ, is small enough to avoid the generation of any wave normal to the direction of propagation. Two unstructured meshes, with spacing $\delta \approx \lambda/15$ and $\delta \approx \lambda/30$, are generated using the stitching method. The majority of the elements are almost equilateral triangles which exhibit all the desired mesh quality properties [ZM06]. To enable a comparison with the results produced by the traditional Yee scheme, two structured triangular grids are generated, using the vertex spacings $\delta = \lambda/15$ and $\delta = \lambda/30$. On these meshes, the co-volume scheme of the equations (11) and (12) reduces to the classical Yee scheme. Figure 2 shows the structured mesh with $\delta = \lambda/15$ and the unstructured mesh with $\delta \approx \lambda/15$. The solution is advanced for 170 cycles, using the maximum allowable time step. For each case considered, the computed distribution of the magnetic field, between $x = 139\lambda$ and $x = 141\lambda$, is compared with the exact distribution in Figure 3. It can be seen that the Yee scheme on the structured grid and the co-volume scheme on the unstructured grid maintain the amplitude of the propagating wave, while the FETD scheme fails to maintain the amplitude. It can also be observed that the phase velocity is under-predicted by both the Yee and the co-volume schemes and is over-predicted by the FETD scheme. However, the phase velocity obtained on the unstructured meshes with the co-volume scheme is more accurate than the phase velocity obtained using the traditional Yee scheme on the structured
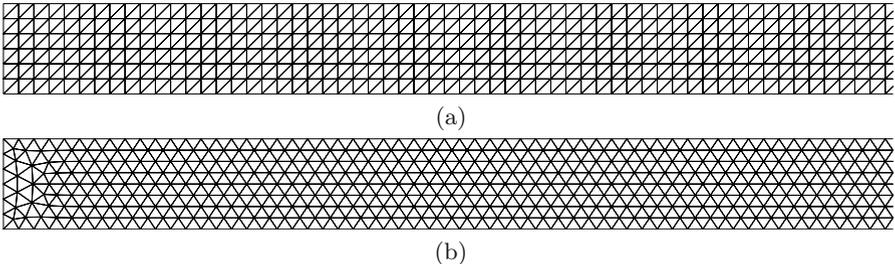

(a)

(b)

Fig. 2. Details of the meshes employed for the propagation of a plane harmonic TE wave in a waveguide: (a) the structured mesh with $\delta = \lambda/15$; (b) the unstructured mesh with $\delta \approx \lambda/15$.
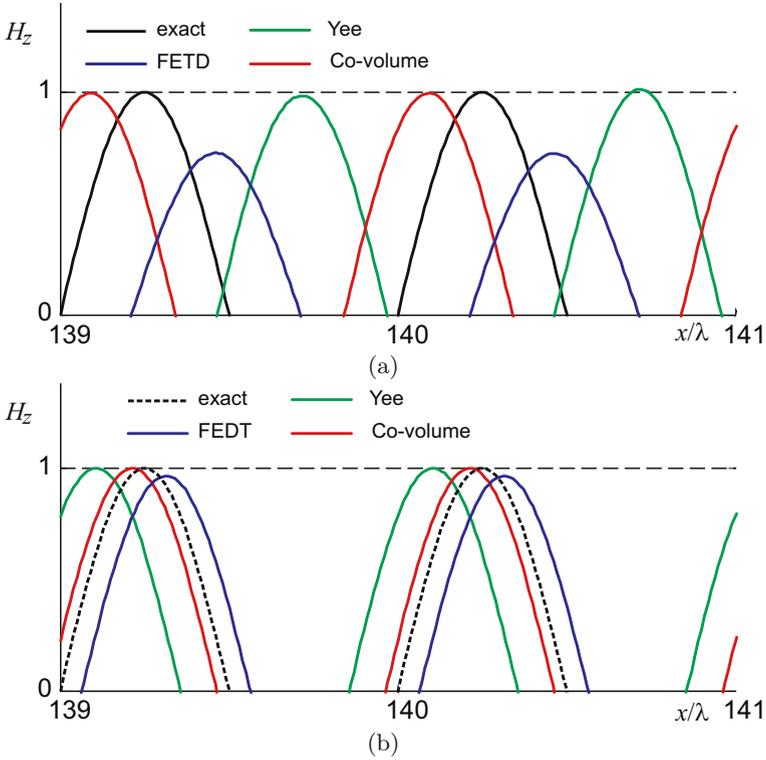
Fig. 3. Propagation of a plane harmonic TE wave in a waveguide: magnetic field after 170 cycles at a distance $x \approx 140\lambda$, using (a) $\delta \approx \lambda/15$, (b) $\delta \approx \lambda/30$.

mesh. Table 1 compares the computational performance of the algorithms, in terms of the required number of steps per cycle (*spc*), the CPU time needed (*time*), the computed phase velocity (C) and the maximum amplitude (A) of the magnetic field in the range $0 \leq x \leq 160\lambda$. This table also enables computation of the speed-up factor, between the co-volume method and FETD, which is achieved on both meshes. The effect of dispersion error on the phase velocity, as a function of time step, is shown in Figure 4. A theoretical phase velocity of one was specified for the present computation. This figure shows the computed phase velocity, for various values of the time step, on the unstructured meshes using the co-volume scheme and the FETD scheme and, on the structured meshes, using the Yee scheme, compared to the theoretically expected Yee values [TH00]. The phase velocity achieved using the co-volume method is much superior to the phase velocity expected from the structured grid implementation.

Table 1. Propagation of a plane harmonic TE wave in a waveguide.

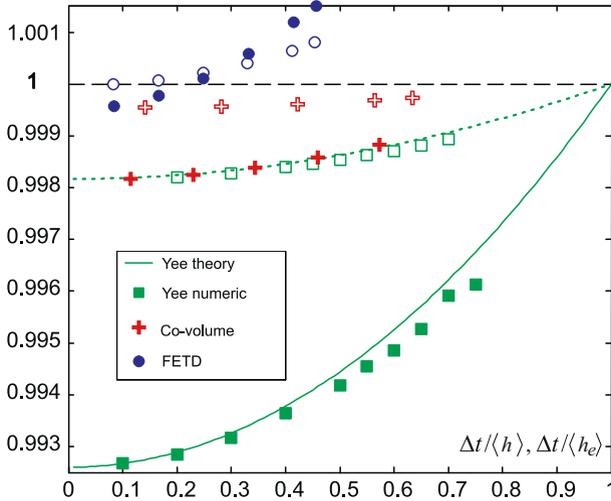|		$\delta \approx \lambda/15$		
Scheme	*spc*	*time*, s	C	A
Yee	21	7	0.99613	1.00
Co-volume	46	29	0.99850	1.00
FETD	44	3151	1.0015	0.723
		$\delta \approx \lambda/30$		
Scheme	*spc*	*time*, s	C	A
Yee	43	61	0.99896	1.00
Co-volume	106	263	0.99964	1.00
FETD	89	23040	1.0008	0.96

Fig. 4. Propagation of a plane harmonic TE wave in a waveguide showing variation of the computed phase velocity with $\Delta t/\langle h \rangle$ ($\Delta t/\langle h_e \rangle$ for FETD). Solid symbols and solid line: $\delta \approx \lambda/15$; open symbols and dotted line: $\delta \approx \lambda/30$. Here $\langle h \rangle$ is the averaged Voronoï edge length, $\langle h_e \rangle$ is the averaged minimal triangle height.

6.2 Scattering by a Circular PEC Cylinder

The second example is the simulation of scattering of a plane single frequency TE wave by a perfectly conducting circular cylinder of diameter λ. The objective is to use this example to illustrate the order of accuracy that can be achieved with the co-volume solution technique and the FETD technique on unstructured meshes. The problem is solved on a series of unstructured meshes, with mesh spacings ranging from $\lambda/8$ to $\lambda/128$. The minimum distance from the rectangular PML to the cylinder is λ. When the spacing is
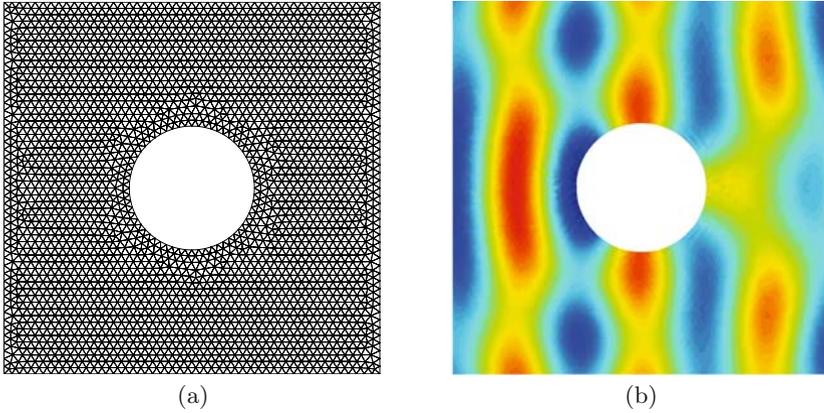
Fig. 5. Scattering of a plane TE wave by a circular PEC cylinder of diameter λ showing (a) an unstructured mesh with $\delta \approx \lambda/16$, (b) the corresponding computed total magnetic field.
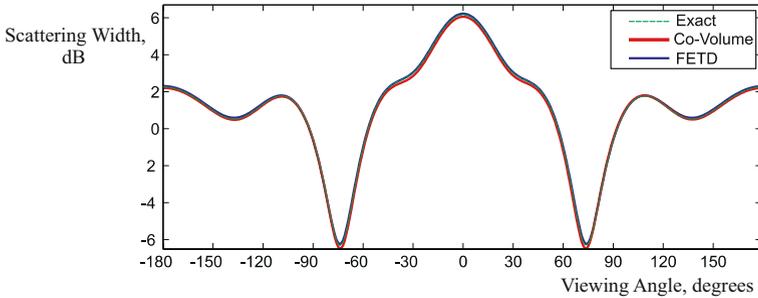

Fig. 6. Scattering of a plane TE wave by a circular PEC cylinder of diameter λ showing a comparison between the computed and analytical scattering width distributions.

$\lambda/16$, the mesh employed, excluding the PML region, and the corresponding distribution of the computed total magnetic field is shown in Figure 5. The computed scattering width distributions are compared to the exact distribution in Figure 6. For each simulation undertaken, the error, E_{SW}, in the solution is determined as the maximum difference, in absolute value, between the computed and analytical scattering width distributions. The variation of this computed error, with the number of elements per wavelength, λ/δ, for both the FETD and co-volume schemes, is shown in Figure 7. It can be observed that a convergence rate of around $\mathcal{O}(\delta^2)$ is obtained with both methods on these unstructured meshes, indicating that second-order accuracy is achieved. It is likely that the error in the FETD results is slightly less because the approach adopted for the evaluation of the scattering width integral requires an interpolation, in the co-volume scheme, to obtain all the field components at
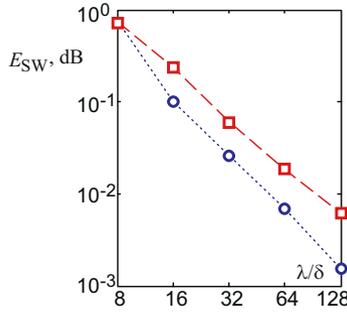
Fig. 7. Scattering of a plane TE wave by a circular PEC cylinder of diameter λ showing the variation of the computed error, with the number of elements, λ/δ, per wavelength, for the co-volume scheme and the FETD scheme on the unstructured meshes.

Table 2. Scattering of a plane TE wave by a circular PEC cylinder of diameter λ.

λ/δ	Co-volume			FETD			Speed up ratio FETD/Co-volume
	spc	*time*	E_{SW}	*spc*	*time*	E_{SW}	
8	21	0.15	0.744	31	1.2	0.750	8
16	42	0.5	0.275	61	15.	0.102	30
32	83	4.0	0.060	122	117	0.026	30
64	165	37	0.019	242	922	0.007	25
128	239	250	0.006	485	7295	0.002	30

one location. The values of *spc*, *time* and E_{SW} are shown in Table 2 for the co-volume scheme and the FETD scheme on these unstructured meshes. It can be observed that, for these simulations, the co-volume scheme is nearly 30 times faster than the FETD scheme.

As a more challenging variation of this example, we also consider scattering of a plane single frequency wave by a perfectly conducting circular cylinder of diameter 15λ. The mesh employed is generated to meet a mesh spacing requirement of $\delta = \lambda/15$. Again, the minimum distance from the PML region to the cylinder is λ. The solution is advanced for 50 cycles of the incident wave and the computed and exact scattering width distributions are compared in Figure 8(a). Excellent agreement with the exact solution is observed using both schemes. The distribution of the computed total magnetic field in the complete domain, including the PML, is shown in Figure 8(b). For this example, the co-volume scheme is nearly 34 times faster than the FETD scheme.
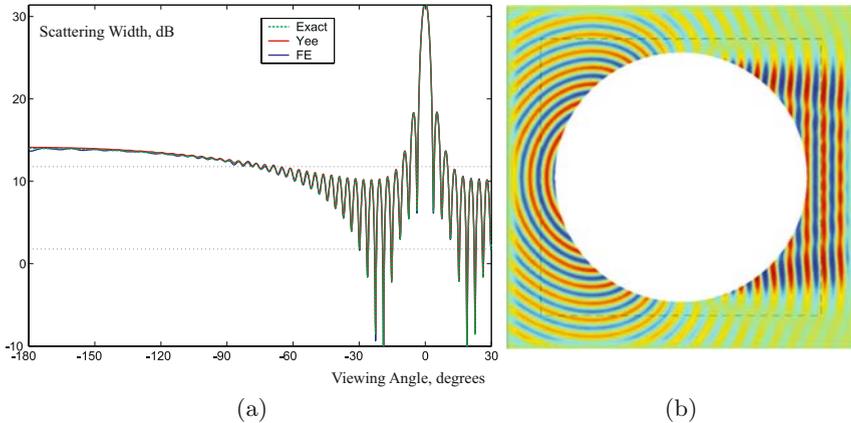
Fig. 8. Scattering of a plane TE wave by a circular PEC cylinder of diameter 15λ showing (a) a comparison between the exact and computed scattering width distributions, (b) computed contour distribution of the total magnetic field in the complete computational domain.

6.3 Scattering by a Square PEC Cylinder

The next example involves scattering of a plane single frequency electromagnetic wave by a perfectly conducting cylinder of square cross section. The sides of the square are of length λ. The objective is to use this example to illustrate the accuracy of the FETD and co-volume schemes in the presence of singularities. This simple geometry means that the computational domain may be discretised using a structured mesh of square elements and, in this case, the co-volume scheme of the equations (11) and (12) reduces to the classical Yee scheme. The distribution of the scattering width obtained using the Yee scheme on a fine Cartesian grid, with 512 elements per wavelength, is taken as the benchmark solution. An unstructured mesh, termed mesh a, is generated with mesh spacing $\lambda/16$. The solution is advanced on this mesh for 8 cycles using both the co-volume and the FETD schemes. In this case, the error E_{SW} is determined as the maximum difference, in absolute value, between the computed and the benchmark scattering width distributions. Table 3 shows the values of *spc*, *time* and E_{SW} for this grid. It is apparent that the error in the FETD scheme is an order of magnitude greater than the error in the co-volume method. This is believed to be due to the singularity in the geometry, where higher mesh resolution will be required in a scheme such as FETD. Two further unstructured meshes are generated, by reducing the spacing by a factor of 2 (termed mesh b) and 4 (termed mesh c), in the vicinity of the corners. Details of the three meshes in the region of one of the corners are shown in Figure 9. Figure 10 shows the variation in the computed error with the near corner resolution that is employed. It can be seen that the error in the FETD results decreases as the mesh is refined. It is also clear that the

Table 3. Simulation of scattering of a plane TE wave by a square PEC cylinder of side length λ.

Mesh resolution	FETD			Co-volume			Speed up ratio
	spc	*time*, s	E_{SW}	*spc*	*time*, s	E_{SW}	
a	61	18.	2.64	45	0.4	0.21	45
b	90	27.	1.66	88	0.8	0.25	34
c	182	58.	0.38	164	1.3	0.14	44

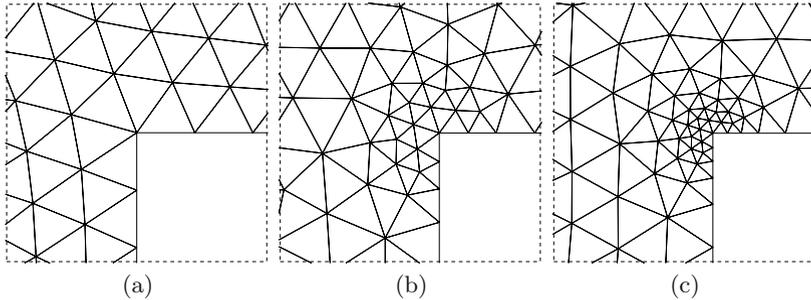(a)   (b)   (c)

Fig. 9. Details of the meshes employed for the simulation of scattering of a plane TE wave by a square PEC cylinder of side length λ showing (a) mesh a, (b) mesh b, (c) mesh c.
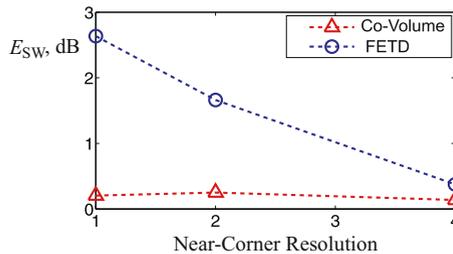
Fig. 10. Simulation of scattering of a plane TE wave by a square PEC cylinder of side length λ showing the variation in the computed error with the near corner mesh resolution.

error in the FETD results on mesh c is similar to the error in the co-volume results obtained on mesh a. The constant error in the co-volume results confirm the belief that no special modification of the scheme is required in the vicinity of geometrical singularities. Table 3 also displays information about the calculations performed on meshes b and c. For this example, the co-volume scheme is faster than FETD by a factor that ranges between 34 and 45. This level of variation in the speed-up factor is probably due to the difficulty in determining exactly the small times required for the co-volume solution.

6.4 Scattering by a PEC NACA0012 Aerofoil

The next example involves the simulation of scattering of a plane single frequency wave, directed along the x-axis, by a perfectly conducting NACA0012 aerofoil of length λ. The aim of this example is to analyse the performance of the numerical schemes when the geometry exhibits high curvature. A benchmark solution is computed using an unstructured mesh with spacing $\lambda/120$. The unstructured mesh is generated, outside the aerofoil, in the region $-\lambda \leq x, y \leq \lambda$. The scattering width distributions computed on this mesh with the co-volume scheme and the FETD scheme proved to be identical. An unstructured mesh was generated to meet the spacing requirement of $\lambda/15$. Another unstructured mesh, providing better representation of the leading edge curvature, is generated by locally reducing the mesh spacing in the vicinity of the leading edge of the airfoil by a factor of 2. A view of both these meshes is shown in Figure 11.

The computed scattering width distributions are compared with the benchmark distribution in Figure 12. It can be observed that the co-volume results are better on the uniform mesh and that the accuracy of the FETD results improve with the local refinement in the leading edge region. For this example, Table 4 shows the values of *spc*, *time* and E_{SW}. The co-volume method is approximately 30 times faster than FETD for this example.

6.5 Scattering by a PEC Cavity

The final example considers the simulation of scattering of a plane single frequency wave by a U-shaped PEC cavity. The thickness of the cavity walls is equal to 0.4λ, the internal cavity width is 2λ and the internal cavity length is 8λ. In the simulation, the wave is incident upon the open end of the cavity and propagates in a direction which lies at an angle $\theta = 30°$ to the main axis of the cavity. An unstructured mesh is employed, with typical spacing $\lambda/15$, in the region that lies within a distance of λ from the scatterer, as

(a)   (b)

Fig. 11. Details of the unstructured meshes employed for the simulation of scattering of a plane TE wave by a PEC NACA0012 aerofoil of length λ showing (a) the uniform mesh, (b) the locally refined mesh.
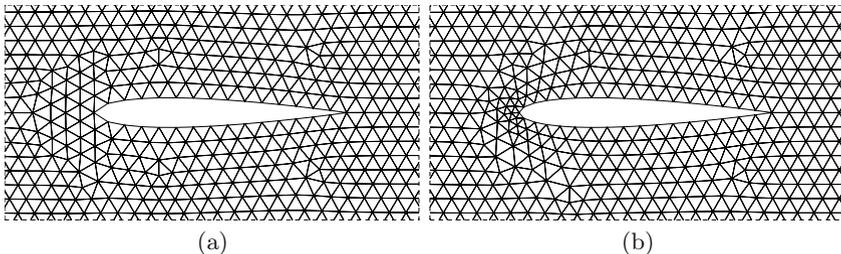
Scattering Width, dB
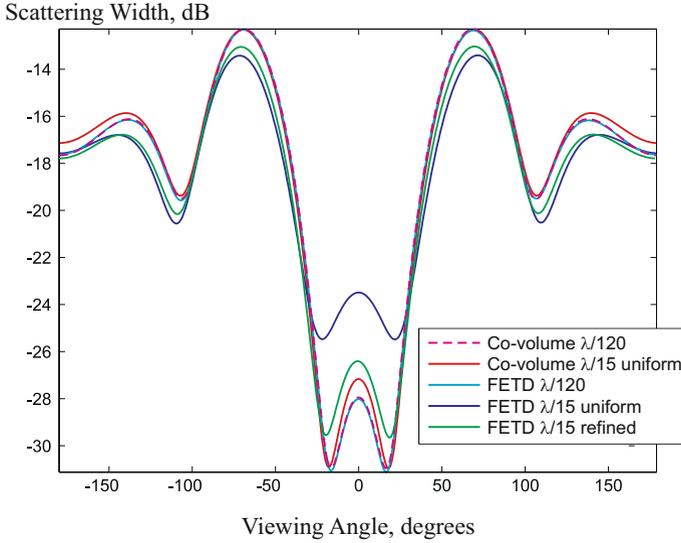

Viewing Angle, degrees

Fig. 12. Simulation of scattering of a plane TE wave by a PEC NACA0012 aerofoil of length λ showing a comparison between the computed and benchmark scattering width distributions.

Table 4. Simulation of scattering of a plane TE wave by a PEC NACA0012 aerofoil of length λ.

Mesh resolution	FETD			Co-volume			Speed up ratio
	spc	*time*, s	E_{SW}	*spc*	*time*, s	E_{SW}	
Uniform	59	12.	6.00	46	0.4	0.9	30
Refined	97	20.	2.14	99	0.6	0.5	33

shown in Fig. 13(a). The simulations are advanced for 150 cycles and the typical distribution of the contours of the computed total magnetic field in the domain, excluding the PML, is shown in Figure 13(b). A comparison of the computed scattering width distributions is given in Figure 14. Also shown on this figure is the scattering width distribution computed using a high order finite element frequency domain (FEFD) simulation [LMHW02]. The number of steps per cycle is 57 for the co-volume scheme and 59 for the FETD method and, for this example, the co-volume scheme requires 31 seconds of cpu time, while the FETD method requires 1980 seconds. This represents a speed-up of a factor of 65.

<center>(a)   (b)</center>
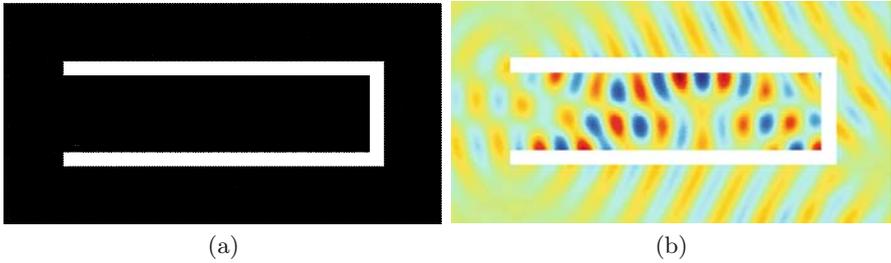
Fig. 13. Simulation of scattering of a plane TE wave by a PEC cavity showing (a) the unstructured mesh employed, (b) the computed total magnetic field after 150 cycles.
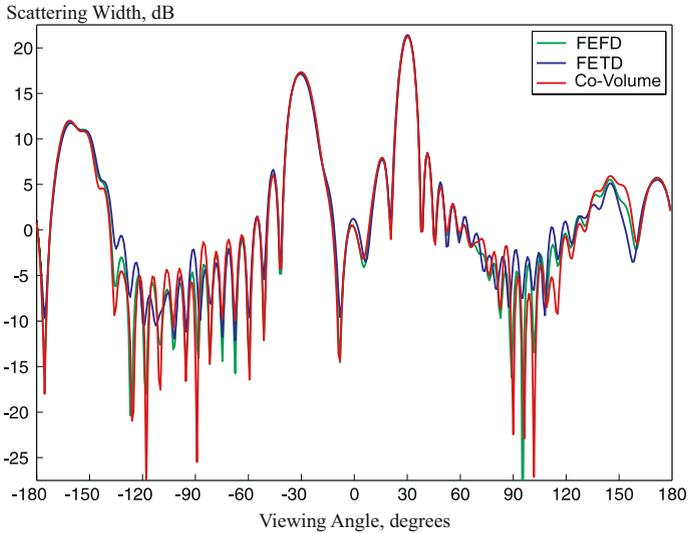
Fig. 14. Simulation of scattering of a plane TE wave by a PEC cavity showing a comparison of the scattering width distributions computed, after 150 cycles, by FETD, the co-volume scheme and a FEFD method.

7 Conclusions

The numerical performance of an explicit unstructured mesh co-volume time domain scheme and a standard finite element time domain method has been compared for a number of electromagnetic wave propagation and scattering examples. To ensure the efficiency of the co-volume approach, the smooth Delaunay–Voronoï dual meshes that are used are generated using a stitching method. The numerical examples that have been considered show that the co-volume method is 30–60 times faster than the finite element method for two-dimensional scattering problems. In addition, the co-volume method

proved to be less sensitive to special geometric features, such as singularities and regions of high curvature. It is anticipated that, for three-dimensional problems, a speed-up factor of three orders of magnitude could be achieved, if the mesh generation method can be extended to provide high quality tetrahedral elements.

References

[Ber94]   J.-P. Berenger. A perfectly matched layer for absorption of electromagnetic waves. *J. Comput. Phys.*, 114:185–200, 1994.

[BP97]   F. Bonnet and F. Poupaud. Berenger absorbing boundary condition with time finite-volume scheme for triangular meshes. *Appl. Numer. Math.*, 25:333–354, 1997.

[CFS93]   J. P. Cioni, L. Fezoui, and H. Steve. A parallel time-domain Maxwell solver using upwind schemes and triangular meshes. *Impact Comput. Sci. Engrg.*, 5:215–247, 1993.

[DBB99]   A. Deraemaeker, I. Babuška, and P. Bouillard. Dispersion and pollution of the FEM solution for the Helmholtz equation in one, two and three dimensions. *Internat. J. Numer. Methods Engrg.*, 46:471–499, 1999.

[DH03]   J. Donéa and A. Huerta. *Finite element methods for flow problems.* John Wiley & Sons, 2003.

[DL97]   E. Darve and R. Löhner. Advanced structured-unstructured solver for electromagnetic scattering from multimaterial objects. AIAA Paper 97–0863, Washington, 1997.

[EHM$^+$03]   M. El hachemi, O. Hassan, K. Morgan, D. P. Rowse, and N. P. Weatherill. Hybrid methods for electromagnetic scattering simulations on overlapping grids. *Comm. Numer. Methods Engrg.*, 19:749–760, 2003.

[EL02]   F. Edelvik and G. Ledfelt. A comparison of time-domain hybrid solvers for complex scattering problems. *Internat. J. Numer. Model.: Elect. Net. Dev. Fields*, 15:475–487, 2002.

[Geo91]   P. L. George. *Automatic mesh generation. Applications to finite element methods.* John Wiley & Sons, 1991.

[GL93]   S. Gedney and F. Lansing. Full wave analysis of printed microstrip devices using a generalized Yee algorithm. In *Proceedings of the IEEE Antenas and Propagation Society International Symposium*, pages 1179–1182, Ann Arbor, 1993. Pennsylvania State University.

[LMHW02]   P. D. Ledger, K. Morgan, O. Hassan, and N. P. Weatherill. Arbitrary order edge elements for electromagnetic scattering simulations using hybrid meshes and a PML. *Internat. J. Numer. Methods Engrg.*, 55:339–358, 2002.

[Mad95]   N. Madsen. Divergence preserving discrete surface integral methods for Maxwell's equations using nonorthogonal unstructured grids. *J. Comput. Phys.*, 119:35–45, 1995.

[MHP94]   K. Morgan, O. Hassan, and J. Peraire. An unstructured grid algorithm for the solution of Maxwell's equations in the time domain. *Internat. J. Numer. Methods Fluids*, 19:849–863, 1994.

[MHP96]   K. Morgan, O. Hassan, and J. Peraire. A time domain unstruc-
            tured grid approach to the simulation of electromagnetic scattering in
            piecewise homogeneous media. *Comput. Methods Appl. Mech. Engrg.*,
            134:17–36, 1996.

[MHPW00]   K. Morgan, O. Hassan, N. E. Pegg, and N. P. Weatherill. The sim-
            ulation of electromagnetic scattering in piecewise homogeneous media
            using unstructured grids. *Comput. Mech.*, 25:438–447, 2000.

[MM98]   A. Monorchio and R. A. Mittra. A hybrid finite-element/finite-
            difference (FE/FDTD) technique for solving complex electromagnetic
            problems. *IEEE Microwave Guided Wave Lett.*, 8:93–95, 1998.

[MSH91]   A. H. Mohammadian, V. Shankar, and W. F. Hall. Computation of
            electromagnetic scattering and radiation using a time-domain finite-
            volume discretization procedure. *Comput. Phys. Comm.*, 68:175–196,
            1991.

[MWH+99]   K. Morgan, N. P. Weatherill, O. Hassan, P. J. Brookes, R. Said, and
            J. Jones. A parallel framework for multidisciplinary aerospace engi-
            neering simulations using unstructured meshes. *Internat. J. Numer.
            Methods Fluids*, 31:159–173, 1999.

[NW98]   R. A. Nicolaides and Q.-Q. Wang. Convergence analysis of a co-volume
            scheme for Maxwell's equations in three dimensions. *Math. Comp.*,
            67:947–963, 1998.

[PLD92]   B. Petitjean, R. Löhner, and C. R. Devore. Finite element solvers for
            radar cross section RCS calculations. AIAA Paper 92–0455, Washing-
            ton, 1992.

[PPM99]   J. Peraire, J. Peiró, and K. Morgan. Advancing front grid generation.
            In J. F. Thompson, B. K. Soni, and N. P. Weatherill, editors, *Handbook
            of Grid Generation*, pages 17.1–17.22. CRC Press, 1999.

[RB00]   T. Rylander and A. Bondeson. Stable FEM–FDTD hybrid method for
            Maxwell's equations. *Comput. Phys. Comm.*, 125:75–82, 2000.

[RBT97]   W. Ruey-Beei and I. Tatsuo. Hybrid finite-difference time-domain mod-
            eling of curved surfaces using tetrahedral edge elements. *IEEE Trans.
            Antennas and Propagation*, 45:1302–1309, 1997.

[SWH+06]   I. Sazonov, D. Wang, O. Hassan, K. Morgan, and N. P. Weatherill. A
            stitching method for the generation of unstructured meshes for use with
            co-volume solution techniques. *Comput. Methods Appl. Mech. Engrg.*,
            195:1826–1845, 2006.

[TH00]   A. Taflove and S. C. Hagness. *Computational electrodynamics: The
            finite-difference time domain method*. Artech House, Boston, 2nd edi-
            tion, 2000.

[WH94]   N. P. Weatherill and O. Hassan. Efficient three-dimensional Delaunay
            triangulation with automatic point creation and imposed boundary
            constraints. *Internat. J. Numer. Methods Engrg.*, 37:2005–2040, 1994.

[Yee66]   K. S. Yee. Numerical solution of initial boundary value problem in-
            volving Maxwell's equation in isotropic media. *IEEE Trans. Antennas
            and Propagation*, 14:302–307, 1966.

[ZM06]   O. C. Zienkiewicz and K. Morgan. *Finite elements and approximation*.
            Dover, 2006.

The von Neumann Triple Point Paradox

Richard Sanders[1]* and Allen M. Tesdall[2]†

[1] Department of Mathematics, University of Houston, Houston, TX 77204, USA
   `sanders@math.uh.edu`
[2] Fields Institute, Toronto, ON M5T 3J1, Canada and Department of
   Mathematics, University of Houston, Houston, TX 77204, USA
   `atesdall@fields.utoronto.ca`

Summary. We describe the problem of weak shock reflection off a wedge and discuss the triple point paradox that arises. When the shock is sufficiently weak and the wedge is thin, Mach reflection appears to be observed but is impossible according to what von Neumann originally showed in 1943. We summarize some recent numerical results for weak shock reflection problems for the unsteady transonic small disturbance equations, the nonlinear wave system, and the Euler equations. Rather than finding a standard but mathematically inadmissible Mach reflection with a shock triple point, the solutions contain a complex structure: there is a sequence of triple points and supersonic patches in a tiny region behind the leading triple point, with an expansion fan originating at each triple point. The sequence of patches may be infinite, and we refer to this structure as Guderley Mach reflection. The presence of the expansion fans at the triple points resolves the paradox. We describe some recent experimental evidence which is consistent with these numerical findings.

Key words: self-similar solutions, two-dimensional Riemann problems, triple point paradox

1 Introduction

Consider a planar normal shock in an inviscid compressible and calorically perfect gas which impinges on a fixed wedge with apex half angle θ_w, see Figure 1. Given an upstream state with density $\rho = \rho_r$, velocity $u = v = 0$ and pressure $p = p_r$, one calculates that downstream of a *fast* (i.e., $u + c$) shock
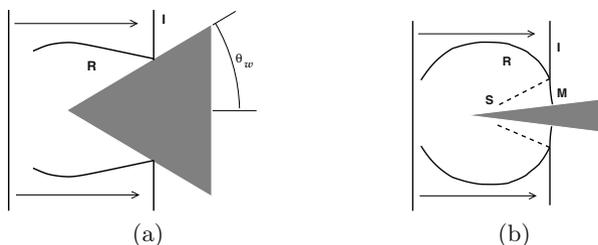
Fig. 1. A planar shock moving from left to right impinges on a wedge. After contact, I indicates the incident shock and R indicates the reflected shock. On the right, the dotted line S indicates a slip line and M is the Mach stem. Regular reflection is depicted on the left. Irregular reflection is depicted on the right.
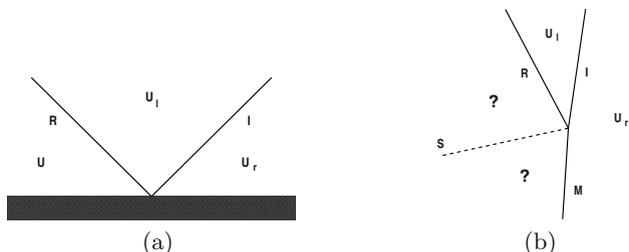

Fig. 2. A blow-up of the incident and reflected shock intersection. Regular reflection is on the left and irregular on the right. The constant states upstream and downstream of the incident shock are denoted by U_r and U_l. Whether or not constant states indicated by the question marks exist depends on the strength of I.

$$\frac{p_l}{p_r} = \frac{2\gamma}{\gamma+1}M^2 - \frac{\gamma-1}{\gamma+1}, \quad \frac{u_l}{c_r} = \frac{2}{\gamma+1}\left(M - \frac{1}{M}\right),$$
$$\frac{\rho_l}{\rho_r} = \frac{(\gamma+1)M^2}{2+(\gamma-1)M^2},$$

(1)

where γ denotes the ratio of specific heats, and $M > 1$ denotes the shock Mach number defined as the Rankine–Hugoniot shock speed divided by the upstream speed of sound $c_r = \sqrt{\gamma p_r/\rho_r}$. Following interaction, a number of self-similar (with respect to the wedge apex) reflection patterns are possible, depending on the values of M and θ_w.

This wedge reflection problem has a rich history, experimentally, analytically, and numerically. Probably the earliest and most significant analytical result was found by von Neumann [Neu43]. In this work were first formulated the equations which describe two and three planar shocks meeting at a point separated by constant states, see Figure 2. The two shock theory leads to what is known as *regular reflection*. The three shock theory leads to *Mach reflection*. For supersonic regular reflection, state U immediately behind the reflected shock R is supersonic and becomes subsonic across a sonic line downstream (toward the wedge's apex). When the incident shock angle is increased

to $\pi/2 - \theta^*(M)$ with respect to the wall, where $\theta_w = \theta^*(M)$ is the critical wedge half angle, state U becomes sonic. Therefore, at $\theta_w = \theta^*(M)$, acoustic signals generated downstream (e.g., from the wedge apex) will overtake the R-I reflection point, conceivably causing transition from regular reflection, depicted in the left figure, to irregular reflection, depicted in the right figure. This is one of several criteria which have been suggested to explain transition from regular to irregular reflection; see Henderson [Hen87] for a thorough and detailed discussion.

Loosely speaking, a *weak* incident shock has M slightly larger than 1, whereas a *strong* incident shock has M substantially larger than one. Theoretical analysis indicates that transition to Mach reflection is impossible when the incident shock is sufficiently weak. In fact, triple point solutions, as depicted in Figure 2(b), do not exist for sufficiently weak shocks. However, experiments in which weak shock waves are reflected off a wedge with $\theta_w \ll \theta^*(M)$ appear to show a standard Mach reflection pattern. This apparent disagreement between theory and experiment was discussed by von Neumann and has since become known as the von Neumann triple point paradox [Neu63, Hen87, SA05].

Guderley [Gud47, Gud62] as far back as 1947 proposed that there is an expansion fan and a supersonic region directly behind the triple point in a steady weak shock Mach reflection. He demonstrated that one could construct local solutions consisting of three plane shocks, an expansion fan, and a contact discontinuity or slip line meeting at a point. However, despite intensive experimental [BT49, STS92, Ste59] and numerical [CH90, BH92, TR94] studies, no evidence of an expansion fan or supersonic patch was observed. The first evidence supporting Guderley's proposed resolution was contained in numerical solutions of shock reflection problems for the unsteady transonic small disturbance equations in [HB00] and the compressible Euler equations in [VK99]. There were presented solutions that contain a tiny supersonic region embedded in the subsonic flow directly behind the triple point in a weak shock Mach reflection. Subsequently, Zakharian et al. [ZBHW00] found a supersonic region in a numerical solution of a shock reflection problem for the Euler equations, for a set of parameter values corresponding to those used in the unsteady transonic small disturbance solution in [HB00]. The supersonic region in the solutions in [VK99, HB00, ZBHW00] is extremely small, which explains why it had not been observed earlier.

This paper is organized as follows. In Section 2 the unsteady transonic small disturbance asymptotic model for a weak shock impinging on a thin wedge is recalled. Numerical evidence is offered to suggest an interesting resolution of the von Neumann paradox. Experimental evidence to support what was found numerically is displayed at the end of this section. In Section 3 a simple 3×3 hyperbolic system is given which exhibits irregular reflection but does not admit Mach reflection. It is solved numerically, displaying very similar structure to what was found in Section 2. Finally, the full compressible Euler equations are solved in Section 4 for a very weak incident shock

impinging on a thin wedge. The numerical solution appears to be in agreement with what is found for the model problems from the previous sections.

2 The Weak Shock Thin Wedge Limit

The compressible Euler equations are given by

$$\frac{\partial \rho}{\partial t} + \nabla \cdot \rho \mathbf{u} = 0,$$

$$\frac{\partial \rho \mathbf{u}}{\partial t} + \nabla \cdot \rho \mathbf{u} \otimes \mathbf{u} + \nabla p = 0, \tag{2}$$

$$\frac{\partial \rho e}{\partial t} + \nabla \cdot (\rho e + p)\mathbf{u} = 0,$$

where ρ is the fluid density, $\mathbf{u} = (u, v)$ is the x-y velocity vector, p is the pressure and e is the total energy per unit mass. The internal energy per unit mass $\varepsilon = e - 1/2|\mathbf{u}|^2$, and we take $p = (\gamma - 1)\rho\varepsilon$ for a calorically perfect gas with the constant ratio of specific heats $\gamma > 1$.

Consider an incident planar shock with Mach number $M = 1 + \varepsilon^2$ striking a thin wedge with half angle $\theta_w = a\varepsilon$, where $\varepsilon > 0$ is destined to vanish. Take the undisturbed upstream state U_r as $\rho = \rho_r$, $u = v = 0$ and $p = p_r$, yielding an upstream speed of sound $c_r = \sqrt{\gamma p_r / \rho_r}$. From (1), calculate that U_l is

$$\frac{p_l}{p_r} = \left(1 + \frac{4\gamma}{\gamma + 1}\varepsilon^2\right) + O(\varepsilon^4), \qquad \frac{u_l}{c_r} = \frac{4}{\gamma + 1}\varepsilon^2 + O(\varepsilon^4),$$

$$\frac{\rho_l}{\rho_r} = \left(1 + \frac{4}{\gamma + 1}\varepsilon^2\right) + O(\varepsilon^4), \qquad \frac{v_l}{c_r} = \frac{-4}{\gamma + 1}a\varepsilon^3 + O(\varepsilon^5). \tag{3}$$

Hunter and Brio [HB00] observed the scales shown in (3) and proposed an asymptotic model based on

$$p = p_r(1 + \varepsilon^2 \hat{p}), \qquad u = c_r \varepsilon^2 \hat{u},$$
$$\rho = \rho_r(1 + \varepsilon^2 \hat{\rho}), \qquad v = c_r \varepsilon^3 \hat{v},$$

and the stretched independent variables

$$\hat{x} = \frac{x - p(t)}{\varepsilon^2}, \qquad \hat{y} = \frac{y}{\varepsilon},$$

where $p(t)$ is the location where the incident shock would (neglecting possible interactions) strike the wedge wall at time t,

$$p(t) = c_r \cos(\theta_w)(1 + \varepsilon^2)\, t = c_r \cos(a\varepsilon)(1 + \varepsilon^2)\, t \approx c_r(1 - (1 - a^2/2)\varepsilon^2)\, t,$$
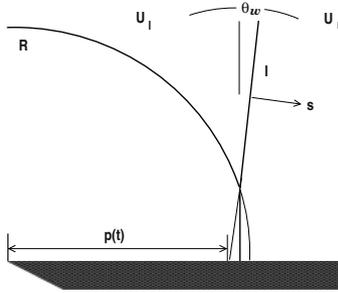
Fig. 3. A weak shock over a thin wedge. U_r and U_l are the states to the right and left of the incident shock I. $\theta_w = a\varepsilon \ll 1$ and the incident shock has Mach number $M = 1 + \varepsilon^2$. $x = p(t)$ is the location where I would intersect the wall at time t, neglecting interaction.

see Figure 3. Inserting these into (2), equating like powers of ε, and making an additional order one change of variable (denoted by $\check{u}$, etc.), they find that $\check{u}$ and $\check{v}$ asymptotically satisfy

$$\check{u}_t + \left(1/2\,\check{u}^2\right)_{\check{x}} + \check{v}_{\check{y}} = 0,$$
$$\check{u}_{\check{y}} - \check{v}_{\check{x}} = 0. \tag{4}$$

This is, of course, the celebrated *unsteady transonic small disturbance equation* (UTSDE). The UTSDE is solved on the upper half plane with a *no-flow* boundary condition $\check{v}(\check{x}, 0, t) = 0$ along $\check{y} = 0$ and initial data

$$(\check{u}(\check{x}, \check{y}, 0), \check{v}(\check{x}, \check{y}, 0)) = \begin{cases} (0, 0) & \text{if } \check{x} > \check{a}\check{y} \\ (1, -\check{a}) & \text{if } \check{x} < \check{a}\check{y}, \end{cases}$$

where

$$\check{a} = \frac{a}{2} = \frac{1}{2}\frac{a\varepsilon}{\sqrt{1 + \varepsilon^2 - 1}} \sim \frac{1}{2}\frac{\theta_w}{\sqrt{M - 1}}.$$

The jump at $\check{x} = \check{a}\check{y}$ corresponds to the incident shock I. The data is vorticity-free but incompatible with the no-flow boundary condition behind. As time advances, the reflected wave pattern R will emerge from the trailing boundary.

For $\check{a}$ in the range $0 < \check{a} < \sqrt{2}$, regular reflection for this initial-boundary value problem is impossible [BH92]. Moreover, it is shown in [BH92] as well as in [TR94] that (4) can never admit triple point solutions. Therefore, this asymptotic model equation is very well designed to investigate the von Neumann triple point paradox.

A numerical solution to (4) was obtained in [HB00] for the value $\check{a} = 0.5$ (a value for which regular reflection does not occur). An irregular reflection pattern globally resembling single Mach reflection was observed. When the region containing the apparent triple point was greatly refined, however, a small supersonic patch located in the subsonic zone directly below the reflected shock and behind the Mach stem was detected, see [HB00, page 242]. This,

along with the contemporaneous work in [VK99], was the first indication that Guderley's resolution of the triple point paradox might be essentially correct. Using a new numerical scheme, a subsequent study by Tesdall and Hunter [TH02], we further investigated the structure of irregular reflection found in the UTSDE asymptotic model.

The supersonic patch detected in [VK99, HB00] appeared to confirm Guderley's four-wave solution. The patch indicates that it is plausible for an expansion wave to be a (unobserved) part of the observed three shock confluence. We briefly summarize the numerical techniques employed by Tesdall and Hunter. First, they used a parabolic grid aligned with the weak reflected shock. They then solved the UTSDE in self-similar variables $\check{x} \to \check{x}/t, \check{y} \to \check{y}/t$. The advantage of using self-similar coordinates is that the problem remains fixed on the computational grid, and a steady self-similar solution is obtained by letting a pseudo-time $t \to \infty$. Following the classical Cole–Murman approach, $(\check{u}, \check{v})$ is written as $\mathrm{grad}\,\phi$. The nonlinearities in the resulting scalar equation are discretized by a min-mod limited Engquist–Osher numerical flux. A steady state solution is obtained by lagged implicit time marching and grid continuation.

We present results obtained by the method of Tesdall and Hunter in Figure 4. The full simulation is carried out on a spatial grid that fits in $[-3, 2] \times [0, 2.5]$, with the inverse slope parameter $\check{a} = 0.5$. The total number of grid points employed is approximately 2.7×10^6, where, by local grid refinement, the region depicted in Figure 4(a) spans $768 \times 608 \approx 4.7 \times 10^5$ points. This yielded a grid size near the triple point of approximately 1.5×10^{-5}.

Clear evidence of an expansion fan is seen at the triple point depicted in Figure 4. What is equally remarkable is what appears to be a sequence of progressively smaller and weaker shock/expansion pairs running a short distance (less than 2%) down the length of the Mach stem. The expansion from wave i appears to terminate through its interaction with the shock from wave $i + 1$. The supersonic region behind the leading triple point is extremely small, which explains why it had not been observed earlier. The results in [TH02] suggest that the sequence of triple points and expansion waves/shocks in a weak shock irregular reflection may be infinite. Whether this sequence is infinite or not is certainly impossible for any numerical simulation to determine. In fact, one could argue that the structure indicated in Figure 4 may be numerical flux dependent (upwind/non-upwind) or that the asymptotic model may predict something that is not physically realized. We address these concerns here and in the following sections.

Experimental confirmation poses a most challenging problem simply because the computed *Guderley Mach reflection* structure is so small and weak. Nevertheless, some experimental evidence has recently been obtained. Following the announcement of the Guderley Mach reflection solution found in [TH02], Skews and Ashworth [SA05] modified an existing shock tube experimental apparatus in order to obtain Mach stem lengths more than an order of magnitude larger than those possible from conventional shock tubes. All

(a)

(b)

(c)
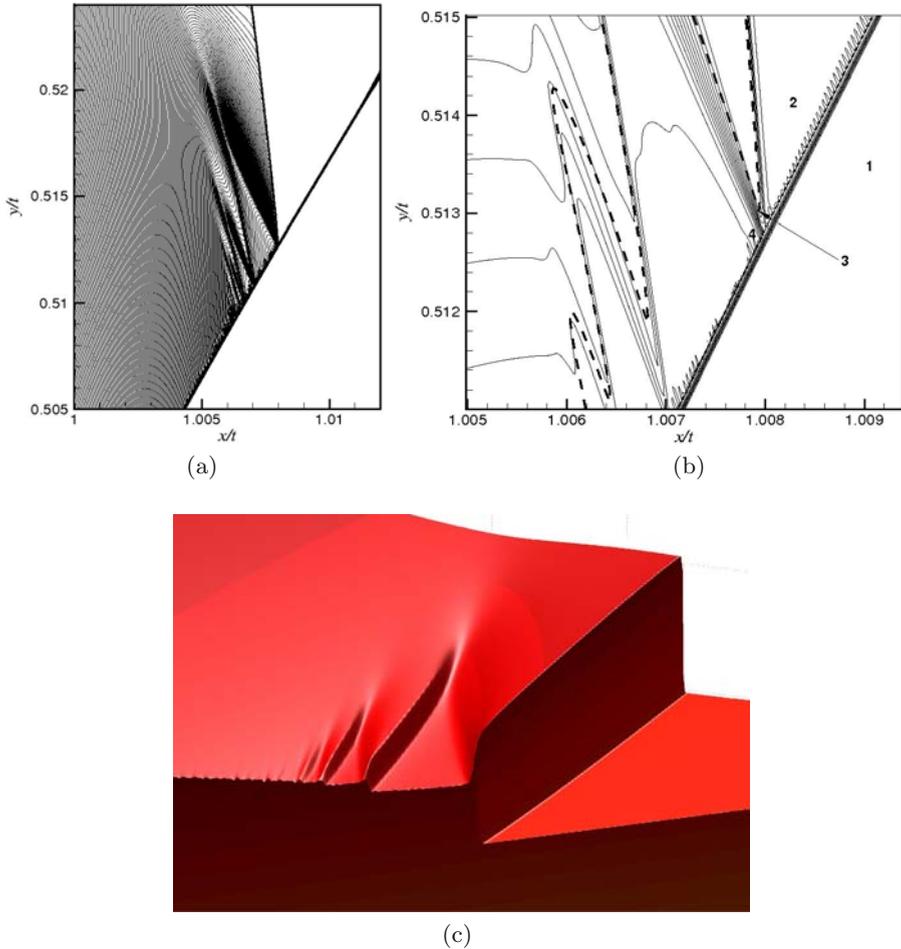
Fig. 4. Closeups of an apparent triple point for the UTSDE using the approach of Tesdall and Hunter. In (a) and (b) the incident shock leaves to the upper right, the reflected shock towards the top, and the Mach stem exits at the bottom. The plot in (a) depicts contour lines of u and shows a sequence of expansions/shocks running down the Mach stem. The plot in (b) shows a detail of v; 1 denotes state $v = 0$, 2 state $v = -\breve{a}$ and 3 points to the expansion wave emanating from what appears macroscopically to be a triple point. The dotted line in (b) delineates the supersonic patches within the subsonic zone behind the Mach stem. The *Guderley Mach reflection* structure can be seen better in the surface plot (c) where the viewer is upstream looking back at the triple point.

experiments were carried out on a 15° ramp with incident shock Mach numbers ranging from 1.05 to 1.1. They present images that "clearly show the existence of an expansion wave immediately behind the reflected wave as proposed by Guderley", and they found "a distinct sharp contrasting line immediately

(a)   (b)
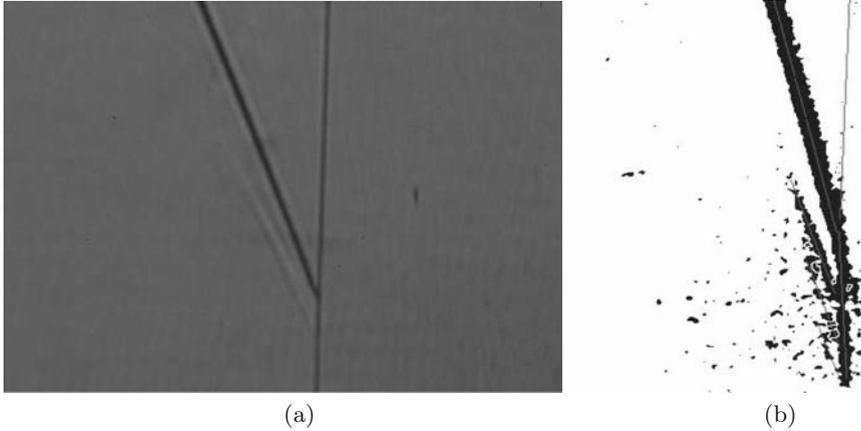
Fig. 5. On the left, a schlieren image of an experimental weak shock reflection. The incident shock (vertical) exits at the top and is moving from left to right. The reflected wave exits to the upper left, and an expansion wave is visible immediately behind it. A highly contrasted image is on the right, showing evidence of a second shocklet behind the first.

after the expansion wave, indicating the existence of a terminating shock". In addition, they obtained evidence in some of their images of a second terminating shocklet behind the first, as predicted by the simulations in [TH02]. Professor Beric Skews graciously supplied us with the images which we give here in Figure 5. Further experimental refinements and data acquisition are currently underway.

3 The Nonlinear Wave System

Here we consider a problem for the nonlinear wave system which is analogous to the reflection of weak shocks discussed in the previous section. The shock reflection problem consists of the nonlinear wave system

$$\frac{\partial \rho}{\partial t} + \nabla \cdot \rho \mathbf{u} = 0,$$

$$\frac{\partial \rho \mathbf{u}}{\partial t} + \operatorname{grad} p = 0,$$

in the half space $x > 0$ with piecewise constant Riemann data consisting of two states separated by a discontinuity located at $x = \kappa y$. Again, ρ should be thought of as density, $\mathbf{u} = (u, v)$ as velocity having x- and y-components, and $p = p(\rho)$ as pressure. For convenience, we assume $p(\rho) = C\rho^{\gamma}$ where C is a constant and $\gamma = 2$. See [TSK06].

The nonlinear wave system is a simplification of the isentropic Euler equations obtained by dropping the momentum transport terms from the

momentum equations. Compared to the UTSDE, the nonlinear wave system is closer in structure to the Euler equations: it is linearly well-posed in space and time, it has a characteristic structure similar to the Euler equations with nonlinear acoustic waves coupled (weakly) to linearly degenerate waves, and it respects the spatial Euclidean symmetries of gas dynamics (excluding space-time Galilean symmetry, of course). In fact (see [KF94]), it may be the simplest system one can construct with these symmetries. It has also served as a prototypical model for the theoretical study of shock wave reflection [ČK98, ČKK05, ČKK01]. However, the greatest attribute of (3) for our purposes is the sheer simplicity of its wave structure. Moreover, the fluxes are quadratic (when $\gamma = 2$), and so its flux Jacobians are linear in conserved variables. The Jacobian's eigenvalues are 0 and $\pm c$, where $c = \sqrt{p_\rho}$, and it has extremely simple eigenvectors. It is very well suited for efficient finite differencing.

Let $U = (\rho, m, n)$ denote the vector of conserved variables, where $m = \rho u$ and $n = \rho v$, and consider the following two-dimensional Riemann data:

$$U(x,y,0) = \begin{cases} U_1 \equiv (\rho_1, 0, 0) & \text{if } x < \kappa y, \\ U_0 \equiv (\rho_0, 0, n_0) & \text{if } x > \kappa y. \end{cases} \tag{5}$$

We choose $\rho_0 > \rho_1$ to obtain an upward moving shock in the far field, and determine n_0 so that the one-dimensional wave between U_0 and U_1 at inverse slope κ consists of a shock and a contact discontinuity with a constant middle state between them. The following expression for n_0 is readily determined:

$$n_0 = \frac{1}{\kappa} \sqrt{(1 + \kappa^2)(p(\rho_0) - p(\rho_1))(\rho_0 - \rho_1)}. \tag{6}$$

There is no physical wall in the Mach reflection simulation below. Rather, reflection occurs because the vertical axis is a line of left-right symmetry, see Figure 6(a). Here, for κ sufficiently large ($\kappa = 1$ will do), regular reflection is impossible. Moreover, as with the UTSDE, (3) can never admit triple point solutions, see [TSK06]. So we now investigate the structure of irregular reflection, this time, however, for a hyperbolic system – one which resembles the Euler equations but is not obtained from them via a limit.

The essential feature of the numerical method employed is the capability to locally refine the grid in the area of the apparent triple point. We again use self-similar variables

$$x \to x/t \equiv \xi, \quad y \to y/t \equiv \eta$$

to cast the problem into one which remains fixed on the grid. Non-uniform, logically rectangular, finite volume grids are constructed so that for a given κ the incident shock is aligned with the grid in the far field. Specifically, each problem with a given incident shock angle has a set of associated finite volume C-grids, each grid in the set corresponds to a level of grid refinement, and we use these to *grid continue* to a steady state.
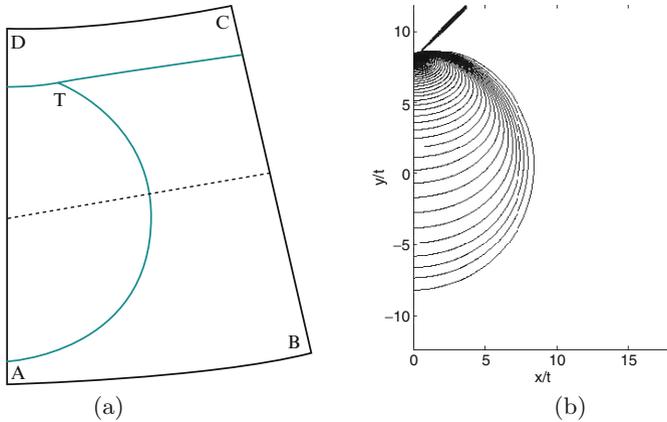
Fig. 6. A schematic diagram of the computational domain is on the left. AD is the line of symmetry. On the right is a computed self-similar solution with $\kappa = 1$.

The basic finite volume schemes used are quite standard. Each grid cell, Ω, is a quadrilateral and, using $\boldsymbol{\nu} = (\nu_\xi, \nu_\eta)$ to denote the normal vector to a typical side of Ω, numerical fluxes are designed to be consistent with

$$\widetilde{F}(U) = (F(U) - \xi U)\,\nu_\xi + (G(U) - \eta U)\,\nu_\eta = \begin{pmatrix} \nu_\xi m + \nu_\eta n - \bar{\xi}\,\rho \\ \nu_\xi p - \bar{\xi}\,m \\ \nu_\eta p - \bar{\xi}\,n \end{pmatrix},$$

where $\bar{\xi} = (\boldsymbol{\xi} \cdot \boldsymbol{\nu})$ and $\boldsymbol{\xi} = (\xi, \eta)$. Since $\boldsymbol{\xi}$ varies in space, numerical flux formulae are evaluated at $\boldsymbol{\xi}$ frozen at the midpoint of each cell side. Two distinctly different numerical fluxes are utilized in the results presented below:

1. Lax–Friedrichs:

$$H_{\mathrm{LF}} = \frac{1}{2}\left(\widetilde{F}(U_l) + \widetilde{F}(U_r) - \Lambda\,(U_r - U_l)\right),$$

where $\Lambda > 0$ is a scalar constant chosen to be larger than the fastest wave speed found on the computational domain.

2. Roe:

$$H_{\mathrm{Roe}} = \frac{1}{2}\left(\widetilde{F}(U_l) + \widetilde{F}(U_r) - R\Lambda L\,(U_r - U_l)\right),$$

where $\Lambda = \mathrm{diag}(|-\bar{\xi} - c|, |-\bar{\xi}|, |-\bar{\xi} + c|)$, and R and L are the matrices of the right and left eigenvectors to the Jacobian of $\widetilde{F}$ evaluated at the midpoint $U_{\mathrm{Roe}} = \frac{1}{2}(U_l + U_r)$. Since we use the equation of state $p = 1/2\rho^2$, the midpoint yields an exact Roe average.

In order to investigate the structure of the solution near the triple point in a manner that has as little numerical bias as possible, we opted to first solve the problem using the classic first-order accurate Lax–Friedrichs finite

(a)

(b)

(c)
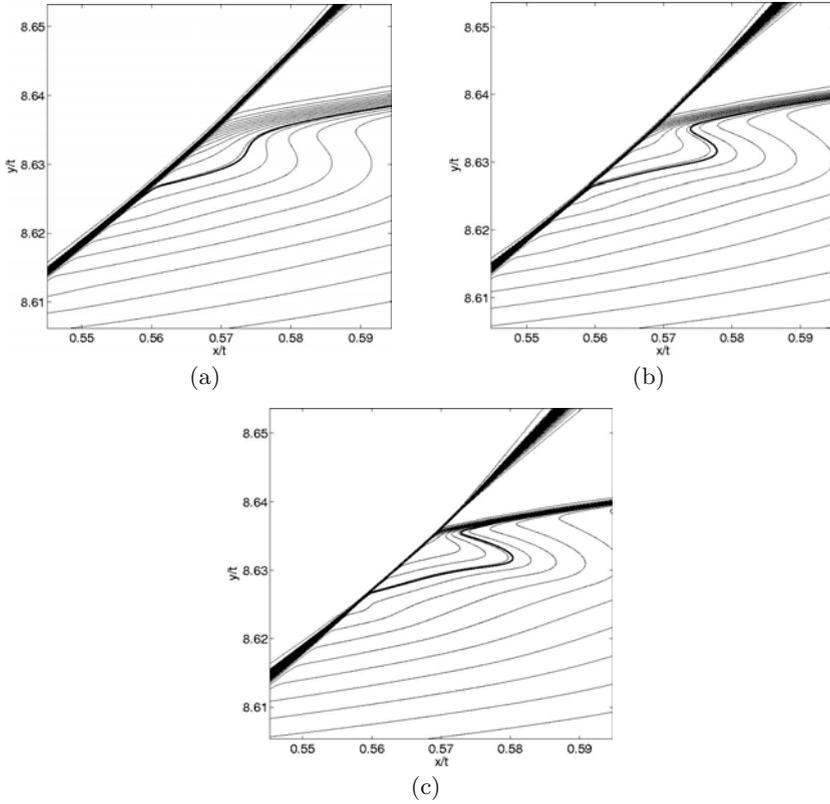
Fig. 7. Density contour plots for the nonlinear wave system using the first order accurate Lax–Friedrichs finite volume scheme in a neighborhood of the triple point. The region shown includes the locally refined 760×760 grid in (a), the 1280×1024 grid in (b) and the 2048×1320 grid in (c). The heavy line below the reflected shock and to the right of the Mach stem delineates a supersonic patch found within the subsonic zone. There is a slight indication of an expansion fan behind the leading triple point in (c).

volume scheme. That is, the Lax–Friedrichs flux is used in conjunction with piecewise constant cell-wise reconstruction. Figure 7 depicts a closeup of what was found on three grids with increasing refinement. The largest grid (c) contains approximately 11 million grid points. Approximately one quarter of these are contained in a square of length 0.05 units centered on the triple point. The solution in (c) clearly resolves a small patch of supersonic flow behind the triple point. This patch is quite small with width of approximately 0.03 and height of approximately 0.01. Note the fattening of the incident and Mach shocks as they leave the region of extreme grid refinement. The much weaker reflected shock is well resolved since it is aligned with the grid, and the grid in the direction normal to the reflected shock is very fine near the triple point.

(a)   (b)
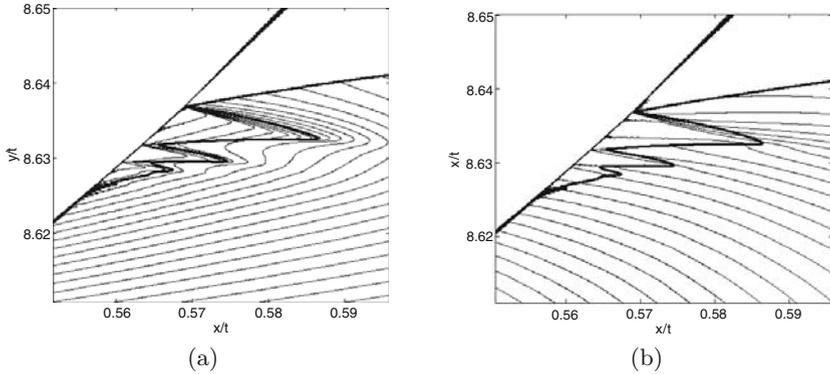
Fig. 8. Density contours (a) and x-momentum contours (b) for the nonlinear wave system using a high-order Roe scheme. These were obtained on the same grid depicted in Figure 7(c). There is now clear evidence of the sequence of interacting shocks and expansions seen earlier for the UTSDE. The heavy line is the sonic line and again delineates the supersonic patch.

The width of the supersonic patch is approximately 5% of the length of the Mach stem. There is a slight indication of an expansion fan at the triple point, but at this level of grid refinement there is no evidence yet of the sequence of shocks and expansions seen in Figure 4.

There comes a time when the results from a first-order scheme are, at best, inadequate, because of hardware limitations. The large grid results just displayed used a grid whose smallest grid size was on the order of one millionth of the extent of the computational domain. Moreover, these problems are steady and, therefore, require hundreds of thousands of pseudo-time iterations. At this stage we, therefore, employed a (perhaps) somewhat less unbiased numerical approach – a high-order scheme based on the Roe numerical flux. High-order accuracy is achieved by using a piecewise quadratic reconstruction limited in characteristic variables. We give the finest grid results from this approach in Figure 8. Three shock/expansion pairs are now clearly evident. The primary wave is at the triple point and two others can be seen along the Mach stem, a pattern very similar to that found for the UTSDE.

4 Weak Shock Irregular Reflection for the Euler Equations

We compute numerical solutions for the Euler equations (2) with $\gamma = 5/3$. A weak $M = 1.04$ vertically aligned incident shock impinges on a $\theta_w = 11.5°$ ramp. These data correspond to parameter $\check{a} \approx 1/2$ in the UTSDE model from Section 2. The grid is defined by a conformal map of the form $z = w^\alpha$, and so it is orthogonal with a singularity at the ramp apex $x = y = 0$. The upstream speed of sound $c_r = 1$, and boundary data on the left, right and top is given
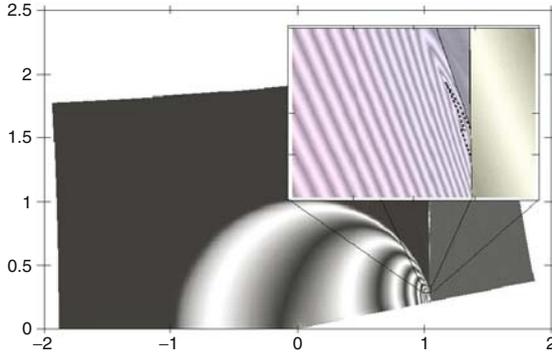
Fig. 9. The geometry of the $M = 1.04/11.5°$ Euler example. The insert indicates the region where extreme local grid refinement is performed.

to exactly agree with this shock located at $x = 1.04$. The lower boundary condition mimics symmetry about the x-axis for $x < 0$ and symmetry with respect to the ramp for $x > 0$. The grid geometry can be seen in Figure 9. This problem is well outside the range where regular reflection solutions are possible. Refer again to the figure to see that its numerical solution (under the insert) clearly resembles single Mach reflection. However, Mach reflection (where three plane shocks meet at a point) is also not possible for a shock this weak [Hen87]. This example demonstrates a classic von Neumann triple point paradox.

This problem is solved in self-similar coordinates by essentially the same high order Roe method discussed in the previous section. However, we simplify the Roe approach by again evaluating the Roe matrix at the midpoint, which for the Euler equations is only an approximation to the Roe average. Also, to avoid spurious expansion shocks, artificial dissipation on the order of $O(|U_r - U_l|)$ is appended to the diagonal part of the Roe dissipation matrix in a field by field manner.

We locally refine a very small neighborhood around the apparent triple point as done earlier. The full finest grid has eleven million grid points with $800 \times 2000 = 1.6 \times 10^6$ ($\Delta x \approx 5 \times 10^{-7}$) devoted to the local refinement. We plot the sonic number $\mathcal{M}$ which is defined as follows. The eigenvalue corresponding to a fast shock in unit direction $\mathbf{n}$ for the self-similar Euler flux Jacobian is

$$\lambda = (u - \xi, v - \eta) \cdot \mathbf{n} + c$$

where $\xi = x/t$ and $\eta = y/t$. Define $r^2 = \xi^2 + \eta^2$ and set $\mathbf{n} = (\xi, \eta)/r$, $u_n = (u, v) \cdot \mathbf{n}$ to find

$$\lambda = c \left(\frac{u_n - r}{c} + 1 \right) = c(1 - \mathcal{M}) \quad \text{where} \quad \mathcal{M} = \frac{r - u_n}{c}.$$
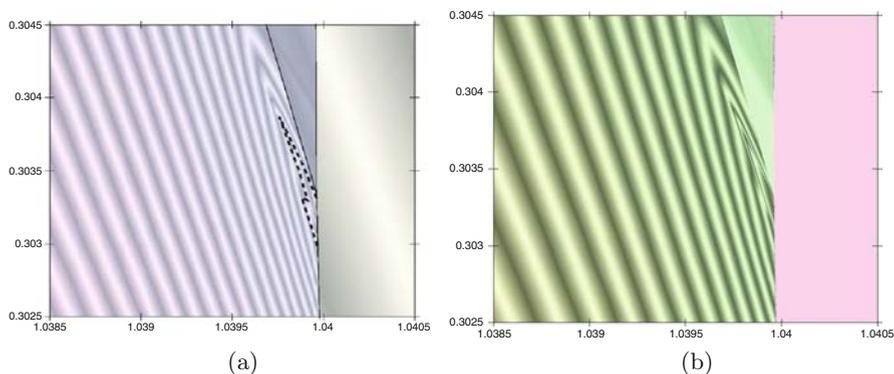
Fig. 10. A closeup of the Euler triple point. The sonic number $\mathcal{M}$ on the left and density ρ on the right. The dotted line on the left delineates the supersonic patch within the subsonic zone behind the Mach stem.
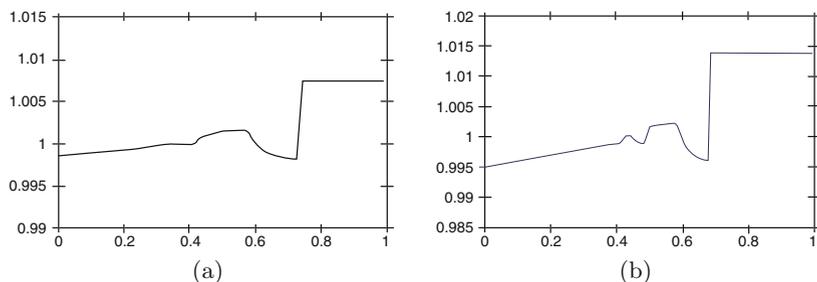
Fig. 11. Vertical cross sections of $\mathcal{M}$ taken bottom-up slightly to the left of the Mach stem. On the left $M = 1.04/11.5°$. The reflected shock is the large jump. Note the crossings at $\mathcal{M} = 1$. On the right, a second example problem with a slightly stronger incident shock $M = 1.075/15.0°$. The evidence of a sequence of shock/expansion wave pairs is stronger for this second example.

When $\mathcal{M} < 1$, the flow is called subsonic. When $\mathcal{M} > 1$, the flow is called supersonic. In this sense, when crossing through a self-similar stationary shock, the fact that $\mathcal{M}$ crosses from subsonic to supersonic is nothing more than the entropy condition $\lambda_l > s > \lambda_r$.

Figure 10 gives a sonic number contour plot (a) and density contours (b) in the triple point neighborhood. Clearly the evidence for Guderley Mach reflection in this example is not nearly as compelling as found for our earlier examples. However, these shocks are extremely weak. In recent work for a $\gamma = 7/5$ gas, we slightly strengthened the incident Mach number, $M = 1.075$, and obtained far more conclusive results. See the sonic number cross sections depicted in Figure 11.

References

[BH92]   M. Brio and J. K. Hunter. Mach reflection for the two-dimensional Burgers equation. *Phys. D*, 60:194–207, 1992.

[BT49]   W. Bleakney and A. H. Taub. Interaction of shock waves. *Rev. Modern Physics*, 21:584–605, 1949.

[CF76]   R. Courant and K. O. Friedrichs. *Supersonic Flow and Shock Waves*. Springer, 1976.

[CH90]   P. Colella and L. F. Henderson. The von Neumann paradox for the diffraction of weak shock waves. *J. Fluid Mech.*, 213:71–94, 1990.

[ČK98]   S. Čanić and B. L. Keyfitz. Quasi-one-dimensional Riemann problems and their role in self-similar two-dimensional problems. *Arch. Rational Mech. Anal.*, 144:233–258, 1998.

[ČKK01]   S. Čanić, B. L. Keyfitz, and E. H. Kim. Mixed hyperbolic-elliptic systems in self-similar flows. *Bol. Soc. Bras. Mat.*, 32:1–23, 2001.

[ČKK05]   S. Čanić, B. L. Keyfitz, and E. H. Kim. Free boundary problems for nonlinear wave systems: Mach stems for interacting shocks. *SIAM J. Math. Anal.*, 37:1947–1977, 2005.

[Gud47]   K. G. Guderley. Considerations of the structure of mixed subsonic-supersonic flow patterns. Air Material Command Tech. Report, F-TR-2168-ND, ATI No. 22780, GS-AAF-Wright Field 39, U.S. Wright-Patterson Air Force Base, Dayton, Ohio, October 1947.

[Gud62]   K. G. Guderley. *The Theory of Transonic Flow*. Pergamon Press, Oxford, 1962.

[HB00]   J. K. Hunter and M. Brio. Weak shock reflection. *J. Fluid Mech.*, 410:235–261, 2000.

[Hen66]   L. F. Henderson. On a class of multi-shock intersections in a perfect gas. *Aero. Q.*, 17:1–20, 1966.

[Hen87]   L. F. Henderson. Regions and boundaries for diffracting shock wave systems. *Z. Angew. Math. Mech.*, 67:73–86, 1987.

[HT04]   J. K. Hunter and A. M. Tesdall. Weak shock reflection. In D. Givoli, M. Grote, and G. Papanicolaou, editors, *A Celebration of Mathematical Modeling*. Kluwer Academic Press, New York, 2004.

[KF94]   B. L. Keyfitz and M. C. Lopes Filho. A geometric study of shocks in equations that change type. *J. Dynam. Differential Equations*, 6:351–393, 1994.

[Neu43]   J. von Neumann. Oblique reflection of shocks. Explosives Research Report 12, Bureau of Ordinance, 1943.

[Neu63]   J. von Neumann. *Collected Works, Vol. 6.* Pergamon Press, New York, 1963.

[Ric81]   R. D. Richtmeyer. *Principles of Mathematical Physics, Vol. 1.* Springer, 1981.

[SA05]   B. Skews and J. Ashworth. The physical nature of weak shock wave reflection. *J. Fluid Mech.*, 542:105–114, 2005.

[Ste59]   J. Sternberg. Triple-shock-wave intersections. *Phys. Fluids*, 2:179–206, 1959.

[STS92]   A. Sasoh, K. Takayama, and T. Saito. A weak shock wave reflection over wedges. *Shock Waves*, 2:277–281, 1992.

[TH02]   A. M. Tesdall and J. K. Hunter. Self-similar solutions for weak shock reflection. *SIAM J. Appl. Math.*, 63:42–61, 2002.

[TR94]   E. G. Tabak and R. R. Rosales. Focusing of weak shock waves and
            the von Neumann paradox of oblique shock reflection. *Phys. Fluids*,
            6:1874–1892, 1994.
[TSK06]   A. M. Tesdall, R. Sanders, and B. L. Keyfitz. The triple point paradox
            for the nonlinear wave system. *SIAM J. Appl. Math.*, 67:321–336, 2006.
[VK99]   E. Vasil'ev and A. Kraiko. Numerical simulation of weak shock diffrac-
            tion over a wedge under the von Neumann paradox conditions. *Comput.
            Math. Math. Phys.*, 39:1335–1345, 1999.
[ZBHW00]   A. Zakharian, M. Brio, J. K. Hunter, and G. Webb. The von Neumann
            paradox in weak shock reflection. *J. Fluid Mech.*, 422:193–205, 2000.

Part III

Domain Decomposition Methods

A Lagrange Multiplier Based Domain Decomposition Method for the Solution of a Wave Problem with Discontinuous Coefficients

Serguei Lapin[1], Alexander Lapin[2], Jacques Périaux[3,4], and Pierre-Marie Jacquart[5]

[1] Department of Mathematics, Washington State University, Pullman WA 99164 USA `slapin@math.wsu.edu`
[2] Kazan State University, Department of Computational Mathematics and Cybernetics, 18 Kremlyovskaya St., Kazan 420008, Russia `alapin@ksu.ru`
[3] Pole Scientifique Dassault/UPMC `jperiaux@free.fr`
[4] University of Jyväskylä, Department of Mathematical Information Technology, P.O. Box 35 (Agora), FI-40014 University of Jyväskylä, Finland
[5] Dassault Aviation, 78, Quai Marcel Dassault, Cedex 300, Saint-Cloud 92552, France `pierre-marie.jacquart@dassault-aviation.fr`

Summary. In this paper we consider the numerical solution of a linear wave equation with discontinuous coefficients. We divide the computational domain into two subdomains and use explicit time difference scheme along with piecewise linear finite element approximations on semimatching grids. We apply boundary supported Lagrange multiplier method to match the solution on the interface between subdomains. The resulting system of linear equations of the "saddle-point" type is solved efficiently by a conjugate gradient method.

1 Problem Formulation

Let $\Omega \subset \mathbb{R}^2$ be a rectangular domain with sides parallel to the coordinate axes and boundary Γ_{ext} (see Fig. 1). Now let $\Omega_2 \subset \Omega$ be a proper subdomain of Ω with a curvilinear boundary and $\Omega_1 = \Omega \setminus \bar{\Omega}_2$.

We consider the following linear wave problem:

$$
\begin{cases}
\varepsilon \dfrac{\partial^2 u}{\partial t^2} - \nabla \cdot (\mu^{-1} \nabla u) = f & \text{in } \Omega \times (0, T), \\[2mm]
\sqrt{\varepsilon\,\mu^{-1}}\dfrac{\partial u}{\partial t} + \mu^{-1}\dfrac{\partial u}{\partial \mathbf{n}} = 0 & \text{on } \Gamma_{\text{ext}} \times (0, T), \\[2mm]
u(x, 0) = \dfrac{\partial u}{\partial t}(x, 0) = 0.
\end{cases}
\tag{1}
$$
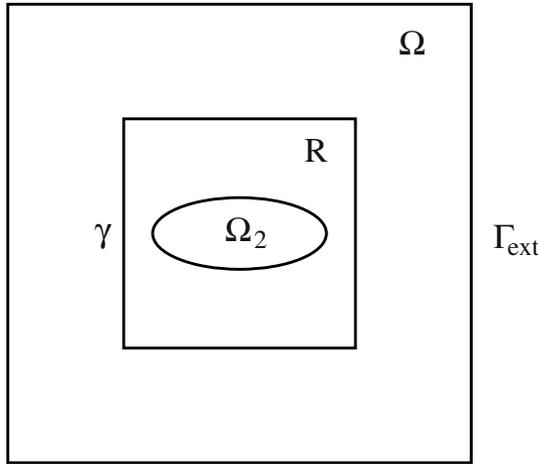
Fig. 1. Computational domain.

Here $\nabla u = (\frac{\partial u}{\partial x_1}, \frac{\partial u}{\partial x_2})$, $\mathbf{n}$ is the unit outward normal vector on Γ_{ext}. We suppose that $\mu_i = \mu|_{\Omega_i}$, $\varepsilon_i = \varepsilon|_{\Omega_i}$ are positive constants for all $i = 1, 2$ and $f_i = f|_{\Omega_i} \in C(\bar{\Omega}_i \times [0, T])$.

Let

$$\varepsilon(x) = \begin{cases} \varepsilon_1 & \text{if } x \in \Omega_1, \\ , \varepsilon_2 & \text{if } x \in \Omega_2, \end{cases} \quad \text{and} \quad \mu(x) = \begin{cases} \mu_1 & \text{if } x \in \Omega_1, \\ \mu_2 & \text{if } x \in \Omega_2. \end{cases}$$

We define a weak solution of problem (1) as a function u such that

$$u \in L^\infty(0, T; H^1(\Omega)), \ \frac{\partial u}{\partial t} \in L^\infty(0, T; L^2(\Omega)), \ \frac{\partial u}{\partial t} \in L^2(0, T; L^2(\Gamma_{\text{ext}})) \quad (2)$$

for a.a. $t \in (0, T)$ and for all $w \in H^1(\Omega)$ satisfying the equation

$$\int_\Omega \varepsilon(x) \frac{\partial^2 u}{\partial t^2} w dx + \int_\Omega \mu^{-1}(x) \nabla u \cdot \nabla w dx + \sqrt{\varepsilon_1 \mu_1^{-1}} \int_{\Gamma_{\text{ext}}} \frac{\partial u}{\partial t} w d\Gamma = \int_\Omega f w dx$$
$$(3)$$

with the initial conditions

$$u(x, 0) = \frac{\partial u}{\partial t}(x, 0) = 0.$$

Note that the first term in (3) means the duality between $(H^1(\Omega))^*$ and $H^1(\Omega)$.

Now, using the Faedo–Galerkin method (as in [DL92]), one can prove the following:

Theorem 1. *Under the assumptions (2) there exists a unique weak solution of problem (1).*

Let

$$E(t) = \frac{1}{2} \int_\Omega \varepsilon(x) \left| \frac{\partial u}{\partial t} \right|^2 dx + \frac{1}{2} \int_\Omega \mu^{-1}(x) |\nabla u|^2 dx$$

be the energy of the system. We take $w = \frac{\partial u}{\partial t}$ in (3) and obtain:

$$\frac{dE(t)}{dt} + \sqrt{\varepsilon_1 \mu_1^{-1}} \int_{\Gamma_{\text{ext}}} \left(\frac{\partial u}{\partial t} \right)^2 d\Gamma = \int_\Omega f \frac{\partial u}{\partial t} dx \leq \|f\|_{L^2(\Omega)} \| \frac{\partial u}{\partial t} \|_{L^2(\Omega)},$$

since $E(0) = 0$, the following stability inequality holds:

$$E(t) \leq \text{const} \, T \|f\|_{L^2(\Omega \times (0,T))}, \quad \forall t \in (0,T).$$

In order to use a structured grid in a part of the domain Ω, we introduce a rectangular domain R with sides parallel to the coordinate axes, such that $\Omega_2 \subset R \subset \Omega$ with γ the boundary of R (Fig. 1).

Define $\tilde{\Omega} = \Omega \setminus \bar{R}$ and let the subscript 1 of a function v_1 mean that this function is defined over $\tilde{\Omega} \times (0,T)$, while v_2 is a function defined over $R \times (0,T)$.

Now we formulate the problem (3) variationally as follows: Let

$$W_1 = \left\{ v \in L^\infty(0,T; H^1(\tilde{\Omega})), \ \frac{\partial v}{\partial t} \in L^\infty(0,T; L^2(\tilde{\Omega})), \ \frac{\partial v}{\partial t} \in L^2(0,T; L^2(\Gamma_{\text{ext}})) \right\},$$

$$W_2 = \left\{ v \in L^\infty(0,T; H^1(R)), \ \frac{\partial v}{\partial t} \in L^\infty(0,T; L^2(R))) \right\},$$

Find a pair $(u_1, u_2) \in W_1 \times W_2$, such that $u_1 = u_2$ on $\gamma \times (0,T)$ and for a.a. $t \in (0,T)$

$$\begin{cases} \displaystyle \int_{\tilde{\Omega}} \varepsilon_1 \frac{\partial^2 u_1}{\partial t^2} w_1 dx + \int_{\tilde{\Omega}} \mu_1^{-1} \nabla u_1 \cdot \nabla w_1 dx + \int_R \varepsilon(x) \frac{\partial^2 u_2}{\partial t^2} w_2 dx \\ \displaystyle + \int_R \mu^{-1}(x) \nabla u_2 \cdot \nabla w_2 dx + \sqrt{\varepsilon_1 \mu_1^{-1}} \int_{\Gamma_{\text{ext}}} \frac{\partial u_1}{\partial t} w_1 d\Gamma = \int_{\tilde{\Omega}} f_1 w_1 dx + \int_R f_2 w_2 dx, \\ \text{for all } (w_1, w_2) \in H^1(\tilde{\Omega}) \times H^1(R) \text{ such that } w_1 = w_2 \text{ on } \gamma, \\ \displaystyle u(x,0) = \frac{\partial u}{\partial t}(x,0) = 0. \end{cases}$$

(4)

Now, introducing the interface supported Lagrange multiplier λ (a function defined over $\gamma \times (0,T)$), the problem (4) can be written in the following way:

Find a triple $(u_1, u_2, \lambda) \in W_1 \times W_2 \times L^\infty(0,T; H^{-1/2}(\gamma))$, which for a.a. $t \in (0,T)$ satisfies

$$\int_{\tilde{\Omega}} \varepsilon_1 \frac{\partial^2 u_1}{\partial t^2} w_1 dx + \int_{\tilde{\Omega}} \mu_1^{-1} \nabla u_1 \cdot \nabla w_1 dx + \int_R \varepsilon(x) \frac{\partial^2 u_2}{\partial t^2} w_2 dx$$

$$+ \int_R \mu^{-1}(x) \nabla u_2 \cdot \nabla w_2 dx + \sqrt{\varepsilon_1 \mu_1^{-1}} \int_{\Gamma_{\text{ext}}} \frac{\partial u_1}{\partial t} w_1 d\Gamma + \int_\gamma \lambda(w_2 - w_1) d\gamma$$

$$= \int_{\tilde{\Omega}} f_1 w_1 dx + \int_R f_2 w_2 dx \quad \text{for all } w_1 \in H^1(\tilde{\Omega}), \ w_2 \in H^1(R), \tag{5}$$

$$\int_\gamma \zeta(u_2 - u_1) d\gamma = 0 \quad \text{for all } \zeta \in H^{-1/2}(\gamma), \tag{6}$$

and the initial conditions from (1).

Remark 1. We selected the time dependent approach to capture harmonic solutions since it substantially simplifies the linear algebra of the solution process. Furthermore, there exist various techniques to speed up the convergence of transient solutions to periodic ones (see, e.g., [BDG+97]).

2 Time Discretization

In order to construct a finite difference approximation in time of the problem (5), (6), we partition the segment $[0, T]$ into N intervals using a uniform discretization step $\Delta t = T/N$. Let $u_i^n \approx u_i(n \Delta t)$ for $i = 1, 2$, $\lambda^n \approx \lambda(n \Delta t)$. The explicit in time semidiscrete approximation to the problem (5), (6) reads as follows:

$$u_i^0 = u_i^1 = 0$$

for $n = 1, 2, \ldots, N - 1$. Find $u_1^{n+1} \in H^1(\tilde{\Omega})$, $u_2^{n+1} \in H^1(R)$ and $\lambda^{n+1} \in H^{-1/2}(\gamma)$ such that

$$\int_{\tilde{\Omega}} \varepsilon_1 \frac{u_1^{n+1} - 2u_1^n + u_1^{n-1}}{\Delta t^2} w_1 dx + \int_{\tilde{\Omega}} \mu_1^{-1} \nabla u_1^n \cdot \nabla w_1 dx +$$

$$+ \int_R \varepsilon(x) \frac{u_2^{n+1} - 2u_2^n + u_2^{n-1}}{\Delta t^2} w_2 dx + \int_R \mu^{-1}(x) \nabla u_2^n \cdot \nabla w_2 dx +$$

$$+ \sqrt{\varepsilon_1 \mu_1^{-1}} \int_{\Gamma_{\text{ext}}} \frac{u_1^{n+1} - u_1^{n-1}}{2\Delta t} w_1 d\Gamma + \int_\gamma \lambda^{n+1}(w_2 - w_1) d\gamma =$$

$$= \int_{\tilde{\Omega}} f_1^n w_1 dx + \int_R f_2^n w_2 dx \quad \text{for all } w_1 \in H^1(\tilde{\Omega}), w_2 \in H^1(R), \tag{7}$$

$$\int_\gamma \zeta(u_2^{n+1} - u_1^{n+1}) d\gamma = 0 \quad \text{for all } \zeta \in H^{-1/2}(\gamma). \tag{8}$$

Remark 2. The integral over γ is written formally; the exact formulation requires the use of the duality pairing $\langle \cdot, \cdot \rangle$ between $H^{-1/2}(\gamma)$ and $H^{1/2}(\gamma)$.

3 Fully Discrete Scheme

To construct a fully discrete space-time approximation to the problem (5), (6), we will use a lowest order finite element method on two grids semimatching on γ (Fig. 2) for the space discretization. Namely, let $\mathcal{T}_{1h}$ and $\mathcal{T}_{2h}$ be triangulations of $\tilde{\Omega}$ and R, respectively. Further we suppose that both triangulations are regular in the sense that

$$\frac{r(e)}{h(e)} \leq q = \text{const}$$

for all $e \in \mathcal{T}_{1h}$ and $e \in \mathcal{T}_{2h}$, where q does not depend on e; $r(e)$ is the radius of the circle inscribed in e, while $h(e)$ is the diameter of e.

We denote by $\mathcal{T}_{1h}$ a coarse triangulation and by $\mathcal{T}_{2h}$ a fine one. Every edge $\partial e \subset \gamma$ of a triangle $e \in \mathcal{T}_{1h}$ is supposed to consist of m_e edges of triangles from $\mathcal{T}_{2h}$, $1 \leq m_e \leq m$ for all $e \in \mathcal{T}_{1h}$

Moreover, let a triangulation $\mathcal{T}_{2h}$ be such that the curvilinear boundary $\partial \Omega_2$ is approximated by a polygonal line consisting of the edges of triangles from $\mathcal{T}_{2h}$ whose vertices belong to $\partial \Omega_2$. Further, we say that a triangle $e \in \mathcal{T}_{2h}$ lies in Ω_2 if its larger part lies in Ω_2, i.e. $\text{meas}(e \cap \Omega_2) > \text{meas}(e \cap (R \setminus \Omega_2))$, otherwise this triangle lies in $R \setminus \Omega_2$.

Let $V_{1h} \subset H^1(\tilde{\Omega})$ be the space of the functions globally continuous, and affine on each $e \in \mathcal{T}_{1h}$, i.e. $V_{1h} = \{u_h \in H^1(\tilde{\Omega}) \mid u_h \in P_1(e) \quad \forall e \in \mathcal{T}_{1h}\}$. Similarly, $V_{2h} \subset H^1(R)$ is the space of the functions globally continuous, and affine on each $e \in \mathcal{T}_{2h}$.

For approximating the Lagrange multipliers space $\Lambda = H^{-1/2}(\gamma)$ we proceed as follows. Assume that on γ, $\mathcal{T}_{1h}$ is two times coarser than $\mathcal{T}_{2h}$. Then let us divide every edge ∂e of a triangle e from the coarse grid $\mathcal{T}_{1h}$, which is located on γ ($\partial e \subset \gamma$), into two parts using its midpoint. Now, we consider the space of the piecewise constant functions, which are constant on every union of half-edges with a common vertex (see Fig. 3).

Further, we use quadrature formulas for approximating the integrals over the triangles from $\mathcal{T}_{1h}$ and $\mathcal{T}_{2h}$, as well as over Γ_{ext}. For a triangle e we set
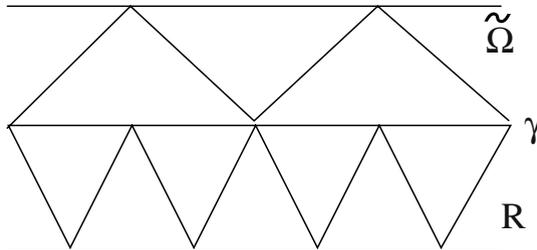
Fig. 2. Semimatching mesh on γ.
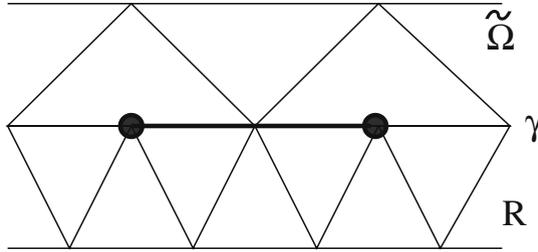
Fig. 3. Space Λ is the space of the piecewise constant functions defined on every union of half-edges with common vertex.

$$\int_e \phi(x)dx \approx \frac{1}{3}\,\mathrm{meas}(e)\sum_{i=1}^{3}\phi(a_i) \equiv S_e(\phi),$$

where the a_i's are the vertices of e and $\phi(x)$ is a continuous function on e. Similarly,

$$\int_{\partial e} \phi(x)dx \approx \frac{1}{2}\,\mathrm{meas}(\partial e)\sum_{i=1}^{2}\phi(a_i) \equiv S_{\partial e}(\phi),$$

where a_i's are the endpoints of the segment ∂e and $\phi(x)$ is a continuous function on this segment.

We use the notations:

$$S_i(\phi) = \sum_{e \in \mathcal{T}_{ih}} S_e(\phi),\ \ i=1,2, \quad \text{and} \quad S_{\Gamma_{\mathrm{ext}}}(\phi) = \sum_{\partial e \subset \Gamma_{\mathrm{ext}}} S_{\partial e}(\phi).$$

Now, the fully discrete problem reads as follows: Let $u_{ih}^0 = u_{ih}^1 = 0$, $i = 1,2$. For $n = 1, 2, \ldots, N-1$, find $(u_{1h}^{n+1}, u_{2h}^{n+1}, \lambda_h^{n+1}) \in V_{1h} \times V_{2h} \times \Lambda_h$ such that

$$\frac{\varepsilon_1}{\Delta t^2}S_1((u_{1h}^{n+1} - 2u_{1h}^n + u_{1h}^{n-1})w_{1h}) + S_1(\mu_1^{-1}\nabla u_{1h}^n \cdot \nabla w_{1h})+$$

$$+ \frac{1}{\Delta t^2}S_2(\varepsilon(x)(u_{2h}^{n+1} - 2u_{2h}^n + u_{2h}^{n-1})w_{2h}) + S_2(\mu^{-1}(x)\nabla u_{2h}^n \cdot \nabla w_{2h})+$$

$$+ \frac{\sqrt{\varepsilon_1 \mu_1^{-1}}}{2\Delta t}S_{\Gamma_{\mathrm{ext}}}((u_{1h}^{n+1} - u_{1h}^{n-1})w_{1h}) + \int_\gamma \lambda_h^{n+1}(w_{2h} - w_{1h})d\gamma =$$

$$= S_1(f_1^n w_{1h}) + S_2(f_2^n w_{2h}) \quad \text{for all } w_{1h} \in V_{1h}, w_{2h} \in V_{2h}, \tag{9}$$

$$\int_\gamma \zeta_h(u_{2h}^{n+1} - u_{1h}^{n+1})d\gamma = 0 \quad \text{for all } \zeta_h \in \Lambda_h. \tag{10}$$

Note that in $S_2(\varepsilon(x)(u_{2h}^{n+1} - 2u_{2h}^n + u_{2h}^{n-1})w_{2h})$ we take $\varepsilon(x) = \varepsilon_2$ if a triangle $e \in \mathcal{T}_{2h}$ lies in Ω_2 and $\varepsilon(x) = \varepsilon_1$ if it lies in $R \setminus \Omega_2$, and similarly for $S_2(\mu^{-1}(x)\nabla u_{2h}^n \nabla w_{2h})$.

Denote by $\mathbf{u}_1$, $\mathbf{u}_2$ and $\boldsymbol{\lambda}$ the vectors of the nodal values of the corresponding functions u_{1h}, u_{2h} and λ_h. Then, in order to find $\mathbf{u}_1^{n+1}$, $\mathbf{u}_2^{n+1}$ and $\boldsymbol{\lambda}^{n+1}$ for a fixed time t^{n+1}, we have to solve a system of linear equations such as

$$\mathbf{Au} + \mathbf{B}^T \boldsymbol{\lambda} = \mathbf{F}, \tag{11}$$

$$\mathbf{Bu} = 0, \tag{12}$$

where matrix $\mathbf{A}$ is diagonal, positive definite and defined by

$$(\mathbf{Au}, \mathbf{w}) = \frac{\varepsilon_1}{\Delta t^2} S_1(u_{1h} w_{1h}) + \frac{1}{\Delta t^2} S_2(\varepsilon(x) u_{2h} w_{2h}) + \frac{\sqrt{\varepsilon_1 \mu_1^{-1}}}{2\Delta t} S_{\Gamma_{\text{ext}}}(u_{1h} w_{1h}),$$

and where the rectangular matrix $\mathbf{B}$ is defined by

$$(\mathbf{Bu}, \boldsymbol{\lambda}) = \int_\gamma \lambda_h (u_{2h} - u_{1h}) d\Gamma,$$

and vector $\mathbf{F}$ depends on the nodal values of the known functions u_{1h}^n, u_{2h}^n, u_{1h}^{n-1} and u_{2h}^{n-1}.

Eliminating $\mathbf{u}$ from the equation (11), we obtain

$$\mathbf{B}\mathbf{A}^{-1}\mathbf{B}^T \boldsymbol{\lambda} = \mathbf{B}\mathbf{A}^{-1}\mathbf{F}, \tag{13}$$

with a symmetric matrix $\mathbf{C} \equiv \mathbf{B}\mathbf{A}^{-1}\mathbf{B}^T$. Let us prove that $\mathbf{C}$ is positive definite. Obviously, $\ker \mathbf{C} = \ker \mathbf{B}^T$. Suppose, that $\mathbf{B}^T \boldsymbol{\lambda} = 0$, then a function $\lambda_h \in \Lambda_h$ corresponding to vector $\boldsymbol{\lambda}$ satisfies

$$I \equiv \int_\gamma \lambda_h u_h d\gamma = 0$$

for all $u_h \in V_{1h}$. Choose u_h equal to λ_h in the nodes of $\mathcal{T}_{1h}$ located on γ. Direct calculations give

$$I = \frac{1}{2} \sum_{i=1}^{N_\lambda} \left[\frac{h_i + h_{i+1}}{2} \lambda_i^2 + h_{i+1} \frac{(\lambda_i + \lambda_{i+1})^2}{2} \right],$$

where N_λ is the number of edges of $\mathcal{T}_{1h}$ on γ, h_i is the length of i-th edge and $h_{N_\lambda+1} \equiv h_1$, $\lambda_{N_\lambda+1} \equiv \lambda_1$. Thus, the equality $I = 0$ implies that $\boldsymbol{\lambda} = \mathbf{0}$, i.e. $\ker \mathbf{B}^T = \{\mathbf{0}\}$.

As a consequence we have

Theorem 2. *The problem (9), (10) has a unique solution* (u_h, λ_h).

Remark 3. A closely related domain decomposition method applied to the solution of linear parabolic equations is discussed in [Glo03].

4 Energy Inequality

Theorem 3. *Let $h_{\min}$ denote the minimal diameter of the triangles from $\mathcal{T}_{1h} \cup \mathcal{T}_{2h}$. There exists a positive number c such that the condition*

$$\Delta t \le c \, \min\{\sqrt{\varepsilon_1\mu_1}, \sqrt{\varepsilon_2\mu_2}\} \, h_{\min} \tag{14}$$

ensures the positive definiteness of the quadratic form

$$\mathcal{E}^{n+1} = \frac{1}{2}\varepsilon_1 S_1\left(\left(\frac{u_{1h}^{n+1} - u_{1h}^n}{\Delta t}\right)^2\right) + \frac{1}{2}S_2\left(\varepsilon\left(\frac{u_{2h}^{n+1} - u_{2h}^n}{\Delta t}\right)^2\right) +$$

$$+ \frac{1}{2}S_1\left(\mu_1^{-1}\left|\nabla\left(\frac{u_{1h}^{n+1} + u_{1h}^n}{2}\right)\right|^2\right) + \frac{1}{2}S_2\left(\mu^{-1}\left|\nabla\left(\frac{u_{2h}^{n+1} + u_{2h}^n}{2}\right)\right|^2\right) -$$

$$- \frac{\Delta t^2}{8}S_1\left(\mu_1^{-1}\left|\nabla\left(\frac{u_{1h}^{n+1} - u_{1h}^n}{\Delta t}\right)\right|^2\right) - \frac{\Delta t^2}{8}S_2\left(\mu^{-1}\left|\nabla\left(\frac{u_{2h}^{n+1} - u_{2h}^n}{\Delta t}\right)\right|^2\right), \tag{15}$$

which we call the discrete energy.

   The system (9), (10) satisfies the energy identity

$$\mathcal{E}^{n+1} - \mathcal{E}^n + \frac{\sqrt{\varepsilon_1\mu_1^{-1}}}{4\Delta t}S_{\Gamma_{\text{ext}}}((u_{1h}^{n+1} - u_{1h}^{n-1})^2) =$$

$$= \frac{1}{2}S_1(f_1^n(u_{1h}^{n+1} - u_{1h}^{n-1})) + \frac{1}{2}S_2(f_2^n(u_{2h}^{n+1} - u_{2h}^{n-1})) \tag{16}$$

and the numerical scheme is stable: There exists a positive number $M = M(T)$ such that

$$\mathcal{E}^n \le M\Delta t \sum_{k=1}^{n-1}(S_1((f_1^k)^2) + S_2((f_2^k)^2)), \quad \forall n. \tag{17}$$

Proof. Let $n \ge 1$. From the equation (10) written for t_{n+1} and t_{n-1} we obtain

$$\int_\gamma \zeta_h((u_{2h}^{n+1} - u_{2h}^{n-1}) - (u_{1h}^{n+1} - u_{1h}^{n-1}))d\gamma = 0 \quad \text{for all } \zeta_h \in \Lambda_h. \tag{18}$$

   Choosing

$$w_{1h} = \frac{u_{1h}^{n+1} - u_{1h}^{n-1}}{2}, \quad w_{2h} = \frac{u_{2h}^{n+1} - u_{2h}^{n-1}}{2}$$

in (9) and

$$\zeta_h = -\frac{\lambda_h^{n+1}}{2}$$

in (18), we add these equalities. Using the identities

$$(u_{ih}^{n+1} - 2u_{ih}^n + u_{ih}^{n-1})(u_{ih}^{n+1} - u_{ih}^{n-1}) = (u_{ih}^{n+1} - u_{ih}^n)^2 - (u_{ih}^n - u_{ih}^{n-1})^2$$

and

$$u_{ih}^n u_{ih}^{n+1} = \frac{1}{4}((u_{ih}^{n+1} + u_{ih}^n)^2 - (u_{ih}^{n+1} - u_{ih}^n)^2),$$

after several technical transformations we obtain

$$\mathcal{E}^{n+1} - \mathcal{E}^n + \frac{\sqrt{\varepsilon_1 \mu_1^{-1}}}{4\Delta t} S_{\Gamma_{\text{ext}}}((u_{1h}^{n+1} - u_{1h}^{n-1})^2) =$$
$$\frac{1}{2} S_1(f_1^n(u_{1h}^{n+1} - u_{1h}^{n-1})) + \frac{1}{2} S_2(f_2^n(u_{2h}^{n+1} - u_{2h}^{n-1})).$$

Therefore,

$$\mathcal{E}^{n+1} \leq \mathcal{E}^n + \frac{1}{2}\Delta t S_1^{1/2}((f_1^n)^2) \left[S_1^{1/2}\left(\left(\frac{u_{1h}^{n+1} - u_{1h}^n}{\Delta t} \right)^2 \right) + \right.$$
$$\left. +S_1^{1/2}\left(\left(\frac{u_{1h}^{n+1} - u_{1h}^n}{\Delta t} \right)^2 \right) \right] + \frac{1}{2}\Delta t S_2^{1/2}((f_2^n)^2) \left[S_2^{1/2}\left(\left(\frac{u_{2h}^{n+1} - u_{2h}^n}{\Delta t} \right)^2 \right) + \right.$$
$$\left. +S_2^{1/2}\left(\left(\frac{u_{2h}^{n+1} - u_{2h}^n}{\Delta t} \right)^2 \right) \right]. \quad (19)$$

Now, we will show that under the condition (14) the quadratic form $\mathcal{E}^n$ is positive definite; more precisely, that there exists a positive constant δ such that

$$\mathcal{E}^n \geq \delta \left(S_1\left(\left(\frac{u_{1h}^{n+1} - u_{1h}^n}{\Delta t} \right)^2 \right) + S_2\left(\left(\frac{u_{2h}^{n+1} - u_{2h}^n}{\Delta t} \right)^2 \right) \right). \quad (20)$$

Obviously, it is sufficient to prove the inequality

$$4\varepsilon_e \mu_e S_e(v_h^2) \geq \Delta t^2 S_e(|\nabla v_h|^2) \quad \forall e \in \mathcal{T}_{1h} \cup \mathcal{T}_{2h}, \ \forall v_h \in P_1(e), \quad (21)$$

where ε_e and μ_e are defined by $\varepsilon_e = \varepsilon_1$ or $\varepsilon_e = \varepsilon_2$ (respectively, $\mu_e = \mu_1$ or $\mu_e = \mu_2$). It is known that for a regular triangulation

$$S_e(|\nabla v_h|^2) \leq 1/c_1^2 h_e^{-2} S_e(v_h^2) \quad (22)$$

with a positive constant c_1, universal for all triangles e, where h_e is the minimal length of the sides of e. Combining (21) and (22), we observe that the time step Δt should satisfy the inequality

$$\Delta t \leq c\sqrt{\varepsilon_e \mu_e}\, h_e, \quad (c = \sqrt{2}c_1), \quad (23)$$

for all $e \in \mathcal{T}_{1h} \cup \mathcal{T}_{2h}$. Evidently, (14) ensures the validity of (23).

Further, using the relation (20), $\mathcal{E}^1 = 0$ and summing the inequalities (19), one obtains the stability inequality (17):

$$\mathcal{E}^n \leq M\Delta t \sum_{k=1}^{n-1}(S_1((f_1^k)^2) + S_2((f_2^k)^2)), \quad \forall n.$$

5 Numerical Experiments

In order to solve the system of linear equations (11)–(12) at each time step we use a Conjugate Gradient Algorithm in the form given by Glowinski and LeTallec [GL89]:

Step 1. $\boldsymbol{\lambda}^0$ given.
Step 2. $\mathbf{A}\mathbf{u}^0 = \mathbf{F} - \mathbf{B}\boldsymbol{\lambda}^0$.
Step 3. $\mathbf{g}^0 = -\mathbf{B}^T\mathbf{u}^0$.
Step 4. If $\|\mathbf{g}^0\| \leq \varepsilon_0$ take $\boldsymbol{\lambda} = \boldsymbol{\lambda}^0$,
       else $\mathbf{w}^0 = \mathbf{g}^0$.
Step 5. For $m \geq 0$, assuming that $\boldsymbol{\lambda}^m$, $\mathbf{g}^m$, $\mathbf{w}^m$ are known,
$$\mathbf{A}\bar{\mathbf{u}}^m = \mathbf{B}\mathbf{w}^m.$$
$$\bar{\mathbf{g}}^m = \mathbf{B}^T\bar{\mathbf{u}}^m.$$
$$\rho_m = \frac{|\mathbf{g}^m|^2}{(\bar{\mathbf{g}}^m, \bar{\mathbf{w}}^m)}.$$
$$\boldsymbol{\lambda}^{m+1} = \boldsymbol{\lambda}^m - \rho_m\mathbf{w}^m.$$
$$\mathbf{u}^{m+1} = \mathbf{u}^m + \rho_m\bar{\mathbf{v}}^m.$$
$$\mathbf{g}^{m+1} = \mathbf{g}^m - \rho_m\bar{\mathbf{g}}^m.$$
Step 6. If $\dfrac{\mathbf{g}^{m+1} \cdot \mathbf{g}^{m+1}}{\mathbf{g}^0 \cdot \mathbf{g}^0} \leq \varepsilon$ then take $\boldsymbol{\lambda} = \boldsymbol{\lambda}^{m+1}$,
       else $\gamma_m = \dfrac{\mathbf{g}^{m+1} \cdot \mathbf{g}^{m+1}}{\mathbf{g}^m \cdot \mathbf{g}^m}$.
Step 7. $\mathbf{w}^{m+1} = \mathbf{g}^{m+1} + \gamma_m\mathbf{w}^m$.
Step 8. Do $m = m + 1$ and go to Step 5.

We consider the problem (9)–(10) with a source term given by the harmonic planar wave

$$u^{inc} = -e^{ik(t-\boldsymbol{\alpha}\cdot\mathbf{x})}, \tag{24}$$

where $\{x_j\}_{j=1}^2$, $\{\alpha_j\}_{j=1}^2$, k is the angular frequency and $|\boldsymbol{\alpha}| = 1$.

For our numerical simulation we consider two cases: the first with the frequency of the incident wave $f = 0.6$ GHz and the second with $f = 1.2$ GHz, which gives us wavelengths $L = 0.5$ meters and $L = 0.25$ meters, respectively.

We performed a series of numerical experiments: scattering by a perfectly reflecting obstacle, wave propagation through a domain with an obstacle completely consisting of a coating material and scattering by an obstacle with coating.

First, we consider the scattering by a perfectly reflecting obstacle. For the experiment we have chosen Ω_2 to be in a form of a perfectly reflecting airfoil, and Ω is a 2 meter × 2 meter rectangle. We used a finite element mesh with 8019 nodes and 15324 elements in the case of $f = 0.6$ GHz (Fig. 4) and 19246 nodes and 37376 elements for $f = 1.2$ GHz.

Figure 5 shows the contour plot for the case when the incident wave is coming from the left and Figure 6 shows the case when the incident wave is coming from the lower left corner with an angle of 45°. For all the experiments
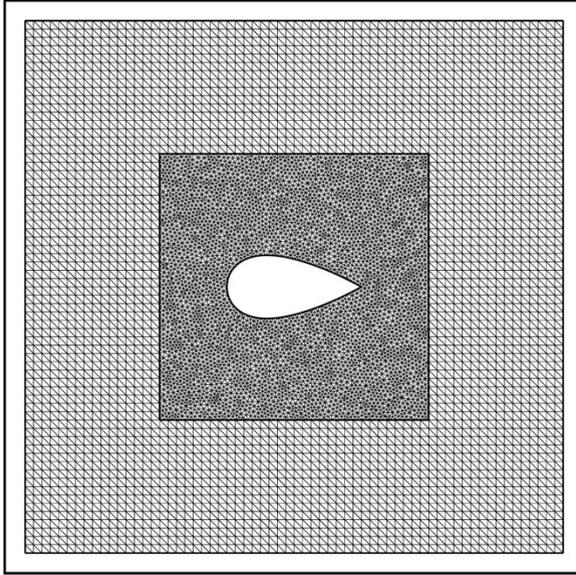
Fig. 4. Example of a finite element mesh.

Fig. 5. Contour plot of the real part of the solution for $L = 0.5$ (left) and $L = 0.25$ (right) meters. Incident wave coming from the left.

we chose the time step to be $\Delta t = T/50$, where $T = 1/f = 1.66 \times 10^{-9}$ sec is a time period corresponding to $L = 0.5$ meters and $T = 1/f = 0.83 \times 10^{-9}$ sec for $L = 0.25$ meters.

The next set of numerical experiments contains the simulations of wave propagation through a domain with an obstacle completely consisting of a

Fig. 6. Contour plot of the real part of the solution for $L = 0.5$ (left) and $L = 0.25$ (right) meters. Incident wave coming from the lower left corner with an angle of 45 degrees.

coating material. We have taken the coating material coefficients to be $\varepsilon_2 = 1$ and $\mu_2 = 9$, implying that the speed of propagation in the coating material is three times slower than in air. As before Ω is a 2 meter $\times$ 2 meter rectangle and Ω_2 has the shape of an airfoil.

For the solution of this problem for an incident frequency $f = 0.6$ GHz we have used a mesh with a total of 8435 nodes and 16228 elements. The time step was taken to be $\Delta t = T/50$, where $T = 1/f = 1.66 \times 10^{-9}$ sec is a time period. We used a mesh consisting of 20258 nodes (39514 elements) for solving the problem for an incident wave with the frequency $f = 1.2$ GHz. The time step was equal to $T/50$, $T = 1/f = 0.83 \times 10^{-9}$ sec.

In Figures 7 and 8 we present the contour plot of the real part of the solution for the incident frequency $L = 0.5$ and $L = 0.25$. We also performed numerical computations for the case when the obstacle is an airfoil with a coating (Figure 9). The coating region is moon shaped and, as before, $\varepsilon_2 = 1$ and $\mu_2 = 9$. We show in Figure 10 the contour plot of the real part of the solution for the incident frequency $L = 0.5$ meters and $L = 0.25$ meters for the case when the incident wave is coming from the left. Figure 11 presents the contour plot of the real part of the solution for incident frequency, $L = 0.5$ meters and $L = 0.25$ meters for the case when incident wave is coming from the lower left corner with angle equal to $45°$.

An important observation for all of the numerical experiments mentioned is that, despite the fact that a mesh discontinuity takes place over γ together with a weak forcing of the matching conditions, we do not observe a discontinuity of the computed fields.
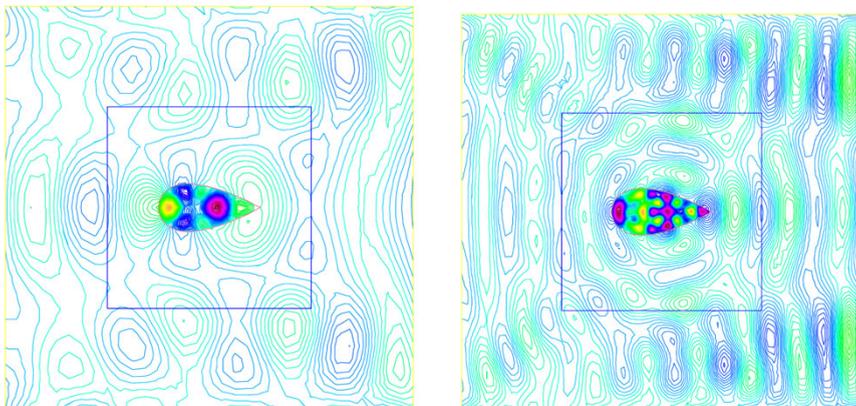
Fig. 7. Contour plot of the real part of the solution for $L = 0.5$ (left) and $L = 0.25$ (right). Incident wave coming from the left.
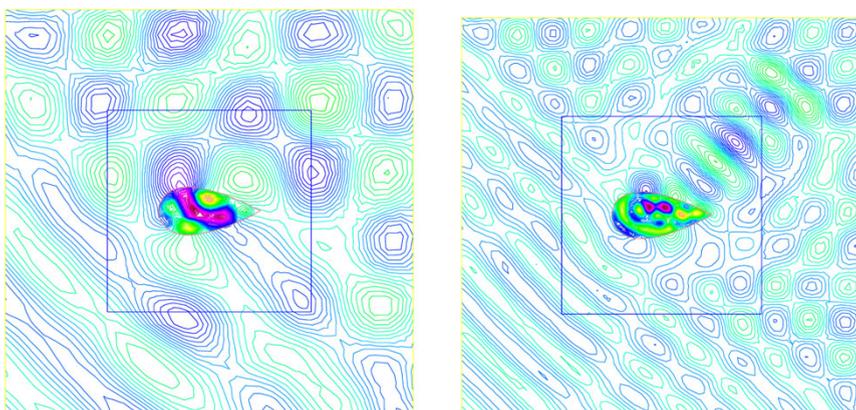
Fig. 8. Contour plot of the real part of the solution for $L = 0.5$ (left) and $L = 0.25$ (right). Incident wave coming from the lower left corner with an angle of 45 degrees.
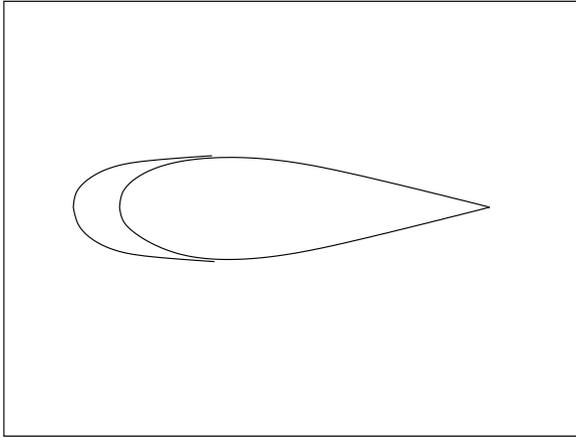
Fig. 9. Obstacle in a form of an airfoil with a coating.
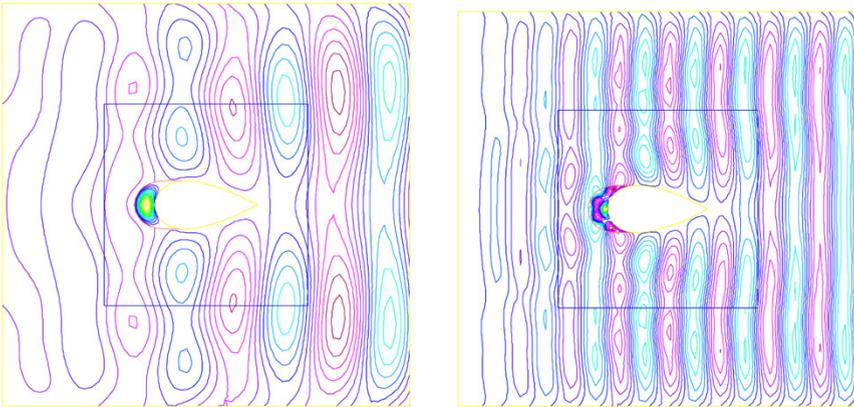
Fig. 10. Contour plot of the real part of the solution for $L = 0.5$ (left) and $L = 0.25$ (right). Incident wave coming from the left.
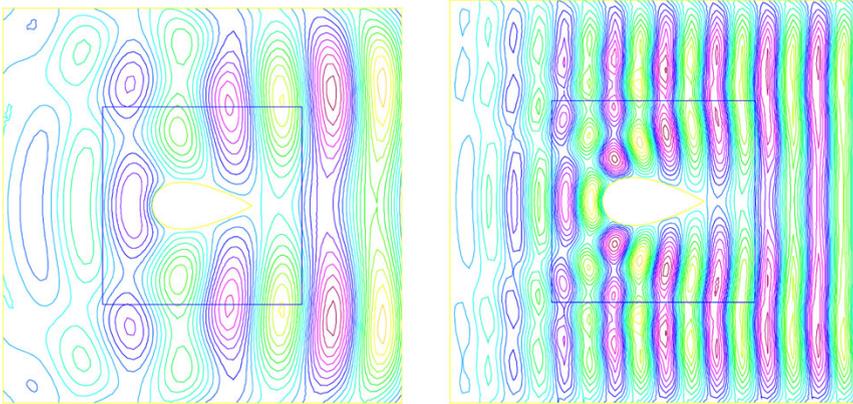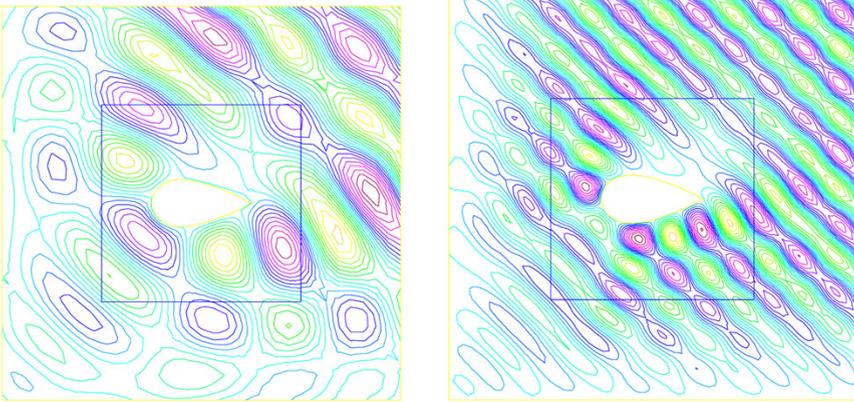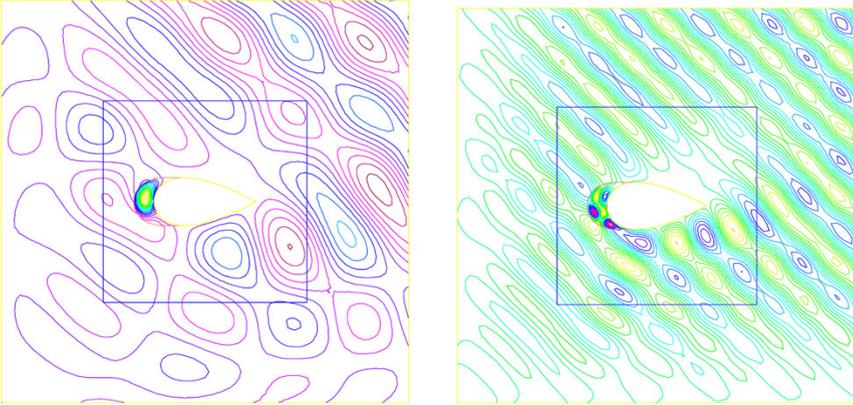
Fig. 11. Contour plot of the real part of the solution for $L = 0.5$ (left) and $L = 0.25$ (right). Incident wave coming from the left lower corner with a 45 degrees angle.

References

[BDG+97] M. O. Bristeau, E. J. Dean, R. Glowinski, V. Kwok, and J. Périaux. Exact controllability and domain decomposition methods with nonmatching grids for the computation of scattering waves. In R. Glowinski, J. Périaux, and Z. Shi, editors, *Domain Decomposition Methods in Sciences and Engineering*, pages 291–307. John Wiley & Sons, 1997.

[DL92]   R. Dautray and J.-L. Lions. *Mathematical Analysis and Numerical Methods for Science and Technology*, volume 5. Springer-Verlag, 1992.

[GL89]   R. Glowinski and P. LeTallec. *Augmented Lagrangian and Operator Splitting Methods in Nonlinear Mechanics*. SIAM, Philadelphia, PA, 1989.

[Glo03]   R. Glowinski. Finite element methods for incompressible viscous flow. In P. G. Ciarlet and J.-L. Lions, editors, *Handbook of Numerical Analysis, Vol. IX*, pages 3–1176. North-Holland, Amsterdam, 2003.

Domain Decomposition and Electronic Structure Computations: A Promising Approach

Guy Bencteux[1,4], Maxime Barrault[1], Eric Cancès[2,4], William W. Hager[3], and Claude Le Bris[2,4]

[1] EDF R&D, 1 avenue du Général de Gaulle, 92141 Clamart Cedex, France
    `{guy.bencteux,maxime.barrault}@edf.fr`
[2] CERMICS, École Nationale des Ponts et Chaussées, 6 & 8, avenue Blaise Pascal, Cité Descartes, 77455 Marne-La-Vallée Cedex 2, France,
    `{cances,lebris}@cermics.enpc.fr`
[3] Department of Mathematics, University of Florida, Gainesville, FL 32611-8105, USA, `hager@math.ufl.edu`
[4] INRIA Rocquencourt, MICMAC project, Domaine de Voluceau, B.P. 105, 78153 Le Chesnay Cedex, France

Summary. We describe a domain decomposition approach applied to the specific context of electronic structure calculations. The approach has been introduced in [BCHL07]. We survey here the computational context, and explain the peculiarities of the approach as compared to problems of seemingly the same type in other engineering sciences. Improvements of the original approach presented in [BCHL07], including algorithmic refinements and effective parallel implementation, are included here. Test cases supporting the interest of the method are also reported.

It is our pleasure and an honor to dedicate this contribution to Olivier Pironneau, on the occasion of his sixtieth birthday. With admiration, respect and friendship.

1 Introduction and Motivation

1.1 General Context

Numerical simulation is nowadays an ubiquitous tool in materials science, chemistry and biology. Design of new materials, irradiation induced damage, drug design, protein folding are instances of applications of numerical simulation. For convenience we now briefly present the context of the specific computational problem under consideration in the present article. A more detailed, mathematically-oriented, presentation is the purpose of the monograph [CDK$^+$03] or of the review article [LeB05].

For many problems of major interest, empirical models where atoms are represented as point particles interacting with a parameterized force-field are

adequate models. On the other hand, when electronic structure plays a role in the phenomenon under consideration, an explicit quantum modelling of the electronic wavefunctions is required. For this purpose, two levels of approximation are possible.

The first category is the category of *ab initio* models, which are general purpose models that aim at solving sophisticated approximations of the Schrödinger equation. Such models only require the knowledge of universal constants and require a, ideally null but practically limited, number of adjustable parameters. The most commonly used models in this category are Density Functional Theory (DFT) based models and Hartree–Fock type models, respectively. Although these two families of models have different theoretical grounding, they share the same mathematical nature. They are *constrained minimization problems*, of the form

$$\inf \left\{ E(\psi_1, \ldots, \psi_N), \ \psi_i \in H^1(\mathbb{R}^3), \ \int_{\mathbb{R}^3} \psi_i \psi_j = \delta_{ij}, \ \forall 1 \le i, j \le N \right\} \quad (1)$$

The functions ψ_i are called the *molecular orbitals* of the system. The energy functional E, which of course depends on the model employed, is parametrized by the charges and positions of the nuclei of the system under consideration. With such models, systems with up to 10^4 electrons can be simulated.

Minimization problems of the type (1) are not approached by minimization algorithms, mainly because they are high-dimensional in nature. In contrast, the numerical scheme consists in solving their Euler–Lagrange equations, which are nonlinear eigenvalue problems. The current practice is to iterate on the nonlinearity using fixed-point type algorithms, called in this framework *Self Consistent Field* iterations, with reference to the mean-field nature of DFT and HF type models.

The second category of models is that of semi-empirical models, such as Extended Hückel Theory based and tight-binding models, which contain additional approximations of the above DFT or HF type models. They consist in solving linear eigenvalue problems. State-of-the-art simulations using such models address systems with up to 10^5–10^6 electrons.

Finite-difference schemes may be used to discretize the above problems. They have proved successful in some very specific niches, most of them related to solid-state science. However, in an overwhelming number of contexts, the discretization of the nonlinear or linear eigenvalue problems introduced above is performed using a Galerkin formulation. The molecular orbitals ψ_i are developed on a Galerkin basis $\{\chi_i\}_{1 \le i \le N_b}$, with size $N_b > N$, the number of electrons in the system. Basis functions may be plane waves. This is often the case for solid state science applications and then N_b is very large as compared to N, typically one hundred times as large or more. They may also be *localized* functions, namely compactly supported functions or exponentially decreasing functions. Such basis sets correspond to the so-called *Linear Combination of Atomic Orbitals* (LCAO) approach. Then the dimension of the basis set needed to reach the extremely demanding accuracy required for

electronic calculation problems is surprisingly small. Such basis sets, typically in the spirit of spectral methods, or modal synthesis, are, indeed, remarkably efficient. The domain decomposition method described in the present article is restricted to the LCAO approach. Indeed, it strongly exploits the locality of the basis functions.

In both categories of models, linear or nonlinear, the elementary brick is the solution to a (generalized) linear eigenvalue problem of the following form:

$$
\begin{cases}
Hc_i = \varepsilon_i Sc_i, \qquad \varepsilon_1 \leq \ldots \leq \varepsilon_N \leq \varepsilon_{N+1} \leq \ldots \leq \varepsilon_{N_b}, \\
c_i^t Sc_j = \delta_{ij}, \\
D_\star = \displaystyle\sum_{i=1}^{N} c_i c_i^t.
\end{cases} \tag{2}
$$

The matrix H is a $N_b \times N_b$ symmetric matrix, called the *Fock matrix*. When the linear system above is one iteration of a nonlinear cycle, this matrix is computed from the result of the previous iteration. The matrix S is a $N_b \times N_b$ symmetric positive definite matrix, called the *overlap* matrix, which depends only on the basis set used (it corresponds to the mass matrix in the language of finite element methods).

One searches for the solution of (2), that is the matrix $D_\star$ called the *density matrix*. This formally requires the knowledge of the first N (generalized) eigenelements of the matrix H (in fact, we shall see below this statement is not exactly true).

The system of the equations (2) is generally viewed as a generalized eigenvalue problem, and most of the computational approaches consist in solving the system via the computation of each individual vector c_i (discretizing the wavefunction ψ_i of (1)), using a direct diagonalization procedure.

1.2 Specificities of the Approaches for Large Systems

The procedure mentioned above may be conveniently implemented for systems of limited size. For large systems, however, the solution procedure for the linear problem suffers from two computational bottlenecks. The first one is the need for assembling the Fock matrix. It *a priori* involves $O(N_b^3)$ operations in DFT models and $O(N_b^4)$ in HF models. Adequate approaches, which lower the complexity of this step, have been proposed. Fast multipole methods (see [SC00]) are one instance of such approaches. The second practical bottleneck is the diagonalization step itself. This is the focus of the present contribution. Because of the possibly prohibitive $O(N_b^3)$ cost of direct diagonalization procedures, the so-called *alternatives to diagonalization* have been introduced. The method introduced in the present contribution aims at competing with such methods, and eventually outperforming them. With a view to understanding the problem under consideration, let us briefly review some peculiarities of electronic structure calculation problems.

The situation critically depends on the type of basis set employed. With plane wave basis sets, the number N of eigenelements to determine can be considered as small, compared to the size N_b of the matrix H ($N_b \sim 100N$). Then, iterative diagonalization methods, based on the inverse power paradigm, are a natural choice. In contrast, in the case of localized basis sets we deal with in this article, N_b varies from 2 to 10 times N. In any case it remains strictly proportional to N. Hence, the problem (2) can be rephrased as follows: identify say one half of the eigenelements of a given matrix. This makes the problem very specific as compared to other linear eigenvalue problems encountered in other fields of the engineering sciences (see [AHLT05, HL07], for instance). The sparsity of the matrices in the present context is another peculiarity of the problem. Although the matrices H and S are sparse for large molecular systems, they are not as sparse as the stiffness and mass matrices usually encountered when using finite difference or finite element methods. For example, the bandwidth of H and S is of the order of 10^2 in the numerical examples reported in Section 5.

1.3 Alternative Methods Towards Linear Scaling

In addition to the above mentioned peculiarities, a crucial specificity of the problem (2) is that the eigenelements do not need to be explicitly identified. As expressed by the last line of (2), only the knowledge of the density matrix $D_\star$ is required, both for the evaluation of the Fock operator associated to the next iteration, in a nonlinear context, and for the evaluation of relevant output quantities, in the linear context or at the last step of the iteration loop.

From a geometrical viewpoint, $D_\star$ is the S-orthogonal projector (in the sense that $D_\star S D_\star = D_\star$ and $D_\star^t = D_\star$) on the vector subspace generated by the eigenvectors associated with the lowest N eigenvalues of the generalized eigenvalue problem $Hc = \varepsilon Sc$.

The above elementary remark is the bottom line for the development of the alternative to diagonalization methods, also often called *linear scaling* methods because their claimed purpose is to reach a linear complexity of the solution procedure (either in terms of N the number of electrons, or N_b the dimension of the basis set). For practical reasons, which will not be further developed here, such methods assume that:

(H1) The matrices H and S are sparse, in the sense that, for large systems, the number of non-zero coefficients scales as N. This assumption is not restrictive. In particular, it is automatically satisfied for DFT and HF models as soon as the basis functions are localized;

(H2) The matrix $D_\star$ built from the solution to (2) is also sparse. This condition seems to be fulfilled as soon as the relative gap

$$\gamma = \frac{\varepsilon_{N+1} - \varepsilon_N}{\varepsilon_{N_b} - \varepsilon_1}. \tag{3}$$

deduced from the solution of (2) is large enough. This observation can be supported by qualitative physical arguments [Koh96], but has seemingly no mathematical grounding to date (see, however, [Koh59]).

State-of-the-art surveys on such methods are [BMG02, Goe99]. One of the most commonly used linear scaling method is the Density Matrix Minimization (DMM) method [LNV93].

2 A New Domain Decomposition Approach

Our purpose is now to expose a method, based on the *domain decomposition* paradigm, which we have recently introduced in [BCHL07], and for which we also consider a setting where the above two assumptions are valid. Although still in its development, we have good hope that this approach will outperform existing ones in a near future. Preliminary test cases support this hope.

The approach described below is not the first occurrence of a method based on a "geographical" decomposition of the matrix H in the context of quantum chemistry (see, e.g., [YL95]). A significant methodological improvement is, however, fulfilled with the present method. To the best of our knowledge, existing methods in the context of electronic calculations that may be recast as domain decomposition methods only consist of local solvers complemented by a crude global step. Our method seems to be the first one really exhibiting the local/global paradigm in the spirit of methods used in other fields of the engineering sciences.

In the following, we expose and make use of the method on one-dimensional systems, typically nanotubes or linear hydrocarbons. Generalizations to three-dimensional systems do not really bring up new methodological issues. They are, however, much more difficult in terms of implementation.

For simplicity, we now present our method assuming that $S = I_{N_b}$, i.e. that the Galerkin basis $\{\chi_i\}_{1 \leq i \leq N_b}$ is orthonormal. The extension of the method to the case when $S \neq I_{N_b}$ is straightforward. The space $\mathcal{M}^{k,l}$ denotes the vector space of the $k \times l$ real matrices.

Let us first notice that a solution $D_\star$ of (2) reads

$$D_\star = C_\star C_\star^t \tag{4}$$

where $C_\star$ is a solution to the minimization problem

$$\inf \left\{ \mathrm{Tr}(HCC^t), \quad C \in \mathcal{M}^{N_b, N}(\mathbb{R}), \ C^t C = I_N \right\}. \tag{5}$$

Our approach consists in solving an *approximation* of the problem (5). The latter is obtained by minimizing the exact energy $\mathrm{Tr}(HCC^t)$ on the set of the matrices C that have the block structure displayed on Figure 1 and satisfy the constraint $C^t C = I_N$.
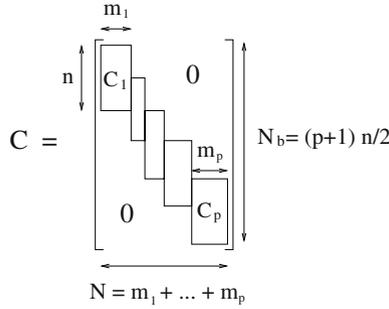
Fig. 1. Block structure of the matrices C.
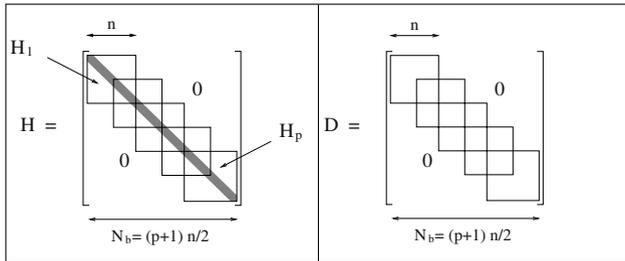
Fig. 2. Block structure of the matrices H and D.

A detailed justification of the choice of this structure is given in [BCHL07]. Let us only mention here that the decomposition is suggested from the localization of electrons and the use of a localized basis set. Note that each block overlaps only with its first neighbors. Again for simplicity, we expose the method in the case where overlapping is exactly $n/2$, but it could be any integer smaller than $n/2$.

The resulting minimization problem can be recast as

$$\inf \left\{ \sum_{i=1}^{p} \mathrm{Tr}\left(H_i C_i C_i^t\right),\ C_i \in \mathcal{M}^{n,m_i}(\mathbb{R}),\ m_i \in \mathbf{N},\ C_i^t C_i = I_{m_i}\ \forall\, 1 \le i \le p,\right.$$

$$\left. C_i^t T C_{i+1} = 0 \quad \forall\, 1 \le i \le p-1,\ \sum_{i=1}^{p} m_i = N \right\}. \quad (6)$$

In the above formula, $T \in \mathcal{M}^{n,n}(\mathbb{R})$ is the matrix defined by

$$T_{kl} = \begin{cases} 1 & \text{if } k - l = \frac{n}{2}, \\ 0 & \text{otherwise,} \end{cases} \quad (7)$$

$H_i \in \mathcal{M}^{n,n}(\mathbb{R})$ is a symmetric submatrix of H (see Figure 2), and

$$\mathrm{Tr}\left(\begin{bmatrix} H_1 & & \\ & \ddots & \\ & & H_p \end{bmatrix}\begin{bmatrix} C_1 & 0 \\ & \ddots \\ 0 & C_p \end{bmatrix}\begin{bmatrix} C_1 & 0 \\ & \ddots \\ 0 & C_p \end{bmatrix}^t\right) = \sum_{i=1}^{p} \mathrm{Tr}\left(\begin{bmatrix} H_i \end{bmatrix}\begin{bmatrix} C_i \end{bmatrix}\begin{bmatrix} C_i \end{bmatrix}^t\right)$$

$$\begin{bmatrix} C_1 & 0 \\ & \ddots \\ 0 & C_p \end{bmatrix}^t\begin{bmatrix} C_1 & 0 \\ & \ddots \\ 0 & C_p \end{bmatrix} = \begin{bmatrix} C_i^t C_i & \cdots & C_i^t T C_{i+1} \\ & \ddots & \\ 0 & & \end{bmatrix}.$$

In this way, we replace the $\frac{N(N+1)}{2}$ *global* scalar constraints $C^t C = I_N$ involving vectors of size N_b, by the $\sum_{i=1}^{p} \frac{m_i(m_i+1)}{2}$ *local* scalar constraints $C_i^t C_i = I_{m_i}$ and the $\sum_{i=1}^{p-1} m_i m_{i+1}$ *local* scalar constraints $C_i^t T C_{i+1} = 0$, involving vectors of size n. We would like to emphasize that we can only obtain in this way a basis of the vector space generated by the lowest N eigenvectors of H. This is the very nature of the method, which consequently cannot be applied for the search for the eigenvectors themselves.

Before we describe in details the procedure employed to solve the Euler–Lagrange equations of (6) in a greater generality, let us consider, for pedagogic purpose, the following oversimplified problem:

$$\inf\left\{\langle H_1 Z_1, Z_1\rangle + \langle H_2 Z_2, Z_2\rangle, \ Z_i \in \mathbb{R}^{N_b}, \ \langle Z_i, Z_i\rangle = 1, \ \langle Z_1, Z_2\rangle = 0\right\}. \quad (8)$$

We have denoted by $\langle\cdot,\cdot\rangle$ the standard Euclidean scalar product on $\mathbb{R}^{N_b}$.

The problem (8) is not strictly speaking a particular occurrence of (6), but it shows the same characteristics and technical difficulties: a separable functional is minimized, there are constraints on variables of each term and there is a cross constraint between the two terms.

The bottom line for our decomposition algorithm is to attack (8) as follows. Choose (Z_1^0, Z_2^0) satisfying the constraints and construct the sequence $(Z_1^k, Z_2^k)_{k\in\mathbf{N}}$ by the following iteration procedure. Assume (Z_1^k, Z_2^k) is known, then

Local step: Solve

$$\begin{cases} \widetilde{Z}_1^k = \arg\inf\left\{\langle H_1 Z_1, Z_1\rangle, \ Z_1 \in \mathbb{R}^{N_b}, \ \langle Z_1, Z_1\rangle = 1, \ \langle Z_1, Z_2^k\rangle = 0\right\}, \\ \widetilde{Z}_2^k = \arg\inf\left\{\langle H_2 Z_2, Z_2\rangle, \ Z_2 \in \mathbb{R}^{N_b}, \ \langle Z_2, Z_2\rangle = 1, \ \langle \widetilde{Z}_1^k, Z_2\rangle = 0\right\}; \end{cases} \quad (9)$$

Global step: Solve

$$\alpha^* = \arg\inf\left\{\langle H_1 Z_1(\alpha), Z_1(\alpha)\rangle + \langle H_2 Z_2(\alpha), Z_2(\alpha)\rangle, \ \alpha \in \mathbb{R}\right\} \quad (10)$$

where

$$Z_1(\alpha) = \frac{\widetilde{Z}_1^k + \alpha \widetilde{Z}_2^k}{\sqrt{1+\alpha^2}}, \quad Z_2(\alpha) = \frac{-\alpha \widetilde{Z}_1^k + \widetilde{Z}_2^k}{\sqrt{1+\alpha^2}}, \quad (11)$$

and set

$$Z_1^{k+1} = \frac{\widetilde{Z}_1^k + \alpha^* \widetilde{Z}_2^k}{\sqrt{1 + (\alpha^*)^2}}, \quad Z_2^{k+1} = \frac{-\alpha^* \widetilde{Z}_1^k + \widetilde{Z}_2^k}{\sqrt{1 + (\alpha^*)^2}}. \tag{12}$$

This algorithm operates at two levels: a fine level where two problems of dimension N_b are solved (rather than one problem of dimension $2N_b$); a coarse level where a problem of dimension 2 is solved.

The local step monotonically reduces the objective function; however, it may not converge to the global optimum. The technical problem is that the Lagrange multipliers associated with the constraint $\langle Z_1, Z_2 \rangle = 0$ may converge to different values in the two subproblems associated with the local step. The global step again reduces the value of the objective function since $\tilde{Z}_1^k$ and $\tilde{Z}_2^k$ are feasible in the global step. The combined algorithm (local step + global step), therefore, makes the objective function monotonically decrease. The simple case $H_1 = H_2$ is interesting to consider. First, if the algorithm is initialized with $Z_2^0 = 0$ in the first line of (9), it is easily seen that the local step is sufficient to converge to the global minimizer, in one single step. Second, it has been proved in [Bar05] that for a more general initial guess and under some assumption on the eigenvalues of the matrix H_1, this algorithm globally converges to an optimal solution of (8). Ongoing work [BCHL] aims at generalizing the above proof when the additional assumption on eigenvalues is omitted. The analysis of the convergence in the case $H_1 \neq H_2$ is a longer term goal.

3 The Multilevel Domain Decomposition (MDD) Algorithm

We define, for all p-tuple $(C_i)_{1 \leq i \leq p}$,

$$\mathcal{E}\left((C_i)_{1 \leq i \leq p}\right) = \sum_{i=1}^{p} \text{Tr}\left(H_i C_i C_i^t\right), \tag{13}$$

and set by convention

$$U_0 = U_p = 0. \tag{14}$$

It has been shown in [BCHL07] that updating the block sizes m_i along the iterations is crucial to make the domain decomposition algorithm converge toward a good approximation of the solution to (5). It is, however, observed in practice that after a few iterations, the block sizes have converged (they do not vary in the course of the following iterations). This is why, for the sake of clarity, we have chosen to present here a simplified version of the algorithm where block sizes are held constant along the iterations. For a description of the complete algorithm with variable block sizes, we refer to [BCHL07].

At iteration k, we have at hand a set of matrices $(C_i^k)_{1 \leq i \leq p}$ such that $C_i^k \in \mathcal{M}^{n,m_i}(\mathbb{R})$, $[C_i^k]^t C_i^k = I_{m_i}$, $[C_i^k]^t T C_{i+1}^k = 0$. We now explain how to compute the new iterate $(C_i^{k+1})_{1 \leq i \leq p}$.

Step 1: *Local fine solver.*
  (a) For each i, find

$$\inf \left\{ \mathrm{Tr}\left(H_i C_i C_i^t\right),\ C_i \in \mathcal{M}^{n,m_i}(\mathbb{R}),\ C_i^t C_i = I_{m_i}, \right.$$
$$\left. [C_{i-1}^k]^t T C_i = 0, \quad C_i^t T C_{i+1}^k = 0 \right\}. \quad (15)$$

This is done via diagonalization of the matrix H_i in the subspace

$$V_i^k = \left\{ x \in \mathbb{R}^n,\ \left[C_{i-1}^k\right]^t T x = 0,\ x^t T C_{i+1}^k = 0 \right\},$$

i.e. diagonalize $P_i^k H_i P_i^k$ where P_i^k is the orthogonal projector on V_i^k.
This provides (at least) $n - m_{i-1} - m_{i+1}$ real eigenvalues and associated orthonormal vectors $x_{i,j}^k$. The latter are T-orthogonal to the column vectors of C_{i-1}^k and C_{i+1}^k.
  (b) Collect the lowest m_i vectors $x_{i,j}^k$ in the $n \times m_i$ matrix $\widetilde{C}_i^k$.
Step 2: *Global coarse solver.* Solve

$$\mathcal{U}^* = \arg\inf \left\{ f(\mathcal{U}),\ \mathcal{U} = (U_i)_i,\ \forall 1 \le i \le p-1,\ U_i \in \mathcal{M}^{m_{i+1}, m_i}(\mathbb{R}) \right\}, \quad (16)$$

where

$$f(\mathcal{U}) = \mathcal{E}\left(\left(C_i(\mathcal{U}) \left(C_i(\mathcal{U})^t C_i(\mathcal{U}) \right)^{-\frac{1}{2}} \right)_i \right) \quad (17)$$

and

$$C_i(\mathcal{U}) = \widetilde{C}_i^k + T\widetilde{C}_{i+1}^k U_i \left([\widetilde{C}_i^k]^t T T^t \widetilde{C}_i^k \right) - T^t \widetilde{C}_{i-1}^k U_{i-1}^t \left([\widetilde{C}_i^k]^t T^t T \widetilde{C}_i^k \right). \quad (18)$$

Next set, for all $1 \le i \le p$,

$$C_i^{k+1} = C_i(\mathcal{U}^*) \left(C_i(\mathcal{U}^*)^t C_i(\mathcal{U}^*) \right)^{-1/2}. \quad (19)$$

Notice that in Step 1, the computations of each odd block is independent from the other odd blocks, and obviously the same for even blocks. Thus, we use here a red/black strategy.

In the global step, we perturb each variable by a linear combination of the adjacent variables. The matrices $\mathcal{U} = (U_i)_i$ in (16) play the same role as the real parameter α in the toy example, the equation (10). The perturbation is designed so that the constraints are satisfied. However, our numerical experiments show that this is not exactly the case, in the sense that, for some i, $[C_i^{k+1}]^t T [C_{i+1}^{k+1}]$ may present coefficients as large as about 10^{-3}. All linear scaling algorithms have difficulties in ensuring this constraint. We should mention here that in our case, the resulting deviation of $C^t C$ from identity is small, $C^t C$ being in any case block tridiagonal.

In practice, we reduce the computational cost of the global step, by again using a domain decomposition method. The blocks $(C_i)_{1 \le i \le p}$ are collected
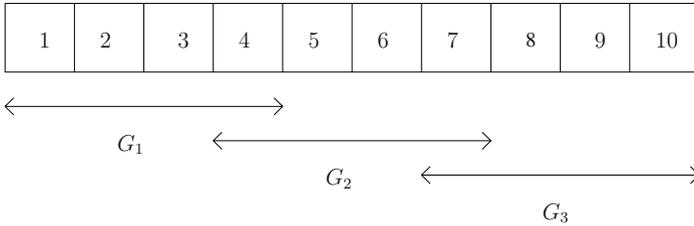
Fig. 3. Collection of $p = 10$ blocks into $r = 3$ groups.

Repeat until convergence:

1a. Local step on blocks: $1, 3, ..., (2i+1), ...$
1b. Local step on blocks: $2, 4, ..., (2i), ...$
2a. Global step on groups: $\{1, 2\}, \{3, 4\}, ..., \{2i-1, 2i\}, ...$
2b. Global step on groups: $\{2, 3\}, \{4, 5\}, ..., \{2i, 2i+1\}, ...$

Fig. 4. Schematic view of the algorithm in the case of 2-block groups ($r = 2$): tasks appearing on the same line are independent from one another. Order between the steps 1a and 1b is reversed from one iteration to the other. The same holds for the steps 2a and 2b.

in r overlapping groups $(G_l)_{1 \le l \le r}$ as shown in Figure 3. As each group only overlaps with its first neighbors, the problem (16) can be solved first for the groups (G_{2l+1}), next for the groups (G_{2l}). We have observed that the number of iterations of the outer loop (local step + global step) does not significantly increase when the 'exact' global step (16) is replaced by the approximate global step consisting in optimizing first the odd groups, then the even groups. The numerical results performed so far (see Section 5) tend to show that the resulting algorithm scales linearly with the system size.

A schematic view of the algorithm is provided in Figure 4.

One important point (not taken into account in [BCHL07]) is that the Hessian of f enjoys a very specific structure. It is a sum of tensor products of square matrices of size m_i. For example, with two-block groups ($r = 2$), we have

$$HU = \sum_{i=1}^{4} A^{(i)} U B^{(i)} \tag{20}$$

with $A^{(i)} \in \mathcal{M}^{m_2, m_2}(\mathbb{R})$ and $B^{(i)} \in \mathcal{M}^{m_1, m_1}(\mathbb{R})$. Consequently, it is possible to compute Hessian-vector products, without assembling the Hessian, in $\mathcal{O}(m_1 \, m_2 \, \max(m_1, m_2))$ elementary operations, instead of $\mathcal{O}(m_1^2 \, m_2^2)$ with a naive implementation. An additional source of acceleration is the fact that this formulation uses only matrix-matrix products. Efficient implementations of matrix-matrix products, taking advantage of higher numbers of floating point operations per memory access, are available in the BLAS 3 library

(see, for instance, [PA04]). This makes Newton-like methods affordable: a good estimation of the Newton direction can be easily computed using an iterative method.

In the current version of our domain decomposition algorithm, the global step is solved approximatively by a single iteration of the Newton algorithm with initial guess $U_i = 0$, the Newton iteration being computed iteratively by means of the SYMMLQ algorithm [PS75]. In a next future, we plan to test the efficiency of advanced first order methods such as the one described in [HZ05]. No definite conclusions about the comparative efficiencies of the various numerical methods for performing the global step can be drawn yet.

4 Parallel Implementation

For parallel implementation, the single-program, multi-data (SPMD) model is used, with message passing techniques using the MPI library, which allows to maintain only one version of the code.

Each processor executes a single instance of the algorithm presented in Section 3 applied to a contiguous subset of blocks. Compared to the sequential version, additional data structures are introduced: each processor needs to access the matrices C_i and H_i corresponding to the last block of the processor located on its left and to the first block of the processor located on its right, as shown in Figure 5. These frontier blocks play the role of ghost nodes in classical domain decomposition without overlapping. For this reason, we will sometimes call them the ghost blocks.

The major part of the communications is performed between neighboring processors at the end of each step of the algorithm (i.e. of each line in the scheme displayed in Figure 4), in order to update the ghost blocks. This occurs only four times per iteration and, as we will see in the next section, the sizes of the exchanged messages are moderate.

Collective communications are needed to compute the current value of the function f appearing in the formula (17) and to check that the maximum deviation from orthogonality remains acceptable. They are also needed to sort the eigenvalues of the different blocks in the local step, in the complete version
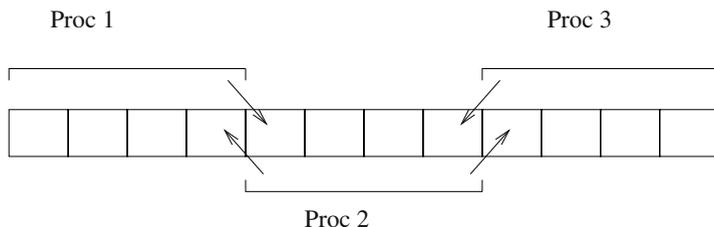

Fig. 5. Distribution of blocks over 3 processors. Arrows indicate the supplementary blocks a processor needs to access.

of the algorithm, allowing variable block sizes (see [BCHL07]). The important point is that the amount of data involved in the collective communications is small as well.

With this implementation we can use up to $nbloc/2$ processors. In order to efficiently use a larger number of processors, sublevels of parallelism should be introduced. For instance, each subproblem (15) (for a given i) can itself be parallelized.

Apart from the very small part of collective communications, the communication volume associated with each single processor remains constant irrespective of the number of blocks per processor and of the number of processors. We can thus expect a very good scalability, except for the situations when load balancing is strongly heterogeneous.

The implementation of the MDD algorithm described above can be easily extended to cover the case of $2D$ and $3D$ molecular systems.

5 Numerical Tests

This section is devoted to the presentation of the performance of the Multilevel Domain Decomposition (MDD) algorithm on matrices actually arising in real-world applications of electronic structure calculations. The benchmark matrices are of the same type of those used in the reference paper [BCHL07].

In the first subsection, we briefly recall how these matrices are generated and we provide some practical details on our implementation of the MDD algorithm. The computational performances obtained on sequential and parallel architectures, including comparisons with the density matrix minimization (DMM) method and with direct diagonalization using LAPACK, are discussed in the second and third subsections, respectively.

5.1 General Presentation

Three families of matrices corresponding to the Hartree–Fock ground state of some polymeric molecules are considered:

- Matrices of type $\mathcal{P}_1$ and $\mathcal{P}_2$ are related to COH-$(CO)_{n_m}$-COH polymeric chains, with interatomic Carbon-Carbon distances equal to 5 and 4 atomic units (a.u.), respectively;
- Matrices of type $\mathcal{P}_1$ are obtained with polyethylen molecules (CH_3-$(CH_2)_{n_m}$-CH_3) with physically relevant Carbon-Carbon distances.

The geometry of the very long molecules is guessed from the optimal distances obtained by geometry optimization (with constraints for $\mathcal{P}_1$ and $\mathcal{P}_2$) on moderate size molecules (about 60 Carbon atoms) and minimal basis sets. All these off-line calculations are performed using the GAUSSIAN package ([FTS+98]). It is then observed that the overlap matrix and Fock matrix obtained exhibit a periodic structure in their bulk. Overlap and Fock matrices for large size

Table 1. Localization parameters, block sizes and asymptotic gaps for the test cases.

	$\mathcal{P}_1$	$\mathcal{P}_2$	$\mathcal{P}_3$
Bandwidth of S	59	79	111
Bandwidth of H	99	159	255
n	130	200	308
q	50	80	126
Asymptotic gap (a.u.)	1.04×10^{-3}	3.57×10^{-3}	2.81×10^{-2}

molecules can then be constructed using this periodicity property. For n_m sufficiently large, bulk periodicity is also observed in the density matrix. This property is used to generate reference solutions for large molecules.

Table 1 gives a synthetic view of the different structure properties of the three families of matrices under examination. The integer q stands for the overlap between two adjacent blocks (note that one could have taken $n = 2q$ if the overlap matrix S was equal to identity, but that one has to take $n > 2q$ in our case since $S \neq I$).

Initial guess generation is of crucial importance for any linear scaling method. The procedure in use here is in the spirit of the domain decomposition method:

1. A first guess of the block sizes is obtained by locating Z electrons around each nucleus of charge Z;
2. A set of blocks C_i is built from the lowest m_i (generalized) eigenvectors associated with the block matrices H_i and S_i (the block matrices H_i are introduced in Section 2; the block matrices S_i are defined accordingly);
3. These blocks are eventually optimized with the local fine solver of the MDD algorithm, including block size update (electron transfer).

Criteria for comparing the results

The quality of the results produced by the MDD and DMM methods is evaluated by computing two criteria. The first criterion is the relative energy difference $e_E = \frac{|E - E_0|}{|E_0|}$ between the energy E of the current iterate D and the energy E_0 of the reference density matrix $D_\star$. The second criterion is the semi-norm

$$e_\infty = \sup_{(i,j)\,\text{s.t.}\,|H_{ij}| \geq \varepsilon} \left| D_{ij} - [D_\star]_{ij} \right| \tag{21}$$

with $\varepsilon = 10^{-10}$. The introduction of the semi-norm (21) is consistent with the cut-off on the entries of H (thus the value chosen for ε). Indeed, in most cases, the matrix D is only used for the calculations of various observables (in particular the electronic energy and the Hellman–Feynman forces), all of them of the form $\text{Tr}(AD)$, where the matrix A shares the same pattern as H (see [CDK+03] for details). The final result of the calculation is, therefore,

insensitive to entries D_{ij} with indices (i, j) such that $|H_{ij}|$ is below some cut-off value.

In all the calculations presented below, the global step is performed with groups consisting of two blocks ($r = 2$), and the algorithm is, therefore, exactly that displayed in Figure 4.

5.2 Sequential Computations

The numerical results presented in this section have been obtained with a single 2.8 GHz Xeon processor.

Density matrices have been computed for a series of matrices H and S of types $\mathcal{P}_1$, $\mathcal{P}_2$, and $\mathcal{P}_3$, using (1) the MDD algorithm, (2) a diagonalization procedure (the *dsbgv.f* routine from the LAPACK library), and (3) the DMM method [LNV93]. The latter method belongs to the class of linear scaling algorithms. An important feature of the DMM method is that linear scaling is achieved through cut-offs on the matrix entries. We have chosen here a cut-off strategy based on *a priori* defined patterns, that may be suboptimal. Our implementation of DMM converges to a fairly good approximation of the exact density matrix and scales linearly, but the prefactor might possibly be improved by more refined cut-off strategies.

A detailed presentation of the comparison between the three methods is provided in [BCHL07]. Our new approach for computing the Newton direction in the global step (see Section 3) further improves the efficiency of MDD: with the new implementation of MDD, and with respect to the former implementation reported on in [BCHL07], CPU time is divided by 2 for $\mathcal{P}_1$ type molecules, by 5 for $\mathcal{P}_2$, and by 10 for $\mathcal{P}_3$, and the memory required is now lower for MDD than for DMM. These results are shown for $\mathcal{P}_2$ in Figures 6 and 7. They clearly demonstrate that the MDD algorithm scales linearly with respect to the parameter n_m (in both CPU time and memory occupancy).

Let us also notice that for $\mathcal{P}_2$, the crossover point between diagonalization and MDD (as far as CPU time is concerned) is now shifted to less than 2,000 basis functions.

5.3 Parallel Computations

We conclude with some tests of our parallel implementation of the MDD algorithm described in Section 4. These tests have been performed on a 8 node Linux cluster in dedicated mode, consisting of 8 biprocessors DELL Precision 450 (Intel(R) Xeon(TM) CPU 2.40GHz), with Gigabit Ethernet connections. They concern the polyethylene family $\mathcal{P}_3$, for which the size of each ghost block is about 150 Ko.

We only test here the highest level of parallelism of the MDD algorithm, consistently with the relatively low number of processors that have been used in this first study. We plan to test multilevel parallelism in a near future.
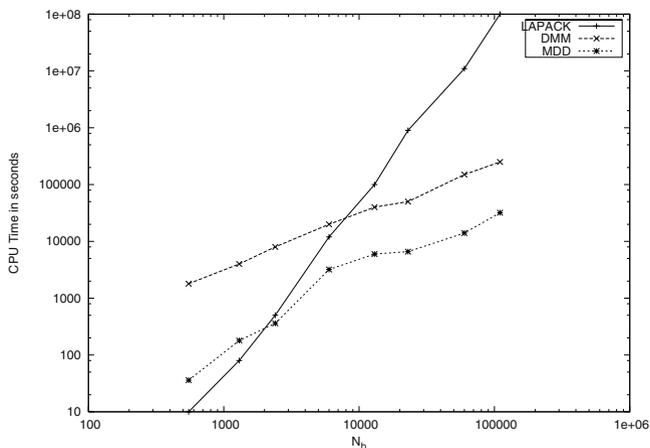
Fig. 6. Requested CPU time for computing the density matrix of a molecule of type $\mathcal{P}_2$ as a function of the number of basis functions.
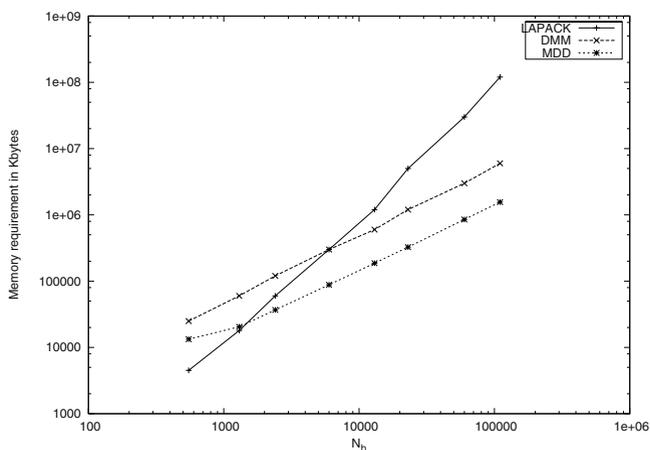
Fig. 7. Requested memory for computing the density matrix of a molecule of type $\mathcal{P}_2$ as a function of the number of basis functions.

In particular, the local step in each block, as well as the global step in each group, will be parallelized.

Tables 2 and 3 report on the speedup (ratio between the wall clock time with one processor and the wall clock time for several processors) and efficiency (ratio between the speedup and the number of processors) of our parallel MDD algorithm.

The scalability, namely the variation of the wall clock time when the number of processors and the size of the matrix proportionally grow, is reported in Table 4, for a molecule of type $\mathcal{P}_3$.

Table 2. Wall clock time as a function of the number of processors for a molecule of type $\mathcal{P}_3$, with $n_m = 3300$ (128 blocks). 8 MDD iterations are necessary to achieve convergence up to 5×10^{-8} in energy and 3×10^{-3} in the density matrix (for the semi-norm (21)).

Number of processors	1	2	4	8	16
Wall clock time (s)	4300	2400	1200	580	360
Speedup		1.8	3.6	7.4	12
Efficiency		0.9	0.9	0.9	0.75

Table 3. Wall clock time as a function of the number of processors for a molecule of type $\mathcal{P}_3$, with $n_m = 13300$ (512 blocks). 7 MDD iterations are necessary to achieve convergence up to 5×10^{-8} in energy and 3×10^{-3} in the density matrix (for the semi-norm (21)).

Number of processors	1	4	8	16
Wall clock time (s)	18460	4820	2520	1275
Speedup		3.8	7.3	14.5
Efficiency		0.96	0.92	0.91

Table 4. Scalability of the MDD algorithm for a molecule of type $\mathcal{P}_3$. The convergence thresholds are 2.5×10^{-7} in energy and 4×10^{-3} in density matrix (for the semi-norm (21)).

Number of processors	1	4	8	16
Wall clock time with 200 atoms (8 blocks) per processor (s)	167	206	222	253
Wall clock time with 800 atoms (32 blocks) per processor (s)	1249	1237	1257	1250

Note that the calculations reported in this article have been performed with minimal basis sets. It is the subject of ongoing works to test the efficiency of the MDD algorithm for larger basis sets.

Let us finally mention that our parallel implementation of the MDD algorithm allows to solve (2) for a polyethylene molecule with $106\,530$ atoms ($372\,862$ basis functions) on 16 processors, in 90 minutes.

6 Conclusion and Perspectives

In its current implementation, the MDD algorithm allows to solve efficiently the linear subproblem for linear molecules (polymers or nanotubes). The following issues will be addressed in a near future:

- Still in the case of 1D systems, we will allow blocks to have more than two neighbors. This should increase the flexibility and efficiency of the MDD

algorithm. For instance, this should render calculations with large basis sets including diffuse atomic orbitals affordable.

- We plan to implement the MDD algorithm in the framework of 2D and 3D molecular systems. Note that even with minimal overlap a given block has typically 8 neighbors in 2D and 26 neighbors in 3D.
- The MDD algorithm will be extended to the cases of the nonlinear Hartree–Fock and Kohn-Sham problems.
- The present version of the MDD algorithm is restricted to insulators (i.e. to matrices H with a sufficiently large gap). The possibility of extending the MDD methodology to cover the case of metallic systems is a challenging issue that will be studied.

References

[AHLT05] P. Arbenz, U. L. Hetmaniuk, R. B. Lehoucq, and R. S. Tuminaro. A comparison of eigensolvers for large-scale 3D modal analysis using AMG-preconditioned iterative methods. *Internat. J. Numer. Methods Engrg.*, 64:204–236, 2005.

[Bar05] M. Barrault. *Développement de méthodes rapides pour le calcul de structures électroniques.* Thèse, l'Ecole Nationale des Ponts et Chaussées, 2005.

[BCHL]   G. Bencteux, E. Cancès, W. W. Hager, and C. Le Bris. Work in progress.

[BCHL07] M. Barrault, E. Cancès, W. W. Hager, and C. Le Bris. Multilevel domain decomposition for electronic structure calculations. *J. Comput. Phys.*, 222(1):86–109, 2007.

[BMG02] D. Bowler, T. Miyazaki, and M. Gillan. Recent progress in linear scaling ab initio electronic structure theories. *J. Phys. Condens. Matter*, 14:2781–2798, 2002.

[CDK+03] E. Cancès, M. Defranceschi, W. Kutzelnigg, C. Le Bris, and Y. Maday. Computational quantum chemistry: a primer. In C. Le Bris, editor, *Handbook of Numerical Analysis, Special volume, Computational Chemistry, Vol. X*, pages 3–270. North-Holland, 2003.

[FTS+98] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, V. G. Zakrzewski, J. A. Montgomery, R. E. Stratmann, J. C. Burant, S. Dapprich, J. M. Millam, A. D. Daniels, K. N. Kudin, M. C. Strain, O. Farkas, J. Tomasi, V. Barone, M. Cossi, R. Cammi, B. Mennucci, C. Pomelli, C. Adamo, S. Clifford, J. Ochterski, G. A. Petersson, P. Y. Ayala, Q. Cui, K. Morokuma, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. Cioslowski, J. V. Ortiz, B. B. Stefanov, G. liu, A. Liashenko, P. Piskorz, I. Kpmaromi, G. Gomperts, R. L.

Martin, D. J. Fox, T. Keith, M. A. Al-Laham, C. Y. Peng, A. Nanayakkara, C. Gonzalez, M. Challacombe, P. M. W. Gill, B. G. Johnson, W. Chen, M. W. Wong, J. L. Andres, M. Head-Gordon, E. S. Replogle, and J. A. Pople. *Gaussian 98 (Revision A.7)*. Gaussian Inc., Pittsburgh, PA, 1998.

[Goe99]   S. Goedecker. Linear scaling electronic structure methods. *Rev. Mod. Phys.*, 71:1085–1123, 1999.

[HL07]   U. L. Hetmaniuk and R. B. Lehoucq. Multilevel methods for eigenspace computations in structural dynamics. In *Domain Decomposition Methods in Science and Engineering XVI*, volume 55 of *Lect. Notes Comput. Sci. Eng.*, pages 103–113, Springer, Berlin, 2007.

[HZ05]   W. Hager and H. Zhang. A new conjugate gradient method with guaranteed descent and an efficient line search. *SIAM J. Optim.*, 16:170–192, 2005.

[Koh59]   W. Kohn. Analytic properties of Bloch waves and Wannier functions. *Phys. Rev.*, 115:809–821, 1959.

[Koh96]   W. Kohn. Density functional and density matrix method scaling linearly with the number of atoms. *Phys. Rev. Lett.*, 76:3168–3171, 1996.

[LeB05]   C. Le Bris. Computational chemistry from the perspective of numerical analysis. In *Acta Numerica, Volume 14*, pages 363–444. 2005.

[LNV93]   X.-P. Li, R. W. Nunes, and D. Vanderbilt. Density-matrix electronic structure method with linear system size scaling. *Phys. Rev. B*, 47:10891–10894, 1993.

[PA04]   W. P. Petersen and P. Arbenz. *Introduction to Parallel Computing*. Oxford University Press, 2004.

[PS75]   C. Paige and M. Saunders. Solution of sparse indefinite systems of linear equations. *SIAM J. Numer. Anal.*, 12:617–629, 1975.

[SC00]   E. Schwegler and M. Challacombe. Linear scaling computation of the Fock matrix. *Theor. Chem. Acc.*, 104:344–349, 2000.

[YL95]   W. Yang and T. Lee. A density-matrix divide-and-conquer approach for electronic structure calculations of large molecules. *J. Chem. Phys.*, 163:5674, 1995.

Part IV

Free Surface, Moving Boundaries and Spectral Geometry Problems

Numerical Analysis of a Finite Element/Volume Penalty Method

Bertrand Maury

Laboratoire de Mathématiques, Université Paris-Sud, FR-91405 Orsay Cedex, France `Bertrand.Maury@math.u-psud.fr`

Summary. The penalty method makes it possible to incorporate a large class of constraints in general purpose Finite Element solvers like freeFEM++. We present here some contributions to the numerical analysis of this method. We propose an abstract framework for this approach, together with some general error estimates based on the discretization parameter ε and the space discretization parameter h. As this work is motivated by the possibility to handle constraints like rigid motion for fluid-particle flows, we shall pay a special attention to a model problem of this kind, where the constraint is prescribed over a subdomain. We show how the abstract estimate can be applied to this situation, in the case where a non-body-fitted mesh is used. In addition, we describe how this method provides an approximation of the Lagrange multiplier associated to the constraint.

1 Introduction

Because of its conceptual simplicity and the fact that it is usually straightforward to implement, the penalty method has been widely used to incorporate constraints in numerical optimization. The general principle can been seen as a relaxed version of the following fact: given a proper functional J over a set X, and K a subset of X, minimizing J over K is equivalent to minimizing $J_K = J + I_K$ over X, where I_K is the indicatrix of K:

$$I_K(x) = \begin{cases} 0 & \text{if } x \in K \\ +\infty & \text{if } x \notin K \end{cases}$$

Assume now that K can be defined as $K = \{x \in X \mid \Psi(x) = 0\}$, where Ψ is a non-negative function, the penalty method consists in considering relaxed functionals J_ε defined as

$$J_\varepsilon = J + \frac{1}{\varepsilon}\Psi, \quad \varepsilon > 0.$$

By definition of K, the function Ψ/ε approaches I_K point-wise:

$$\frac{1}{\varepsilon}\Psi(x) \longrightarrow I_K(x) \quad \text{as } \varepsilon \text{ goes to } 0, \ \forall x \in X.$$

If J_ε admits a minimum u^ε, for any ε, one can expect u^ε to approach a (or *the*) minimum of J over K, if it exists.

In actual Finite Element computations, some u_h^ε is computed as the solution to a finite dimensional problem, where h is a space-discretization parameter. The present work is motivated by the fact that, even if the penalty method for the continuous problem is convergent and the discretization procedure is sound, the rate of convergence of u_h^ε toward the exact solution is not straightforward to obtain.

To our knowledge, the first paper dedicated to the analysis of the penalty method in the Finite Element context dates back to 1973 (see [Bab73]), where this method was used to incorporate Dirichlet boundary conditions in some variants of the Finite Element Method. Since then, this strategy has been followed to integrate obstacles in fluid flow simulations [ABF99], to model the rigidity constraint [JLM05].

The present work is motivated by the handling of rigid particles in a fluid flow. Various approaches have been proposed to incorporate rigid bodies in a Stokes or Navier–Stokes fluid: arbitrary Lagrangian Eulerian approach [JT96, Mau99], fictitious domain approach [PG02]. More recently, a strategy based on augmented Lagrangian principles was proposed to handle a large class of multimaterial flows [VCLR04, RPVC05]. In [JLM05], we tested the raw penalty method to handle the rigidity constraint in a viscous fluid. This approach is not sophisticated: it simply consists in adding to the variational formulation the term

$$\frac{1}{\varepsilon}\int_\Omega \left(\nabla\mathbf{u} + \nabla^T\mathbf{u}\right) : \left(\nabla\mathbf{v} + \nabla^T\mathbf{v}\right).$$

It presents some drawbacks: as it is based on pure penalty and not augmented Lagrangian, the penalty parameter has to be taken very small for the constraint to be fulfilled properly, which may harm the conditioning of the system to solve. Yet, it shows itself to be robust in practice, it allows the use of non-boundary-fitted (e.g., Cartesian) meshes. Besides, it is straightforward to implement, so that a full Navier–Stokes solver (in 2D) with circular rigid particles can be written in about 50 lines, by using, for example, FreeFem++ (created by O. Pironneau, see [FFp]). Note that new tools for 3D problems are already available (see, e.g., [ff3, DPP03] or [lif]), which enable to perform computations of three dimensional fluid-particle flows.

We shall actually focus here on a simpler problem (see the problem (8)), which is a scalar version of the rigidity constraint. The fluid velocity is indeed replaced by a temperature field, and the rigid particle is replaced by a zone with infinite conductivity. The Lagrange multiplier can be interpreted in this context as a heat source term (see Remark 6).

We begin by presenting some standard properties of the penalty method for quadratic optimization (Section 2.1), and some convergence results. Then

we present the model problem, describe how we penalize and discretize it, and we show how the abstract framework applies to this situation. We finish by presenting an error estimate for the primal and dual components of the solutions, in terms of the quantities ε (the penalty parameter) and h (the mesh step size), whose proof is postponed to another paper.

2 Preliminaries, Abstract Framework

2.1 Continuous Problem

We recall here some standard properties concerning the penalty method applied to infinite dimensional problems. Most of those properties are established in [BF91], with a slightly different formalism. We shall consider the following set of assumptions:

$$
\left.
\begin{aligned}
& V \text{ is a Hilbert space, } \varphi \in V', \\
& a(\cdot,\cdot) \text{ bilinear, symmetric, continuous, elliptic } (a(v,v) \geq \alpha \left|v\right|^2), \\
& b(\cdot,\cdot) \text{ bilinear, symmetric, continuous, non-negative,} \\
& K = \{u \in V \mid b(u,u) = 0\} = \ker b, \\
& J(v) = \frac{1}{2}a(v,v) - \langle\varphi,v\rangle, \quad u = \arg\min_K J, \\
& J_\varepsilon(v) = \frac{1}{2}a(v,v) + \frac{1}{2\varepsilon}b(v,v) - \langle\varphi,v\rangle, \quad u^\varepsilon = \arg\min_V J_\varepsilon.
\end{aligned}
\right\}
\tag{1}
$$

Proposition 1. *Under the assumptions (1), the solution u^ε to the penalized problem converges to u.*

Proof. We write the variational formulation for the penalized problem:

$$
a(u^\varepsilon, v) + \frac{1}{\varepsilon}b(u^\varepsilon, v) = \langle\varphi, v\rangle \quad \forall v \in V.
\tag{2}
$$

Taking $v = u^\varepsilon$, we get

$$
\alpha \left|u^\varepsilon\right|^2 \leq a(u^\varepsilon, u^\varepsilon) \leq \|\varphi\| \left|u^\varepsilon\right|
$$

so that $\left|u\right|^\varepsilon$ is bounded. We extract a subsequence, still denoted by (u^ε), which converges weakly to some $z \in V$. As $J_\varepsilon \geq J$ and $b(u,u) = 0$, we have

$$
J(u^\varepsilon) \leq J_\varepsilon(u^\varepsilon) \leq J_\varepsilon(u) = J(u) \quad \forall \varepsilon > 0,
\tag{3}
$$

so that (J is convex and continuous) $J(z) \leq \liminf J(u^\varepsilon) \leq J(u)$. As

$$
J(u^\varepsilon) + \frac{1}{2\varepsilon}b(u^\varepsilon, u^\varepsilon) \leq J(u),
$$

$b(u^\varepsilon, u^\varepsilon)/\varepsilon$ is bounded, so that $b(u^\varepsilon, u^\varepsilon)$ goes to 0 with ε. Consequently, it holds $0 \le b(z,z) \le \liminf b(u^\varepsilon, u^\varepsilon) = 0$, which implies $z \in K$, so that $z = u$.

To establish the strong character of the convergence, we show that u^ε converges toward u for the norm associated to $a(\cdot, \cdot)$, which is equivalent to the original norm. As u^ε converges weakly to u for this scalar product $(a(u^\varepsilon, v) \to a(u, v)$ for any $v \in V)$, it is sufficient to establish the convergence of $|u^\varepsilon|_a = a(u^\varepsilon, u^\varepsilon)^{1/2}$ towards $|u|_a$. Firstly $|u|_a \le \liminf |u^\varepsilon|_a$, and the other inequality comes from (3):

$$\frac{1}{2}a(u^\varepsilon, u^\varepsilon) - \langle \varphi, u^\varepsilon \rangle \le \frac{1}{2}a(u, u) - \langle \varphi, u \rangle,$$

so that $\limsup |u^\varepsilon|_a \le |u|_a$.

The proposition does not say anything about the rate of convergence, and it can be very poor, as the following example illustrates.

Example 1. Consider $I =]0, 1[$, $V = H^1(I)$, and the problem which consists in minimizing the functional

$$J(v) = \frac{1}{2}\int_I |u'|^2,$$

over $K = \{v \in V \mid v(x) = 0 \text{ a.e. in } \mathcal{O} =]0, 1/2[\}$. The solution to that problem is obviously $u = \max\{0, 2(x - 1/2)\}$. Now let us denote by u^ε the minimum of the penalized functional

$$J_\varepsilon = \frac{1}{2}\int_I |u'|^2 + \frac{1}{2\varepsilon}\int_{\mathcal{O}} |u|^2,$$

The solution to the penalized problem can be computed exactly:

$$u^\varepsilon = k_\varepsilon(x) \operatorname{sh}\left(\frac{x}{\sqrt{\varepsilon}}\right) \text{ in }]0, 1/2[\text{ with } k_\varepsilon(x) = \left(\operatorname{sh}\left(\frac{x}{\sqrt{\varepsilon}}\right) + \frac{1}{2\sqrt{\varepsilon}}\operatorname{ch}\left(\frac{x}{\sqrt{\varepsilon}}\right)\right)^{-1},$$

and u^ε affine in $]1/2, 1[$, continuous at $1/2$. This makes it possible to estimate $|u^\varepsilon - u|$, which turns out to behave like $\varepsilon^{1/4}$.

Yet, in many situations, convergence can be shown to be of order 1, given some assumptions are fulfilled. Let us introduce $\xi \in V'$ as the unique linear functional such that

$$a(u, v) + \langle \xi, v \rangle = \langle \varphi, v \rangle \quad \forall v \in V. \tag{4}$$

Before stating the first order convergence result, we show here that the penalty method provides an approximation of ξ.

Proposition 2. *Let $\xi^\varepsilon \in V'$ be defined by*

$$v \in V \longmapsto \langle \xi^\varepsilon, v \rangle = \frac{1}{\varepsilon}b(u^\varepsilon, v).$$

Then ξ^ε converges (strongly) to ξ in V', at least as fast as u^ε converges to u.

Proof. This is a direct consequence of the identity which we obtain by subtracting (4) and (2):

$$\langle \xi, v \rangle - \frac{1}{\varepsilon} b(u^\varepsilon, v) = a(u - u^\varepsilon, v) \quad \forall v \in V$$

which yields $\|\xi - \xi^\varepsilon\|_{V'} \leq C \, |u - u^\varepsilon|$.

Let us now establish the first order convergence, provided an extra compatibility condition between $b(\cdot, \cdot)$ and ξ is met.

Proposition 3. *Under the assumptions (1), we assume, in addition, that there exists $\tilde{\xi} \in V$ such that*

$$\langle \xi, v \rangle = b(\tilde{\xi}, v) \quad \forall v \in V.$$

Then $|u^\varepsilon - u| = \mathcal{O}(\varepsilon)$.

Proof. First of all, notice that it is possible to pick $\tilde{\xi}$ in $K^\perp$ (if not, we project it onto $K^\perp$). Now following the idea which is proposed in [Bab73] in a slightly different context (see the proof of Theorem 3.2 therein), we introduce

$$R_\varepsilon(v) = \frac{1}{2} a(u - v, u - v) + \frac{1}{2\varepsilon} b(\varepsilon \tilde{\xi} - v, \varepsilon \tilde{\xi} - v)$$

and we develop

$$R_\varepsilon(v) = \frac{1}{2} a(u, u) + \frac{\varepsilon}{2} b(\tilde{\xi}, \tilde{\xi}) + \frac{1}{2} a(v, v) + \frac{1}{2\varepsilon} b(v, v) - a(u, v) - b(\tilde{\xi}, v).$$

As $b(\tilde{\xi}, v) = \langle \xi, v \rangle$ and $-a(u, v) - \langle \xi, v \rangle = -\langle \varphi, v \rangle$, the functional R_ε is equal to J_ε up to a constant. Therefore, minimizing R_ε or minimizing J_ε are equivalent tasks. Let us now introduce $w = \varepsilon \tilde{\xi} + u$. It comes

$$R_\varepsilon(w) = \frac{\varepsilon^2}{2} a(\tilde{\xi}, \tilde{\xi}) + 0 \quad \text{because } u \in K = \ker b,$$

so that $|R_\varepsilon(w)| \leq C \varepsilon^2$. As u^ε minimizes R_ε,

$$0 \leq R_\varepsilon(u^\varepsilon) = \frac{1}{2} a(u - u^\varepsilon, u - u^\varepsilon) + \frac{1}{2\varepsilon} b(\varepsilon \tilde{\xi} - u^\varepsilon, \varepsilon \tilde{\xi} - u^\varepsilon) \leq C \varepsilon^2,$$

from which we deduce, as $a(\cdot, \cdot)$ is elliptic, $|u - u^\varepsilon| = \mathcal{O}(\varepsilon)$.

Corollary 1. *Under the assumptions (1), we assume, in addition, that $b(\cdot, \cdot)$ can be written $b(u, v) = (Bu, Bv)$, where B is a linear continuous operator onto a Hilbert space Λ, with closed range. Then $|u^\varepsilon - u| = \mathcal{O}(\varepsilon)$.*

Proof. Let us show that the assumption of Proposition 3 is met. It is sufficient to prove that any $\xi \in V'$ which vanishes over K identifies through $b(\cdot, \cdot)$ to some $\tilde{\xi} \in V$, i.e. there exists $\tilde{\xi} \in V$ such that

$$\langle \xi, v \rangle = b(\tilde{\xi}, v) \quad \forall v \in V.$$

Note that, as ξ vanishes over K, it can be seen as a linear functional defined on $K^{\perp}$, so that it is equivalent to establish that $T : V \longrightarrow (K^{\perp})'$ defined by

$$\tilde{\xi} \longmapsto \zeta : \langle \zeta, v \rangle = b(\tilde{\xi}, v) \quad \forall v \in K^{\perp}$$

is surjective. We denote by $T^{\star} \in \mathcal{L}(K^{\perp}, V)$ the adjoint of T. For all $w \in K^{\perp}$,

$$|T^{\star} w| = \sup_{v \neq 0} \frac{(T^{\star} w, v)}{|v|} = \sup_{v \neq 0} \frac{b(w, v)}{|v|} = \sup_{v \neq 0} \frac{(Bw, Bv)}{|v|} \geq \frac{|Bw|^2}{|w|}.$$

As B has closed range, $|Bw| \geq C |w|$ for all w in $(\ker B)^{\perp} = K^{\perp}$, so that

$$|T^{\star} w| \geq C^2 |w| \quad \forall w \in K^{\perp},$$

from which we conclude that T is surjective.

Remark 1. Note that Proposition 3 is strictly stronger than its corollary. Consider the standard situation $V = H^1(\Omega)$ where Ω is a smooth, bounded domain, $a(u, v) = \int \nabla u \cdot \nabla v$, and $\langle \varphi, v \rangle = \int f v$, where f is L^2, and $b(v, v) = \int_{\partial \Omega} v^2$. In this situation the corollary cannot be used, because the trace operator from $H^1(\Omega)$ onto $L^2(\partial \Omega)$ does not have a close range. On the other hand, one can establish that

$$\langle \xi, v \rangle = \int_{\partial \Omega} \frac{\partial u}{\partial n} v$$

and, as the solution u is regular ($u \in H^2(\Omega)$), its normal derivative (in $H^{1/2}(\partial \Omega)$) can be built as the trace of a function $\tilde{\xi}$ in $H^1(\Omega)$, so that Proposition 3 holds true.

We conclude this section by some considerations concerning the saddle-point formulation of the constrained problem. We consider again the closed situation:

Proposition 4. *Under the assumptions of Corollary 1, there exists $\lambda \in \Lambda$ such that*

$$a(u, v) + (\lambda, Bv) = \langle \varphi, v \rangle \quad \forall v \in V. \tag{5}$$

The solution is unique in $B(V)$ (which identifies to $\Lambda / \ker B^{\star}$).

Proof. The proof of this standard property can be found in [BF91]. In fact, it has just been established in the proof of Corollary 1: $\boldsymbol{\lambda}$ is simply $B\tilde{\xi}$. As for uniqueness in $B(V)$, consider two solutions $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2$. The equation (5) implies that $\boldsymbol{\lambda}_2 - \boldsymbol{\lambda}_1$ is in $\ker B^{\star} = B(V)^{\perp}$.

Proposition 5. *Under the assumptions of Proposition 4 (the assumptions (1) and $B(V)$ is closed), we introduce*

$$\lambda^\varepsilon = \frac{1}{\varepsilon} B u^\varepsilon.$$

Then $|\lambda^\varepsilon - \lambda| = \mathcal{O}(\varepsilon)$, where λ is the unique solution of (5) in $B(V)$.

Proof. Subtracting the variational formulations for u and u^ε, we get

$$(\lambda^\varepsilon - \lambda, Bv) = a(u^\varepsilon - u, v) \quad \forall v \in V.$$

Now, as the range of B is closed, and $\lambda^\varepsilon - \lambda \in B(V) = (\ker B^\star)^\perp$, we have the inf-sup condition (see, e.g., [BF91])

$$\sup_{v \in V} \frac{(\lambda^\varepsilon - \lambda, Bv)}{|v|} \geq \beta \, |\lambda^\varepsilon - \lambda| \, ,$$

so that

$$\beta \, |\lambda^\varepsilon - \lambda| \leq \sup \frac{(\lambda^\varepsilon - \lambda, Bv)}{|v|} = \sup \frac{a(u^\varepsilon - u, v)}{|v|} \leq \|a\| \, |u^\varepsilon - u| \, ,$$

which ensures the first order convergence thanks to Corollary 1.

2.2 Discretized Problem

We consider now a family $(V_h)_h$ of inner approximation spaces $(V_h \subset V)$ and the associated penalized/discretized problems

$$\begin{cases} \text{Find } u_h^\varepsilon \in V_h \text{ such that } J_h^\varepsilon(u_h^\varepsilon) = \inf_{v \in V} J_h^\varepsilon(v), \\[2mm] J_h^\varepsilon(v_h) = \frac{1}{2} a(v_h, v_h) + \frac{1}{2\varepsilon} b(v_h, v_h) - \langle \varphi, v_h \rangle. \end{cases} \tag{6}$$

As far as we know, there does not exist any general theory which would give an upper bound for the error $|u - u_h^\varepsilon|$ as the sum of a discretization error (typically h of $h^{1/2}$ for volume penalty, depending on whether the mesh is boundary-fitted or not), and a penalty error (typically ε for closed-range penalty terms). We propose here two general properties which are direct consequences of standard arguments. They are suboptimal in the sense that neither of them is optimal from both standpoints (discretization and penalty), but, at least, they make it possible to recover the behavior in extreme situations (when ε goes to 0 much quicker than h, and the opposite situation).

We shall need the following lemma:

Lemma 1. *Under the assumptions (1), there exists $C > 0$ such that*

$$b(u^\varepsilon, u^\varepsilon) \leq C\varepsilon \, |u - u^\varepsilon| \, .$$

Proof. By the definition of u^ε,

$$J_\varepsilon(u^\varepsilon) = \frac{1}{2}a(u^\varepsilon, u^\varepsilon) - \langle \varphi, u^\varepsilon \rangle + \frac{1}{2\varepsilon}b(u^\varepsilon, u^\varepsilon) \leq J_\varepsilon(u) = \frac{1}{2}a(u, u) - \langle \varphi, u \rangle,$$

so that

$$0 \leq \frac{1}{2\varepsilon}b(u_\varepsilon, u_\varepsilon) \leq \frac{1}{2}a(u, u) - \frac{1}{2}a(u^\varepsilon, u^\varepsilon) + \langle \varphi, u^\varepsilon - u \rangle$$

$$\leq \frac{1}{2}a(u + u^\varepsilon, u - u^\varepsilon) + \langle \varphi, u^\varepsilon - u \rangle,$$

which yields the estimate by continuity of $a(\cdot, \cdot)$ and φ.

Proposition 6. *Under the assumptions (1), we denote by u_h^ε the solution to the problem (6). Then*

$$|u_h^\varepsilon - u| \leq C \left(\min_{v_h \in V_h \cap K} |v_h - u| + \sqrt{|u^\varepsilon - u|} \right).$$

Proof. As u_h^ε minimizes $a(v - u^\varepsilon, v - u^\varepsilon) + b(v - u^\varepsilon, v - u^\varepsilon)/\varepsilon$ over V_h,

$$\alpha |u_h^\varepsilon - u^\varepsilon|^2 \leq a(u_h^\varepsilon - u^\varepsilon, u_h^\varepsilon - u^\varepsilon)$$

$$\leq a(u_h^\varepsilon - u^\varepsilon, u_h^\varepsilon - u^\varepsilon) + \frac{1}{\varepsilon}b(u_h^\varepsilon - u^\varepsilon, u_h^\varepsilon - u^\varepsilon)$$

$$\leq \min_{v_h \in V_h} \left(a(v_h - u^\varepsilon, v_h - u^\varepsilon) + \frac{1}{\varepsilon}b(v_h - u^\varepsilon, v_h - u^\varepsilon) \right)$$

$$\leq \min_{v_h \in V_h \cap K} \left(a(v_h - u^\varepsilon, v_h - u^\varepsilon) + \frac{1}{\varepsilon}b(v_h - u^\varepsilon, v_h - u^\varepsilon) \right).$$

As v_h is in K, the second term is $b(u^\varepsilon, u^\varepsilon)/\varepsilon$, which is bounded by $C|u^\varepsilon - u|$ (by Lemma 1). Finally, we get

$$|u_h^\varepsilon - u^\varepsilon| \leq C \left(\min_{v_h \in V_h \cap K} |v_h - u^\varepsilon| + \sqrt{|u^\varepsilon - u|} \right),$$

from which we conclude.

Proposition 7. *Under the assumptions (1), it holds*

$$|u_h^\varepsilon - u| \leq \frac{C}{\sqrt{\varepsilon}} \inf_{v_h \in V_h} |u^\varepsilon - v_h| + |u^\varepsilon - u|,$$

where u_h^ε is the solution to (6).

Proof. One has

$$|u_h^\varepsilon - u| \leq |u_h^\varepsilon - u^\varepsilon| + |u^\varepsilon - u|,$$

and we control the first term by Céa's lemma applied to the bilinear form $a + b/\varepsilon$, whose ellipticity constant behaves like $1/\varepsilon$.

Example 2. The simplest example of penalty formulation one may think about is the following: the constraint to vanish on a subdomain $\mathcal{O} \subset\subset \Omega$ is handled by minimizing the functional

$$J_\varepsilon(v) = \frac{1}{2} \int_\Omega |\nabla v|^2 - \int_\Omega fv + \frac{1}{2\varepsilon} \int_{\mathcal{O}} u^2. \tag{7}$$

Example 1 (which is a one-dimensional version of this situation) suggests that $|u^\varepsilon - u|$ behaves like $\varepsilon^{1/4}$. If we admit this convergence rate, Proposition 6 provides an estimate in $h^{1/2} + \varepsilon^{1/8}$. This estimate is optimal in h: the natural space discretization order is obtained if ε is small enough ($\varepsilon = h^4$ in the present case).

Symmetrically, the natural order in ε can be recovered if h is small enough. Indeed, if we admit that u^ε can be approximated at the same order as u over Ω, which is $1/2$, then the choice $\varepsilon = h^{4/3}$ in Proposition 7 gives

$$|u_h^\varepsilon - u| \leq \frac{C}{\varepsilon^{1/2}} \varepsilon^{3/4} + \varepsilon^{1/4} = \mathcal{O}(\varepsilon^{1/4}).$$

Notice that if we replace u^2 by $u^2 + |\nabla u|^2$ in the integral over $\mathcal{O}$ in (7), the assumptions of Corollary 1 are fulfilled, so that convergence holds at the first order in ε. As a consequence, $|u - u_h^\varepsilon|$ is bounded by $C(h^{1/2} + \varepsilon^{1/2})$ (by Proposition 6), which suggests the choice $\varepsilon = h$.

3 Model Problem

This section is dedicated to a special situation, which can be seen as a scalar version of the rigidity constraint in a Stokes flow.

We introduce a domain $\Omega \subset \mathbb{R}^2$ (smooth, or polygonal and convex), and $\mathcal{O} \subset\subset \Omega$ which we suppose circular (see Remark 4, at the end of this paper, for extensions to more general situations). We consider the following problem:

$$\begin{cases} -\triangle u = f & \text{in } \Omega \setminus \overline{\mathcal{O}}, \\ u = 0 & \text{on } \partial\Omega, \\ u = U & \text{on } \partial\mathcal{O}, \\ \displaystyle\int_{\partial\mathcal{O}} \frac{\partial u}{\partial n} = 0, \end{cases} \tag{8}$$

where U is an unknown constant, and $f \in L^2(\Omega \setminus \overline{\mathcal{O}})$. The scalar field u can be seen as a temperature, and $\mathcal{O}$ as a zone with infinite conductivity.

Definition 1. *We say that u is a weak solution to (8) if $u \in V = H_0^1(\Omega)$, there exists $U \in \mathbb{R}$ such that $u = U$ almost everywhere in $\mathcal{O}$, and*

$$\int_\Omega \nabla u \cdot \nabla v = \int_\Omega fv \quad \forall v \in \mathbf{D}(\Omega).$$

Proposition 8. *The problem (8) admits a unique weak solution* $u \in V = H_0^1(\Omega)$, *which is characterized as the solution to the minimization problem*

$$
\begin{cases}
\text{Find } u \in K \text{ such that} \\
J(u) = \inf_{v \in K} J(v), \quad \text{with } J(v) = \frac{1}{2} \int_\Omega |\nabla u|^2 - \int_\Omega fv \\
K = \{v \in H_0^1(\Omega) \mid \nabla v = 0 \text{ a.e. in } \mathcal{O}\},
\end{cases}
\tag{9}
$$

where f *has been extended by* 0 *inside* $\mathcal{O}$.
  Furthermore, the restriction of u *to* $\Omega \setminus \overline{\mathcal{O}}$ *is in* $H^2(\Omega \setminus \overline{\mathcal{O}})$.

Proof. Existence and uniqueness are direct consequences of the Lax–Milgram theorem applied in $K = \{v \in V \mid \nabla v = 0 \text{ a.e. in } \mathcal{O}\}$, which gives, in addition, the characterization of u as the solution to (9). Now the restriction of u to $\Omega \setminus \overline{\mathcal{O}}$ satisfies $-\triangle u = f$, with regular Dirichlet boundary conditions on the boundary of $\Omega \setminus \overline{\mathcal{O}}$ which decomposes as $\partial \mathcal{O} \cup \partial \Omega$. As Ω is a convex polygon and $\partial \mathcal{O}$ is smooth, standard theory ensures $u|_{\Omega \setminus \overline{\mathcal{O}}} \in H^2(\Omega \setminus \overline{\mathcal{O}})$.

We introduce the penalized version of the problem (9)

$$
\begin{cases}
\text{Find } u^\varepsilon \in V \text{ such that } J^\varepsilon(u^\varepsilon) = \inf_{v \in V} J^\varepsilon(v), \\
J^\varepsilon(v) = \frac{1}{2} \int_\Omega |\nabla v|^2 + \frac{1}{2\varepsilon} \int_\mathcal{O} |\nabla v|^2 - \int_\Omega fv,
\end{cases}
\tag{10}
$$

for which linear convergence can be expected:

Proposition 9. *Let* u *be the solution to the problem (9),* u^ε *the solution to the problem (10). It holds* $\|u - u^\varepsilon\|_{H^1(\Omega)} = \mathcal{O}(\varepsilon)$.

Proof. Let us show that

$$
B : v \in H_0^1(\Omega) \longmapsto \nabla v \in L^2(\mathcal{O})^2
$$

has a closed range. Consider $\mu \in \Lambda$ with $\mu = \nabla v$. We define $w \in H_0^1(\mathcal{O})$ as $w = v - m(v)$, where $m(v)$ is the mean value of v over $\mathcal{O}$. By the Poincaré–Wirtinger inequality, one has

$$
\|w\|_{H^1(\mathcal{O})} \leq C \|\mu\|_{L^2(\mathcal{O})^2} .
$$

Now, as $\mathcal{O} \subset\subset \Omega$, there exists a continuous extension operator from $H^1(\mathcal{O})$ to $H_0^1(\Omega)$, so that we can extend w to obtain $\tilde{w} \in H_0^1(\Omega)$ with a norm controlled by $\|\mu\|_{L^2(\mathcal{O})^2}$, which proves the closed character of $B(V)$. The linear convergence is then given by Corollary 1.

3.1 Standard Estimate

Now we consider the family of Cartesian triangulations (T_h) of the square Ω (see Fig. 1), and we denote by V_h the standard Finite Element space of continuous, piecewise affine function with respect to T_h.

The discretized/penalized problem reads

$$\begin{cases} \text{Find } u_h^\varepsilon \in V_h \text{ such that } J^\varepsilon(u_h^\varepsilon) = \inf_{v \in V_h} J^\varepsilon(v_h), \\ J^\varepsilon(v_h) = \frac{1}{2} \int_\Omega |\nabla v_h|^2 + \frac{1}{2\varepsilon} \int_{\mathcal{O}} |\nabla v_h|^2 - \int_\Omega f v_h. \end{cases} \tag{11}$$

Proposition 10 (Error estimate for (8)). *Let u be the weak solution to (9) and u_h^ε the solution to (11). There exists $C > 0$ such that*

$$|u - u_h^\varepsilon| \le C(h^{1/2} + \varepsilon^{1/2}). \tag{12}$$

Proof. The proof relies on Proposition 6, which asserts

$$|u_h^\varepsilon - u| \le C \left(\min_{v_h \in V_h \cap K} |v_h - u| + \sqrt{|u^\varepsilon - u|} \right).$$

By Proposition 9 the second term scales like $\varepsilon^{1/2}$. As for the space discretization term, the $h^{1/2}$-estimate is given by Proposition 11 below.

Proposition 11 (Approximation of u). *We make the same assumptions as in the beginning of the section, and we consider $u \in H_0^1(\Omega)$ such that $u = U \in \mathbb{R}$ a.e. in $\mathcal{O}$, $u_{\Omega \setminus \overline{\mathcal{O}}} \in H^2(\Omega \setminus \overline{\mathcal{O}})$. As previously, we consider a*
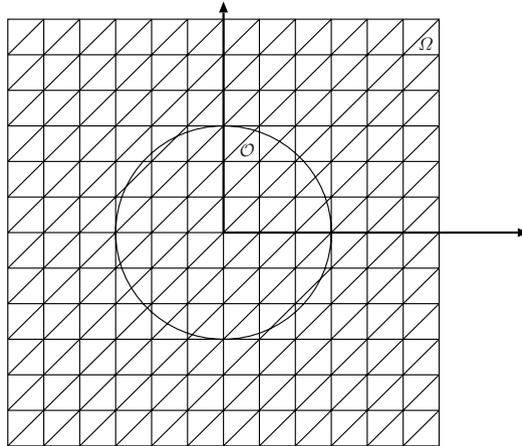

Fig. 1. Domains Ω, $\mathcal{O}$, and the mesh T_h.

Cartesian triangulation T_h of Ω and the associated first order Finite Element space V_h. There exists $C > 0$ such that

$$\inf_{\tilde{u}_h \in V_h} \|u - \tilde{u}_h\|_{H^1(\Omega)} \leq Ch^{1/2}.$$

Proof. We shall use in the proof the following notations: given a domain ω and v a function over ω, we denote by $|u|_{0,\omega}$ the L^2-norm of v over ω, by

$$|v|_{1,\omega} = \left(\int_\omega |\nabla v|^2 \right)^{1/2}$$

the H^1-seminorm, etc...

We denote by I_h is the standard interpolation operator from $C(\Omega)$ onto V_h. Notice that u is continuous over Ω (it is piecewise H^2, and continuous through the interface $\partial\mathcal{O}$). Let us assume here that the constant value U on $\mathcal{O}$ is 0 (which can be achieved by subtracting a smooth extension of this constant outside $\mathcal{O}$). We define $\tilde{u}_h$ as the function in V_h which is 0 at every vertex contained in a triangle which intersects $\mathcal{O}$, and which identifies to $I_h u$ at all other vertices. We introduce a narrow band around $\mathcal{O}$

$$\omega_h = \{ x \in \Omega \mid x \notin \overline{\mathcal{O}}, \ d(x, \overline{\mathcal{O}}) < 2\sqrt{2}h \}.$$

As $u|_{\Omega \setminus \overline{\mathcal{O}}} \in H^2(\Omega \setminus \overline{\mathcal{O}})$, the standard finite element estimate gives

$$|u - \tilde{u}_h|_{0, \Omega \setminus (\mathcal{O} \cup \overline{\omega}_h)} \leq Ch^2 |u|_{2, \Omega \setminus \overline{\mathcal{O}}}, \tag{13}$$

$$|u - \tilde{u}_h|_{1, \Omega \setminus (\mathcal{O} \cup \overline{\omega}_h)} \leq Ch |u|_{2, \Omega \setminus \overline{\mathcal{O}}}. \tag{14}$$

By construction, both L^2- and H^1-errors in $\mathcal{O}$ are zero. There remain to estimate the error in the band ω_h. The principle is the following: $\tilde{u}_h$ is a poor approximation of u in ω_h, but it is not very harmful because ω_h is small. Similar estimates are proposed in [SMSTT05] or [AR]. We shall give here a proof more adapted to our situation. First of all, we write

$$\|u - \tilde{u}_h\| \leq |u|_{0,\omega_h} + |u|_{1,\omega_h} + |u_h|_{0,\omega_h} + |u_h|_{1,\omega_h} = A + B + C + D. \tag{15}$$

We assume here that u is C^2 in $\Omega \setminus \overline{\mathcal{O}}$ (the general case $h \in H^2(\Omega \setminus \overline{\mathcal{O}})$ is obtained immediately by density). Using polar coordinates (we assume here that the radius of $\mathcal{O}$ is 1), we write

$$|u|_{1,\omega_h}^2 = \int_0^{2\pi} \int_1^{1+2h} |\nabla u|^2 \, r \, dr \, d\theta.$$

For $i = 1, 2$, one has

$$\partial_i u(r, \theta) = \partial_i u(1, \theta) + \int_1^r \partial_r \partial_i u \, dr,$$

so that

$$|u|^2_{1,\omega_h} \leq 2 \int_0^{2\pi} \int_1^{1+2h} |\partial_i u(1,\theta)|^2 \, r \, dr \, d\theta + 2 \int_0^{2\pi} \int_1^{1+2h} \left| \int_1^r \partial_r \partial_i u \, ds \right|^2 r \, dr \, d\theta$$

$$\leq C h \int_{\partial \mathcal{O}} \left| \frac{\partial u}{\partial n} \right|^2 + 2h \int_0^{2\pi} \int_1^{1+2h} \left(\int_1^{1+2h} |\partial_r \partial_i u|^2 \, ds \right) r \, dr \, d\theta$$

$$\leq C h \, |u|^2_{2,\Omega \setminus \overline{\mathcal{O}}} + C' h^2 \, |u|^2_{2,\Omega \setminus \overline{\mathcal{O}}},$$

from which we deduce $|u|_{1,\omega_h} \leq C h^{1/2}$. A similar computation on u gives immediately $|u|_{0,\omega_h} \leq C h^{3/2}$. As for $\tilde{u}_h$ (the two last terms in (15)), the proof is less trivial. It relies on three technical lemmas which we give now before ending the proof.

Lemma 2. *There exist constants C and C' such that, for any non-degenerated triangle T, for any function w_h affine in T,*

$$C \, |T| \, \|w_h\|^2_{L^\infty(T)} \leq \|w_h\|^2_{L^2(T)} \leq C' \, |T| \, \|w_h\|^2_{L^\infty(T)} . \tag{16}$$

Proof. It is a consequence of the fact that, when deforming the supporting triangle T, the L^∞-norm is unchanged whereas the L^2-norm scales like $|T|^{1/2}$.

Lemma 3. *There exists a constant C such that, for any non-degenerated triangle T, for any function w_h affine in T,*

$$|w_h|^2_{1,T} \leq C \frac{|T|}{\rho_T^2} \, \|w_h\|^2_{L^\infty(T)},$$

where ρ_T is the diameter of the inscribed circle.

Proof. Again, it is a straightforward consequence of the fact that, when deforming the supporting triangle T, L^∞-norm is unchanged whereas the gradient (which is constant over the triangle) scales like $1/\rho_T$, so that the H^1-seminorm scales like $|T|/\rho_T$.

The last lemma quantifies how one can control the L^2-norm of the interpolate of a regular function on a triangle, by means of the L^2-norm and the H^2-seminorm of the function.

Lemma 4. *There exists a constant C such that, for any regular triangle T (see below), for any $u \in H^2(T)$,*

$$|I_h u|^2_{0,T} \leq C \left(|u|^2_{0,T} + h^4 \, |u|^2_{2,T} \right).$$

By regular, we mean that T runs over a set of triangles such that the flatness $\operatorname{diam} T / \rho_T$ is bounded.

Proof. The interpolation operator $I_h : H^2(T) \to L^2(T)$ is continuous, and $|u|_{2,T}$ scales like $h/\rho_T^2 \approx 1/h$ whereas the L^2-norms scale like h.

We may now complete the proof of Proposition 11. The problematic triangles are those on which $\tilde{u}_h$ neither identifies to 0 nor to $I_h u$. On such triangles, $\tilde{u}_h$ sticks to $I_h u$ at 1 or 2 vertices, and vanishes at 2 or 1 vertices. As a consequence, the L^∞-norm of $\tilde{u}_h$ is less than the L^∞-norm of $I_h u$. Let T be such a triangle. We write (using Lemma 2, the latter remark, the fact that I_h is a contraction from L^∞ onto L^∞, Lemma 2 again, and Lemma 4),

$$|\tilde{u}_h|_{0,T}^2 \leq C' \, |T| \, \|\tilde{u}_h\|_{L^\infty(T)}^2 \leq C' \, |T| \, \|I_h u\|_{L^\infty(T)}^2$$
$$\leq \frac{C'}{C} \, \|I_h u\|_{L^2(T)}^2 \leq C'' \left(|u|_{0,T}^2 + h^4 \, |u|_{2,T}^2 \right).$$

By summing up all these contributions over all triangles outside $\mathcal{O}$ which intersect ω_h (they are all contained in ω_{2h}) and using the fact that the L^2-norm of u on ω_h behaves like $h^{3/2} \, |u|_{2,\omega_h}$, we obtain

$$|\tilde{u}_h|_{0,\omega_h}^2 \leq \sum_{T \cap \omega_h \neq \emptyset} |\tilde{u}_h|_{0,T}^2 \leq C \left(|u|_{0,\omega_{2h}}^2 + h^4 \, |u|_{2,\omega_{2h}}^2 \right) \leq C h^3 \, |u|_{2,\omega_{2h}}^2 \, ,$$

which gives the expected $h^{3/2}$-estimate for $|u_h|_{0,\omega_h}$. The last term of (15) is handled straightforwardly: Thanks to Lemmas 2 and 3, which imply the inverse inequality $|\tilde{u}_h|_{1,\omega_h} \leq C h^{-1} |\tilde{u}_h|_{0,\omega_h}$, we obtain the $h^{1/2}$ bound for $|\tilde{u}_h|_{1,\omega_h}$.

3.2 Primal/Dual Estimate

Proposition 5 asserts the convergence of $\boldsymbol{\lambda}^\varepsilon$ towards $\boldsymbol{\lambda}$, the Lagrange multiplier associated to the constraint. One may wonder whether $\boldsymbol{\lambda}_h^\varepsilon = B u_h^\varepsilon / \varepsilon$ is likely to approximate $\boldsymbol{\lambda}$. In general, a positive answer to that question can be given as soon as a uniform discrete inf-sup condition for B is fulfilled. This condition is not verified in the situation we consider. The non-uniformity of the inf-sup condition is due to the fact that there may exist triangles whose intersection with $\mathcal{O}$ is very small. We propose here a way to overcome this problem by suppressing those tiny areas in the penalty term, which leads us to introduce a discrete version B_h of B. Let us first give some properties for the continuous Lagrange multiplier, and we shall give a precise description of the way the obstacle is lifted.

Proposition 12 (Saddle-point formulation of (9)). *Let u be the weak solution to (8). There exists a unique $\boldsymbol{\lambda} \in \Lambda = L^2(\mathcal{O})^2$ such that $\boldsymbol{\lambda}$ is a gradient, and*

$$\int_\Omega \nabla u \cdot \nabla v + \int_\mathcal{O} \boldsymbol{\lambda} \cdot \nabla v = \int_\Omega f v \quad \forall v \in V.$$

In addition, $\boldsymbol{\lambda}$ is in $H^1(\mathcal{O})^2$.

Proof. The first part is a consequence of Proposition 4 (we established in the proof of Proposition 9 that the range of B is closed), which ensures the existence of $\boldsymbol{\lambda} \in \Lambda$ its uniqueness in $B(V)$.

Let us now describe λ. We have

$$\int_\Omega \nabla u \cdot \nabla v + \int_\mathcal{O} \boldsymbol{\lambda} \cdot \nabla v = \int_\Omega fv,$$

so that, by taking tests functions in $\mathbf{D}(\mathcal{O})$, we get $\boldsymbol{\lambda} \in H_{\mathrm{div}}(\mathcal{O})$ with $\nabla \cdot \boldsymbol{\lambda} = 0$. Taking now test functions which do not vanish on the boundary of $\mathcal{O}$, we identify the normal trace of $\boldsymbol{\lambda}$ as $\partial u / \partial n \in H^{1/2}(\partial \mathcal{O})$. Therefore, $\boldsymbol{\lambda}$ is defined as the unique divergence free vector field in $\mathcal{O}$, with normal derivative equal to $\partial u / \partial n$ on $\partial \mathcal{O}$, which, in addition, is a gradient. In other words: $\boldsymbol{\lambda} = \nabla \Phi$, with

$$\begin{cases} \triangle \Phi = 0 & \text{in } \mathcal{O}, \\ \dfrac{\partial \Phi}{\partial n} = \dfrac{\partial u}{\partial n} & \text{on } \partial \mathcal{O}. \end{cases}$$

As $\mathcal{O}$ is smooth, $\Phi \in H^2(\mathcal{O})$, so that $\boldsymbol{\lambda} = \nabla \Phi \in H^1(\mathcal{O})^2$.

Now we consider again the family of Cartesian triangulations (T_h) of the square Ω (see Fig. 1), and we denote by V_h the standard Finite Element space of continuous, piecewise affine functions with respect to T_h. As indicated in the beginning of this section, we suppress the small areas in the computation of the penalty term by introducing a reduced obstacle $\mathcal{O}_h$:

Definition 2. *The reduced obstacle $\mathcal{O}_h \subset \mathcal{O}$ is defined as the union of the sets $T \cap \mathcal{O}$, where T runs over triangles of T_h such that their intersection with $\mathcal{O}$ compares reasonably with their own size, in the following sense: given $\eta > 0$ a fixed parameter, we set*

$$\mathcal{O}_h = \bigcup_{|T \cap \mathcal{O}| \geq \eta |T|} (T \cap \mathcal{O}). \tag{17}$$

Definition 3. *We recall that $V = H_0^1(\Omega)$, Λ is $L^2(\mathcal{O})^2$, and $B \in \mathcal{L}(V, \Lambda)$ is the gradient operator (see Proposition 12). We define $B_h \in \mathcal{L}(V, \Lambda)$ as*

$$v \in V \longmapsto \mu = B_h v = \chi_{\mathcal{O}_h} \nabla v,$$

where $\chi_{\mathcal{O}_h}$ is the characteristic function of $\mathcal{O}_h$ (see Definition 2). Finally, the discretization space $\Lambda_h \subset \Lambda = L^2(\mathcal{O})^2$ is the set of all those vector fields μ_h such that their restriction to $\mathcal{O}_h$ is the gradient of a scalar field $v_h \in V_h$, and which vanish almost everywhere in $\mathcal{O} \setminus \mathcal{O}_h$, which we can express

$$\Lambda_h = \{\mu_h \in \Lambda \mid \exists v_h \in V_h, \ \mu_h = B_h v_h\} = B_h(V_h).$$

The fully discretized problem reads

$$
\begin{cases}
\text{Find } u_h^\varepsilon \in V_h \text{ such that } J_h^\varepsilon(u^\varepsilon) = \inf_{v_h \in V_h} J_h^\varepsilon(v_h), \\
J_h^\varepsilon(v_h) = \dfrac{1}{2} \int_\Omega |\nabla v_h|^2 + \dfrac{1}{2\varepsilon} \int_{\mathcal{O}_h} |\nabla v_h|^2 - \int_\Omega f v_h.
\end{cases} \tag{18}
$$

We may now state the primal/dual estimate.

Proposition 13 (Primal/dual error estimate for (8)). *Let u be the weak solution to (8), u_h^ε the solution to (11), $\boldsymbol{\lambda}$ the Lagrange multiplier (see Proposition 12), and $\boldsymbol{\lambda}_h^\varepsilon = B_h u_h^\varepsilon / \varepsilon$ (see Definition 3). We have the following error estimate:*

$$
|u - u_h^\varepsilon| + |\boldsymbol{\lambda} - \boldsymbol{\lambda}_h^\varepsilon| \le C(h^{1/2} + \varepsilon^{1/2}). \tag{19}
$$

Proof. The proof of this estimate is quite technical (in particular, the discrete inf-sup condition, see below), and we shall detail it on a forthcoming paper. Let us simply say here that it relies on the following ingredients:

1. some general properties of the continuous penalty method which we established in the beginning of this section,
2. an abstract stability estimate for saddle point-like problems with stabilization, in the spirit of Theorem 1.2 in [BF91],
3. a uniform discrete inf-sup condition for B_h:

$$
\sup_{v_h \in V_h} \frac{(B_h v_h, \boldsymbol{\lambda}_h)}{|v_h|} \ge \beta \, \|\boldsymbol{\lambda}_h\|_{\Lambda_h} , \tag{20}
$$

4. some approximation properties for V_h (Proposition 11 and a similar property for the Lagrange multiplier).

Remark 2 (Optimal estimate, role of η in the definition of $\mathcal{O}_h$). The estimate we establish is still suboptimal in ε: the order $1/2$ is obtained, whereas the continuous method converges linearly. It is due to the fact that we had to introduce a discrete operator B_h, and the difference leads to an extra term which scales like $\varepsilon^{1/2}$. It calls for some comments on the parameter η which appears in the definitions of $\mathcal{O}_h$ and B_h (see Definitions 2 and 3). The smaller η is, the closer B_h approaches B, which reduces the $\varepsilon^{1/2}$ term in the estimate. This observation may suggest to have η go to zero in the theoretical estimate. But, on the other hand, when η goes to 0, so does the inf-sup constant β (see (20)), so that $1/\beta$, which is hidden in the constant C in the error estimate (19), blows up.

Remark 3 (Boundary fitted meshes). Although it is somewhat in contradiction with its original purpose, the penalty method can be used together with a discretization based on a boundary fitted mesh. In that case, the approximation error behaves no longer like $h^{1/2}$, but like h. More important, it is not necessary to get rid of the tiny triangles which were incompatible, in case of a Cartesian mesh, with the uniform discrete inf-sup condition. Now considering that the half order in ε was lost because of the fact we introduced a reduced obstacle, one can expect to recover the optimal order of convergence, both in h and in ε.

Remark 4 (Technical assumptions). Some assumptions we made are only technical and can surely be relaxed without changing the convergence results. For example, the inclusion, which we supposed circular, could be any smooth domain. Note that a convex polygon is not acceptable, as it is seen by the problem from the outside, so that u may no longer be in H^2, which rules out some of the approximation properties we made.

Remark 5 (Convergence in space). The poor rate of convergence in h is optimal for a non-boundary-fitted mesh, at least if we consider the H^1-error overall Ω. Indeed, as the solution is constant inside $\mathcal{O}$, non-constant outside with a jump in the normal derivative, the error within each element intersecting $\partial \mathcal{O}$ is a $\mathcal{O}(1)$ in this L^∞-norm. By summing up over all those triangles, which cover a zone whose measure scales like h, we end up with this $h^{1/2}$-error. Note that a better convergence could be expected, in theory, if one considers only the error in the domain of interest $\Omega \setminus \overline{\mathcal{O}}$, the question being now whether the bad convergence in the neighborhood of $\partial \mathcal{O}$ pollutes the overall approximation. Our feeling is that this pollution actually occurs, because nothing is done in the present approach to distinguish the real domain of interest from the fictitious domain (inside the obstacle), so that the method tends to balance the errors on both sides. An interesting way to privilegiate the side of interest is proposed in [DP02] for a boundary penalty method; it consists in having the diffusion coefficient vanish within Ω. Note that other methods have been proposed to reach the optimal convergence rate on non-boundary-fitted mesh (see [Mau01]), but they are less straightforward to implement.

The simplest way to improve the actual order of convergence is to carry out a local refinement strategy in the neighborhood of $\partial \mathcal{O}$ (see [RAB06], for example). By using elements of scale h^2 in this zone, one recovers the first order convergence in space, at least in practice.

Remark 6 (Meaning of λ). As we already mentioned, the Lagrange multiplier $\boldsymbol{\lambda}$ can often be interpreted as a force or a heat source which ensures the prescribed constraint, depending on the context, and it may be useful to estimate this term with accuracy. For example, the problem we considered can be reformulated as a control problem: find a source term g with zero mean value (no heat is injected into the system) which is subject to vanish outside $\overline{\mathcal{O}}$, such that the solution u to

$$-\triangle u = f + g, \quad u = 0 \quad \text{on } \partial \Omega,$$

is constant over $\mathcal{O}$. This equation is to be considered in the distributional sense, as g is surely not a function. (If it were L^2, for example, u would be in $H^2(\Omega)$, which is surely not true as its normal derivative overcomes a jump through $\partial \mathcal{O}$.) Abstractly speaking, this source term g is simply the opposite of the linear functional ξ which we introduced (see (4)) and it is related to the Lagrange multiplier $\boldsymbol{\lambda}$ (see (5))

$$\langle g, v \rangle = -\langle \xi, v \rangle = -\int_{\mathcal{O}} \boldsymbol{\lambda} \cdot \nabla v = -\int_{\partial \mathcal{O}} \boldsymbol{\lambda} \cdot \mathbf{n}\, v.$$

The source term g is, therefore, a single layer distribution supported by $\partial\mathcal{O}$, with weight $-\lambda \cdot \mathbf{n} = -\partial u/\partial n$ (where $\mathbf{n}$ is the outward normal to $\partial\mathcal{O}$). Note that it is in $H^{1/2}(\partial\mathcal{O})$.

Remark 7. Note that letting ε go to 0 for any $h > 0$ leads to an estimate for a fictitious domain method (*à la* Glowinski, i.e. based on the use of Lagrange multiplier). In [GG95], an error estimate is obtained for such a method; it relies on two independent meshes for the primal and dual components of the solution (conditionally to some compatibility conditions between the sizes of the two meshes). We recover this estimate in the situation where the local mesh is simply the restriction of the covering mesh to the obstacle (to the reduced obstacle $\mathcal{O}_h$, to be more precise).

Remark 8. The approach we presented can be extended to other situations, like the one we already considered in Example 2, as soon as a H^1-penalty is used. The functional to minimize is then

$$J_\varepsilon(v) = \frac{1}{2} \int_\Omega |\nabla v|^2 - \int_\Omega fv + \frac{1}{2\varepsilon} \int_\Omega \left(u^2 + |\nabla u|^2 \right),$$

so that B identifies to the restriction operator from $H_0^1(\Omega)$ to $H^1(\mathcal{O})$. The discrete inf-sup condition, as well as the approximation properties, are essentially the same as in the case we considered here.

Concerning the original problem of simulating fluid-particle flows, an extra difficulty lies in the fact that two constraints of different types must be dealt with (global incompressibility and local rigid motion). It raises additional issues which shall be addressed in the future.

References

[ABF99]   Ph. Angot, Ch.-H. Bruneau, and P. Fabrie. A penalization method to take into account obstacles in incompressible viscous flows. *Numer. Math.*, 81(4):497–520, 1999.

[AR]   Ph. Angot and I. Ramière. Convergence analysis of the Q1-finite element method for elliptic problems with non boundary-fitted meshes. Submitted.

[Bab73]   I. Babuška. The finite element method with penalty. *Math. Comp.*, 27:221–228, 1973.

[BF91]   F. Brezzi and M. Fortin. *Mixed and hybrid finite element methods*, volume 15 of *Springer Series in Computational Mathematics*. Springer-Verlag, New York, 1991.

[DP02]   S. Del Pino. *Une méthode d'éléments finis pour la résolution d'EDP dans des domaines décrits par géométrie constructive.* PhD thesis, Université Pierre et Marie Curie, Paris, 2002.

[DPP03]   S. Del Pino and O. Pironneau. A fictitious domain based general PDE solver. In E. Heikkola, editor, *Numerical Methods for Scientific Computing*, Barcelona, 2003.

[ff3]   freeFEM3D (`http://www.freefem.org/ff3d/`).

[FFp]   freeFEM++ (`http://www.freefem.org/`).

[GG95]   V. Girault and R. Glowinski. Error analysis of a fictitious domain method applied to a Dirichlet problem. *Japan J. Indust. Appl. Math.*, 12(3):487–514, 1995.

[JLM05]   J. Janela, A. Lefebvre, and B. Maury. A penalty method for the simulation of fluid-rigid body interaction. In *CEMRACS 2004—mathematics and applications to biology and medicine*, volume 14 of *ESAIM Proc.*, pages 115–123 (electronic). EDP Sci., Les Ulis, 2005.

[JT96]   A. A. Johnson and T. E. Tezduyar. Simulation of multiple spheres falling in a liquid-filled tube. *Comput. Methods Appl. Mech. Engrg.*, 134(3-4): 351–373, 1996.

[lif]   LifeV (`http://www.lifev.org/`).

[Mau99]   B. Maury. Direct simulations of 2D fluid-particle flows in biperiodic domains. *J. Comput. Phys.*, 156(2):325–351, 1999.

[Mau01]   B. Maury. A fat boundary method for the Poisson problem in a domain with holes. *J. Sci. Comput.*, 16(3):319–339, 2001.

[PG02]   T.-W. Pan and R. Glowinski. Direct simulation of the motion of neutrally buoyant circular cylinders in plane Poiseuille flow. *J. Comput. Phys.*, 181(1):260–279, 2002.

[RAB06]   I. Ramière, Ph. Angot, and M. Belliard. A fictitious domain approach with spread interface for elliptic problems with general boundary conditions. *Comput. Methods App. Mech. Engrg.*, 196(4–6):766–781, 2007.

[RPVC05]   T. N. Randrianarivelo, G. Pianet, S. Vincent, and J. P. Caltagirone. Numerical modelling of solid particle motion using a new penalty method. *Internat. J. Numer. Methods Fluids*, 47:1245–1251, 2005.

[SMSTT05]   J. San Martín, J.-F. Scheid, T. Takahashi, and M. Tucsnak. Convergence of the Lagrange-Galerkin method for the equations modelling the motion of a fluid-rigid system. *SIAM J. Numer. Anal.*, 43(4):1536–1571 (electronic), 2005.

[VCLR04]   S. Vincent, J.-P. Caltagirone, P. Lubin, and T. N. Randrianarivelo. An adaptive augmented Lagrangian method for three-dimensional multimaterial flows. *Comput. & Fluids*, 33(10):1273–1289, 2004.

A Numerical Method for Fluid Flows
with Complex Free Surfaces

Andrea Bonito[1]* and Alexandre Caboussat[2], Marco Picasso[3],
and Jacques Rappaz[3]

[1] Department of Mathematics, University of Maryland, College Park, MD
    20742-4015, USA `andrea.bonito@a3.epfl.ch`
[2] Department of Mathematics, University of Houston, 77204-3008, Houston, TX,
    USA `caboussat@math.uh.edu`
[3] Institute of Analysis and Scientific Computing, Ecole Polytechnique Fédérale de
    Lausanne, 1015 Lausanne, Switzerland
    `{marco.picasso;jacques.rappaz}@epfl.ch`

Summary. A numerical method for the simulation of fluid flows with complex free
surfaces is presented. The liquid is assumed to be a Newtonian or a viscoelastic
fluid. The compressible effects of the surrounding gas are taken into account, as
well as surface tension forces. An Eulerian approach based on the volume-of-fluid
formulation is chosen. A time splitting algorithm, together with a two-grids method,
allows the various physical phenomena to be decoupled. A chronological approach is
adopted to highlight the successive improvements of the model and the wide range
of applications. Numerical results show the potentialities of the method.

1 Introduction

Complex free surface phenomena involving Newtonian and/or non-Newtonian
flows are nowadays a topic of active research in many fields of physics, engi-
neering or bioengineering. The literature contains numerous models for com-
plex liquid-gas free surfaces problems, see, e.g., [FCD+06, SZ99]. For instance,
when considering the injection of a liquid in a complex cavity initially filled
with gas, an Eulerian approach is generally adopted in order to catch the
topology changes of the liquid region.

Such two-phases flows are computationally expensive in three space di-
mensions since (at least) both the velocity and pressure must be computed at
each grid point of the whole liquid-gas domain.

The purpose of this article is to review a numerical model in order to
compute complex free surface flows in three space dimensions. The features

---

of the model are the following. A volume-of-fluid method is used to track
the liquid domain, which can exhibit complex topology changes. The velocity
field is computed only in the liquid region. The incompressible liquid can be
modeled either as a Newtonian or as a viscoelastic fluid. The ideal gas law is
used to compute the external pressure in the surrounding gas and the resulting
force is added on the liquid-gas free surface. Surface tension effects can also
be taken into account on the liquid-gas free surface. The complete description
of the model can be found in [MPR99, MPR03, CPR05, Cab06, BPL06].

The numerical model is based on a time-splitting approach [Glo03] and a
two-grids method. This allows advection, diffusion and viscoelastic phenomena
to be decoupled, as well as the treatment of the liquid and gas phases. Finite
element techniques [FF92] are used to solve the diffusion phenomena using an
unstructured mesh of the cavity containing the liquid. A forward characteristic
method [Pir89] on a structured grid allows advection phenomena to be solved
efficiently.

The article is structured as follows. In Section 2, the simplest model is
presented: the liquid is an incompressible Newtonian fluid, the effects of the
surrounding gas and surface tension are neglected. The effects of the surround-
ing gas are described in Section 3, those of the surface tension in Section 4.
Finally, the case of a viscoelastic liquid is considered in Section 5. Numerical
results are presented throughout the text and illustrate the capabilities and
improvements of the model.

2 Modeling of an Incompressible Newtonian Fluid with a Free Surface

2.1 Governing Equations

The model presented in this section has already been published in [MPR99,
MPR03]. Let Λ, with a boundary $\partial\Lambda$, be a cavity of $\mathbb{R}^3$ in which a liquid
must be confined, and let $T > 0$ be the final time of simulation. For any given
time $t \in (0, T)$, let Ω_t, with a boundary $\partial\Omega_t$, be the domain occupied by
the liquid, let $\Gamma_t = \partial\Omega_t \setminus \partial\Lambda$ be the free surface between the liquid and the
surrounding gas and let Q_T be the space-time domain containing the liquid,
i.e. $Q_T = \{(x, t) : x \in \Omega_t, \ 0 < t < T\}$.

In the liquid region, the velocity field $\mathbf{v} : Q_T \rightarrow \mathbb{R}^3$ and the pressure
field $p : Q_T \rightarrow \mathbb{R}$ are assumed to satisfy the time-dependent, incompressible
Navier–Stokes equations, that is

$$\rho\frac{\partial \mathbf{v}}{\partial t} + \rho(\mathbf{v} \cdot \nabla)\mathbf{v} - 2\operatorname{div}(\mu\mathbf{D}(\mathbf{v})) + \nabla p = \mathbf{f} \quad \text{in } Q_T, \tag{1}$$

$$\operatorname{div}\mathbf{v} = 0 \quad \text{in } Q_T. \tag{2}$$

Here $\mathbf{D}(\mathbf{v}) = 0.5 \cdot (\nabla\mathbf{v} + \nabla\mathbf{v}^T)$ denotes the rate of deformation tensor, ρ the
constant density and $\mathbf{f}$ the external forces.

The dynamic viscosity μ can be constant or, in order to take into account turbulence effects, a turbulent viscosity $\mu_T = \mu_T(\mathbf{v}) = \alpha_T \rho \sqrt{2\mathbf{D}(\mathbf{v}) : \mathbf{D}(\mathbf{v})}$, where α_T is a parameter to be chosen, is added. The use of a turbulent viscosity is required when large Reynolds numbers and thin boundary layers are involved. Otherwise, in order to consider Bingham flows (when considering mud flows or avalanches, for instance), a plastic viscosity $\mu_B = \alpha_0 \rho / \sqrt{2\mathbf{D}(\mathbf{v}) : \mathbf{D}(\mathbf{v})}$, where α_0 is a parameter to be chosen, can be added.

Let $\varphi : \Lambda \times (0,T) \to \mathbb{R}$ be the characteristic function of the liquid domain Q_T. The function φ equals one if the liquid is present, zero if it is not, thus $\Omega_t = \{x \in \Lambda : \varphi(x,t) = 1\}$. In order to describe the kinematics of the free surface, φ must satisfy (in a weak sense)

$$\frac{\partial \varphi}{\partial t} + \mathbf{v} \cdot \nabla \varphi = 0 \quad \text{in } \Lambda \times (0,T), \tag{3}$$

where the velocity $\mathbf{v}$ is extended continuously in the neighborhood of Q_T. At initial time, the characteristic function of the liquid domain φ is given, which defines the initial liquid region $\Omega_0 = \{x \in \Lambda : \varphi(x,0) = 1\}$. The initial velocity field $\mathbf{v}$ is prescribed in Ω_0.

The boundary conditions for the velocity field are the following. On the boundary of the liquid region being in contact with the walls (that is to say the boundary of Λ), inflow, slip or Signorini boundary conditions are enforced, see [MPR99, MPR03]. On the free surface Γ_t, the forces acting on the free surface are assumed to vanish, when both the influence of the external media and the capillary and surface tension effects are neglected on the free surface. If these influences are not neglected, we have to establish the equilibrium of forces on the free surface. In the first case, the following equilibrium relation is then satisfied on the liquid-gas interface:

$$-p\mathbf{n} + 2\mu \mathbf{D}(\mathbf{v})\mathbf{n} = 0 \quad \text{on } \Gamma_t, \ t \in (0,T), \tag{4}$$

where $\mathbf{n}$ is the unit normal of the liquid-gas free surface oriented toward the external gas.

The mathematical description of our model is complete. The model unknowns are the characteristic function φ in the whole cavity, the velocity $\mathbf{v}$ and pressure p in the liquid domain only. These unknowns satisfy the equations (1)–(3). Simplified problems extracted from this model of incompressible liquid flow with a free surface have been investigated theoretically in [CR05, Cab05], in one and two dimensions of space, and existence results and error estimates have been obtained.

2.2 Time Splitting Scheme

An implicit splitting algorithm is proposed to solve (1)–(3) by splitting the advection from the diffusion part of the Navier–Stokes equations. Let $0 = t^0 < t^1 < t^2 < \ldots < t^N = T$ be a subdivision of the time interval $[0,T]$, define
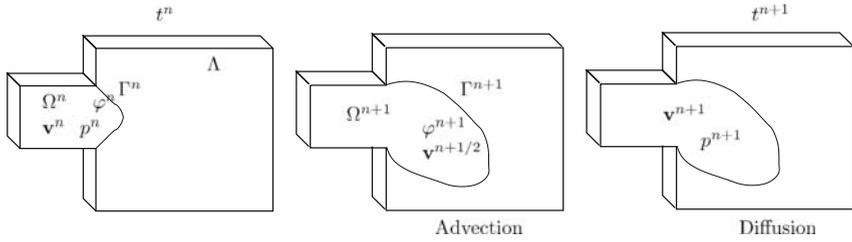
Fig. 1. The splitting algorithm (from left to right). Two advection problems are solved to determine the new approximation of the characteristic function φ^{n+1}, the new liquid domain Ω^{n+1} and the predicted velocity $\mathbf{v}^{n+1/2}$. Then, a generalized Stokes problem is solved in the new liquid domain Ω^{n+1} in order to obtain the velocity $\mathbf{v}^{n+1}$ and the pressure p^{n+1}.

$\delta t^n = t^{n+1} - t^n$ the n-th time step, $n = 0, 1, 2, \ldots, N$, δt the largest time step. Let φ^n, $\mathbf{v}^n$, p^n, Ω^n be approximations of φ, $\mathbf{v}$, p, Ω_t at time t^n, respectively. Then the approximations φ^{n+1}, $\mathbf{v}^{n+1}$, p^{n+1}, Ω^{n+1} at time t^{n+1} are computed by means of an implicit splitting algorithm, as illustrated in Figure 1.

Two advection problems are solved first, leading to a prediction of the new velocity $\mathbf{v}^{n+1/2}$ together with the new approximation of the characteristic function φ^{n+1} at time t^{n+1}, which allows to determine the new liquid domain Ω^{n+1} and the new liquid interface Γ^{n+1}. Then a generalized Stokes problem is solved on Ω^{n+1} with the boundary condition (4) on the liquid interface Γ^{n+1}, Dirichlet, slip or Signorini-type conditions on the boundary of the cavity Λ and the velocity $\mathbf{v}^{n+1}$ and pressure p^{n+1} in the liquid are obtained.

This time-splitting algorithm introduces an additional error on the velocities and pressures which is of order $\mathcal{O}(\delta t)$, see, e.g., [Mar90]. This algorithm allows the motion of the free surface to be decoupled from the diffusion step, which consists in solving a Stokes problem in a fixed domain [Glo03].

Advection Step. Solve between the times t^n and t^{n+1} the two advection problems:

$$\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v} \cdot \nabla)\mathbf{v} = 0, \quad \frac{\partial \varphi}{\partial t} + \mathbf{v} \cdot \nabla \varphi = 0 \tag{5}$$

with initial conditions $\mathbf{v}^n$ and φ^n. This step is solved exactly by the method of characteristics [Mau96, Pir89] which yields a prediction of the velocity $\mathbf{v}^{n+1/2}$ and the characteristic function of the new liquid domain φ^{n+1}:

$$\mathbf{v}^{n+1/2}(x + \delta t^n \mathbf{v}^n(x)) = \mathbf{v}^n(x) \quad \text{and} \quad \varphi^{n+1}(x + \delta t^n \mathbf{v}^n(x)) = \varphi^n(x) \tag{6}$$

for all x belonging to Ω^n. Then, the new liquid domain Ω^{n+1} is defined as the set of points such that φ^{n+1} equals one.

Diffusion Step. The diffusion step consists in solving a generalized Stokes problem on the domain Ω^{n+1} using the predicted velocity $\mathbf{v}^{n+1/2}$ and the boundary condition (4). The following backward Euler scheme is used:

$$\rho\frac{\mathbf{v}^{n+1} - \mathbf{v}^{n+1/2}}{\delta t^n} - 2\operatorname{div}\left(\mu\mathbf{D}(\mathbf{v}^{n+1})\right) + \nabla p^{n+1} = \mathbf{f}(t^{n+1}) \quad \text{in } \Omega^{n+1}, \quad (7)$$

$$\operatorname{div}\mathbf{v}^{n+1} = 0 \qquad \text{in } \Omega^{n+1}, \quad (8)$$

where $\mathbf{v}^{n+1/2}$ is the prediction of the velocity obtained with (6) after the advection step. The boundary conditions on the free surface are given by (4). The weak formulation corresponding to (7), (8) and (4), therefore, consists in finding $\mathbf{v}^{n+1}$ and p^{n+1} such that $\mathbf{v}^{n+1}$ is vanishing on $\partial\Lambda$ and

$$\rho\int_{\Omega^{n+1}}\frac{\mathbf{v}^{n+1} - \mathbf{v}^{n+1/2}}{\delta t^n}\cdot\mathbf{w}\,dx + 2\int_{\Omega^{n+1}}\mu\mathbf{D}(\mathbf{v}^{n+1}):\mathbf{D}(\mathbf{w})\,dx$$

$$-\int_{\Omega^{n+1}}p^{n+1}\operatorname{div}\mathbf{w}\,dx - \int_{\Omega^{n+1}}\mathbf{f}\cdot\mathbf{w}\,dx - \int_{\Omega^{n+1}}q\operatorname{div}\mathbf{v}^{n+1}\,dx = 0, \quad (9)$$

for all test functions $(\mathbf{w}, q)$ such that $\mathbf{w}$ vanishes on the boundary of the cavity where essential boundary conditions are enforced.

2.3 A Two-Grids Method for Space Discretization

Advection and diffusion phenomena being now decoupled, the equations (5) are first solved using the method of characteristics on a structured mesh of small cells in order to reduce numerical diffusion of the interface Γ_t between the liquid and the gas, and have an accurate approximation of the liquid region, see Figure 2 (left).

The bounding box of the cavity Λ is meshed into a structured grid made out of small cubic cells of size h, each cell being labeled by indices (ijk). Let φ_{ijk}^n and $\mathbf{v}_{ijk}^n$ be the approximate values of φ and $\mathbf{v}$ at the center of cell number (ijk) at time t^n. The unknown φ_{ijk}^n is the volume fraction of liquid in the cell ijk and is the numerical approximation of the characteristic function φ at

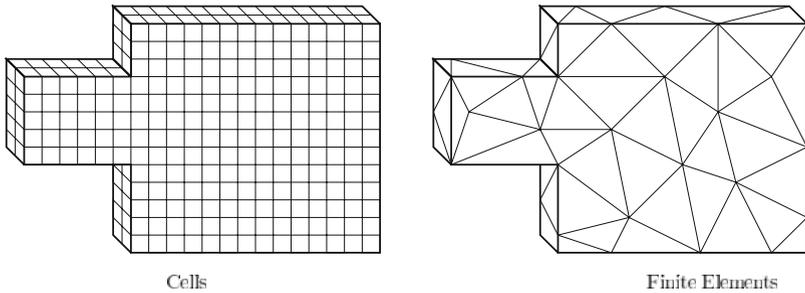Cells   Finite Elements

Fig. 2. Two-grids method. The advection step is solved on a structured mesh of small cubic cells composed of blocks whose union covers the physical domain Ω_h (left), while the diffusion step is solved on a finite element unstructured mesh of tetrahedra (right).
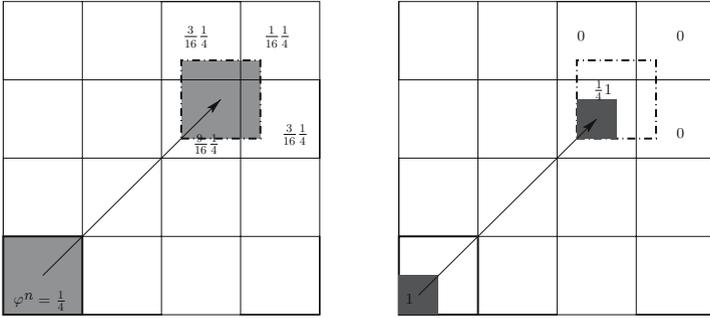
Fig. 3. Effect of the SLIC algorithm on numerical diffusion. An example of two dimensional advection and projection when the volume fraction of liquid in the cell is $\varphi_{ij}^n = \frac{1}{4}$. Left: without SLIC, the volume fraction of liquid is advected and projected on four cells, with contributions (from the top left cell to the bottom right cell) $\frac{3}{16}\frac{1}{4}$, $\frac{1}{16}\frac{1}{4}$, $\frac{9}{16}\frac{1}{4}$, $\frac{3}{16}\frac{1}{4}$. Right: with SLIC, the volume fraction of liquid is first pushed at one corner, then it is advected and projected on one cell only, with contribution $\frac{1}{4}$.

time t^n, which is piecewise constant on each cell of the structured grid. The advection step for the cell number (ijk) consists in advecting φ_{ijk}^n and $\mathbf{v}_{ijk}^n$ by $\delta t^n \mathbf{v}_{ijk}^n$ and then projecting the values on the structured grid, to obtain φ_{ijk}^{n+1} and a prediction of the velocity $\mathbf{v}_{ijk}^{n+\frac{1}{2}}$. A simple implementation of the SLIC (Simple Linear Interface Calculation) algorithm, described in [MPR03] and inspired by [NW76], allows to reduce the numerical diffusion of the domain occupied by the liquid by pushing the fluid along the faces of the cell before advecting it. The choice of how to push the fluid depends on the volume fraction of liquid of the neighboring cells. The cell advection and projection with SLIC algorithm are presented in Figure 3, in two space dimensions for the sake of simplicity. We refer to [AMS04] for a recent improvement of the SLIC algorithm.

Remark 1. A post-processing technique allows to avoid the *compression effects* and guarantees the conservation of the mass of liquid. Related to *global repair algorithms* [SW04], this technique produces final values φ_{ijk}^{n+1} which are between zero and one, even when the advection of φ^n gives values strictly larger than one. The technique consists in moving the fraction of liquid in excess in the cells that are over-filled to *receiver* cells in a global manner by sorting the cells according to φ^{n+1}. Details can be found in [MPR99, MPR03].

Once values φ_{ijk}^{n+1} and $\mathbf{v}_{ijk}^{n+1/2}$ have been computed on the cells, values of the fraction of liquid φ_P^{n+1} and of the velocity field $\mathbf{v}_P^{n+\frac{1}{2}}$ are computed at the nodes P of the finite element mesh with approximated projection methods. We take advantage of the difference of refinement between a coarse finite element

mesh and a finer structured grid of cells. Let $\mathcal{T}_h$ be the triangulation of the cavity Λ. For any vertex P of $\mathcal{T}_h$, let ψ_P be the corresponding finite element basis function (i.e. the continuous, piecewise linear function having value one at P, zero at the other vertices). Then, φ_P^{n+1}, the volume fraction of liquid at vertex P and time t^{n+1} is computed by:

$$\varphi_P^{n+1} = \left(\sum_{\substack{K \in \mathcal{T}_h \\ P \in K}} \sum_{\substack{ijk \\ C_{ijk} \in K}} \psi_P(C_{ijk}) \varphi_{ijk}^{n+1} \right) \Big/ \left(\sum_{\substack{K \in \mathcal{T}_h \\ P \in K}} \sum_{\substack{ijk \\ C_{ijk} \in K}} \psi_P(C_{ijk}) \right), \quad (10)$$

where C_{ijk} is the center of the cell (ijk). The same kind of formula is used to obtain the predicted velocity $\mathbf{v}^{n+\frac{1}{2}}$ at the vertices of the finite element mesh. When these values are available at the vertices of the finite element mesh, the approximation of the liquid region Ω_h^{n+1} used for solving (9) is defined as the union of all elements of the mesh $K \in \mathcal{T}_h$ with (at least) one of its vertices P such that $\varphi_P^{n+1} > 0.5$, the approximation of the free surface being denoted by Γ_h^{n+1}.

Numerical experiments reported in [MPR99, MPR03] have shown that choosing the size of the cells of the structured mesh approximately 5 to 10 times smaller than the size of the finite elements is a good choice to reduce numerical diffusion of the interface Γ_t. Furthermore, since the characteristics method is used, the time step is not restricted by the CFL number (which is the ratio between the time step times the maximal velocity divided by the mesh size). Numerical results in [MPR99, MPR03] have shown that a good choice generally consists in choosing CFL numbers ranging from 1 to 5.

Remark 2. In number of industrial mold filling applications, the shape of the cavity containing the liquid (the mold) is complex. Therefore, a special, hierarchical, data structure has been implemented in order to reduce the memory requirements, see [MPR03, RDG+00]. The cavity is meshed into tetrahedra for the resolution of the diffusion problem. For the advection part, a hierarchical structure of blocks, which cover the cavity and are glued together, is defined. A computation is performed inside a block if and only if it contains cells with liquid. Otherwise the whole block is deactivated.

The diffusion step consists in solving the Stokes problem (9) with finite element techniques. Let $\mathbf{v}_h^{n+1}$ (resp. p_h^{n+1}) be the piecewise linear approximation of $\mathbf{v}^{n+1}$ (resp. p^{n+1}). The Stokes problem is solved with stabilized $\mathbb{P}_1 - \mathbb{P}_1$ finite elements (Galerkin Least Squares, see [FF92]) and consists in finding the velocity $\mathbf{v}_h^{n+1}$ and pressure p_h^{n+1} such that:

$$\rho \int_{\Omega_h^{n+1}} \frac{\mathbf{v}_h^{n+1} - \mathbf{v}_h^{n+1/2}}{\delta t^n} \mathbf{w}\, d\mathbf{x} + 2 \int_{\Omega_h^{n+1}} \mu \mathbf{D}(\mathbf{v}_h^{n+1}) : \mathbf{D}(\mathbf{w})\, d\mathbf{x}$$

$$- \int_{\Omega_h^{n+1}} \mathbf{f}\mathbf{w}\, d\mathbf{x} - \int_{\Omega_h^{n+1}} p_h^{n+1} \operatorname{div} \mathbf{w}\, d\mathbf{x} - \int_{\Omega_h^{n+1}} \operatorname{div} \mathbf{v}_h^{n+1} q\, d\mathbf{x}$$

$$- \sum_{K \subset \Omega_h^{n+1}} \alpha_K \int_K \left(\frac{\mathbf{v}_h^{n+1} - \mathbf{v}_h^{n+1/2}}{\delta t^n} + \nabla p_h^{n+1} - \mathbf{f} \right) \cdot \nabla q\, d\mathbf{x} = 0, \quad (11)$$

for all $\mathbf{w}$ and q the velocity and pressure test functions, compatible with the boundary conditions on the boundary of the cavity Λ. The value of the parameter α_K is discussed in [MPR99, MPR03].

The projection of the continuous piecewise linear approximation $\mathbf{v}_h^{n+1}$ back on the cell (ijk) is obtained by interpolation of the piecewise finite element approximation at the center C_{ijk} of the cell. It allows to obtain a value of the velocity $\mathbf{v}_{ijk}^{n+1}$ on each cell ijk of the structured grid for the next time step.

2.4 Numerical Results

The classical "vortex-in-a-box" test case widely treated in the literature is considered here [RK98]. The initial liquid domain is a circle of radius 0.015 with its center located in $(0.05, 0.075)$. It is stretched by a given velocity, given by the stream function $\psi(x, y) = 0.01\pi \sin^2(\pi x/0.1) \sin^2(\pi y/0.1) \cos(\pi t/2)$. The velocity being periodic in time, the initial liquid domain is reached after a time $T = 2$. Figure 4 illustrates the liquid-gas interface for three structured meshes [CPR05]. The interface with maximum deformation and the interface after one period of time are represented. Numerical results show the efficiency and convergence of the scheme.

An S-shaped channel lying between two horizontal plates is filled. The channel is contained in a $0.17\,\mathrm{m} \times 0.24\,\mathrm{m}$ rectangle. The distance between the two horizontal plates is $0.008\,\mathrm{m}$. Water is injected at one end with
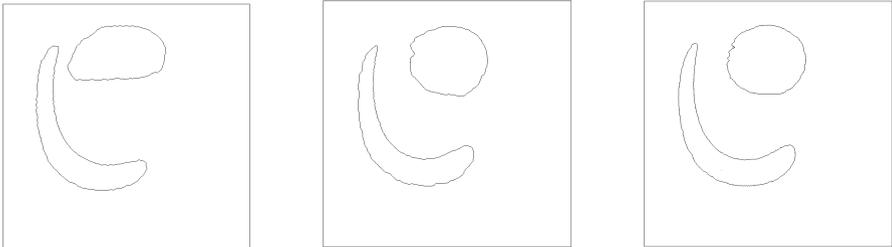

Fig. 4. Single vortex test case, representation of the computed interface at times $t = 1$ (maximal deformation) and $t = 2$ (return to initial shape). Left: coarser mesh, middle: middle mesh, right: finer mesh.
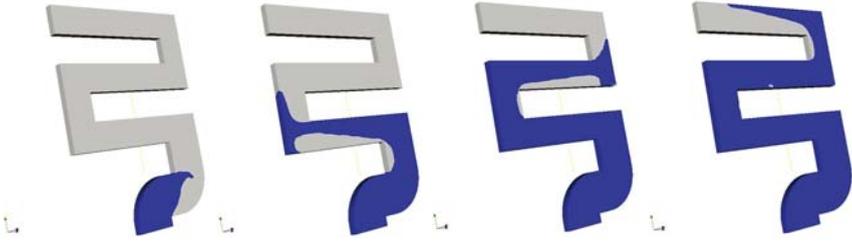
Fig. 5. S-shaped channel: 3D results when the cavity is initially filled with vacuum. Time equals 8.0 ms, 26.0 ms, 44.0 ms and 53.9 ms.

constant velocity $8.7\,\mathrm{m/s}$. Density and viscosity are taken to be respectively $\rho = 1000\,\mathrm{kg/m^3}$ and $\mu = 0.01\,\mathrm{kg/(ms)}$.

Slip boundary conditions are enforced to avoid boundary layers and a turbulent viscosity is added, the coefficient α_T being equal to $4h^2$, as proposed in [CPR05]. Since the ratio between Capillary number and Reynolds number is very small, surface tension effects are neglected.

The final time is $T = 0.0054$ s and the time step is $\tau = 0.0001$ s. The mesh is made out of 96030 elements. In Figure 5, 3D computations are presented when a valve is placed at the end of the cavity, thus allowing the gas to exit. The CPU time for the simulations in three space dimensions is approximately 319 minutes for 540 time steps. Most of the CPU time is spent to solve the Stokes problem. A comparison with experimental results shows that the bubbles of gas trapped by the liquid vanish too rapidly. In order to obtain more realistic results, the effect of the gas compressibility onto the liquid must be considered. This is the scope of the next section.

3 Extension to the Modeling of an Incompressible Liquid Surrounded by a Compressible Gas

3.1 Extension of the Model

In Section 2, the zero force condition (4) was applied on the liquid-gas interface. Going back to the simulation of Figure 5, this corresponds to filling with liquid a cavity under vacuum. When considering industrial mold filling processes, the mold is not initially under vacuum, but contains some compressible air that interacts with the liquid. Therefore, the model has to be extended. The velocity in the gas is disregarded here, since it is CPU time consuming to solve the Euler compressible equations in the gas domain. The model presented in Section 2 is extended by adding the normal forces due to the gas pressure on the free surface Γ_t, still neglecting tangential and capillary forces. The relationship (4) is replaced by

$$-p\mathbf{n} + 2\mu\mathbf{D}(\mathbf{v})\mathbf{n} = -P\mathbf{n} \quad \text{on } \Gamma_t, \ t \in (0, T), \tag{12}$$

where P is the pressure in the gas. For instance, consider the experiment of Figure 5 where the cavity is being filled with liquid. The gas present in the cavity at initial time can either escape if a *valve* is placed at the end of the cavity (in which case the gas does exert very little resistance on the liquid) or be trapped in the cavity. When a bubble of gas is trapped by the liquid, the gas pressure prevents the bubble to vanish rapidly, as it is the case for vacuum.

The pressure in the gas is assumed to be constant in space in each bubble of gas, that is to say in each connected component of the gas domain. Let $k(t)$ be the number of bubbles of gas at time t and let $B_i(t)$ denote the domain occupied by the bubble number i (the i-th connected component). Let $P_i(t)$ denote the pressure in $B_i(t)$. At initial time, $P_i(0)$ is constant in each bubble i. The gas is assumed to be an ideal gas. If $V_i(t)$ is defined as the volume of $B_i(t)$, the pressure in each bubble at time t is thus computed by using the law of ideal gases at constant temperature:

$$P_i(t)V_i(t) = \text{constant} \quad i = 1, \ldots, k(t). \tag{13}$$

The above relationship is an expression of the conservation of the number of molecules of trapped gas (gas that cannot escape through a valve) between time t and $t + \delta t$. However, this simplified model requires the tracking of the position of the bubbles of gas between two time steps.

When δt is small enough, three situations appear between two time steps: first, a single bubble remains a single bubble; or a bubble splits into two bubbles, or two bubbles merge into one. Combinations of these three situations may appear.

For instance, in the case of a single bubble, if the pressure $P(t)$ in the bubble at time t and the volumes $V(t)$ and $V(t + \delta t)$ are known, the gas pressure at time $t + \delta t$ is easily computed from the relation $P(t + \delta t)V(t + \delta t) = P(t)V(t)$. The other cases are described at the discrete level in the following. Details can be found in [CPR05].

The additional unknowns in our model are the bubbles of gas $B_i(t)$ and the constant pressure $P = P_i(t)$ in the bubble of gas number i. The equations (1)–(3) are to be solved together with (12), (13).

3.2 Modification of the Numerical Method

The tracking of the bubbles of gas and the computation of their internal pressure introduce an additional step in our time splitting scheme. This procedure is inserted between the advection step (6) and the diffusion step (7), (8), in order to compute an approximation of the pressure to plug into (12).

Let us denote by k^n, P_i^n, B_i^n, $i = 1, 2, \ldots, k^n$, the approximations of k, P_i, B_i, $i = 1, 2, \ldots, k$, respectively at time t^n. Let $\xi(t)$ be a *bubble numbering*

function, defined as negative in the liquid region Ω_t and equal to i in bubble $B_i(t)$. The approximations k^{n+1}, P_i^{n+1}, B_i^{n+1}, $i = 1, 2, \ldots, k^{n+1}$ and ξ^{n+1} are computed as follows.

Numbering of the Bubbles of gas

Given the new liquid domain Ω^{n+1}, the key point is to find the number of bubbles k^{n+1} (that is to say the number of connected components) and the bubbles B_i^{n+1}, $i = 1, \ldots, k^{n+1}$. Given a point P in the gas domain $\Lambda \setminus \Omega^{n+1}$, we search for a function u such that $-\Delta u = \delta_P$ in $\Lambda \setminus \Omega^{n+1}$, with $u = 0$ on Ω^{n+1} and u continuous. Since the solution u to this problem is strictly positive in the connected component containing point P and vanishes outside, the first bubble is found. The procedure is repeated iteratively until all the bubbles are recognized. The algorithm is written as follows:

Set $k^{n+1} = 0$, $\xi^{n+1} = 0$ in $\Lambda \setminus \Omega^{n+1}$ and $\xi^{n+1} = -1$ in Ω^{n+1}, and $\Theta^{n+1} = \{x \in \Lambda : \xi^{n+1}(x) = 0\}$.
While $\Theta^{n+1} \neq \emptyset$, do:
1. Choose a point P in Θ^{n+1};
2. Solve the following problem: Find $u : \Lambda \to \mathbb{R}$ which satisfies:

$$\begin{cases} -\Delta u = \delta_P, & \text{in } \Theta^{n+1}, \\ \quad u = 0, & \text{in } \Lambda \setminus \Theta^{n+1}, \\ \quad [u] = 0, & \text{on } \partial\Theta^{n+1}, \end{cases} \qquad (14)$$

where δ_P is Dirac delta function at point P, $[u]$ is the jump of u through $\partial\Theta^{n+1}$;
3. Increase the number of bubbles k^{n+1} at time t^{n+1}: $k^{n+1} = k^{n+1} + 1$;
4. Define the bubble of gas number k^{n+1}: $B_{k^{n+1}}^{n+1} = \{x \in \Theta^{n+1} : u(x) \neq 0\}$;
5. Update the bubble numbering function $\xi^{n+1}(x) = k^{n+1}$, for all $x \in B_{k^{n+1}}^{n+1}$;
6. Update Θ^{n+1} for the next iteration: $\Theta^{n+1} = \{x \in \Lambda : \xi^{n+1}(x) = 0\}$.

The cost of this original numbering algorithm is bounded by the cost of solving k^{n+1} times a Poisson problem in the gas domain. The corresponding CPU time used to solve the Poisson problems is usually less than 10 percent of the total CPU time. This numbering algorithm is implemented on the finite element mesh. The Poisson problems (14) are solved on $\mathcal{T}_h$, using standard continuous, piecewise linear finite elements.

Computation of the Pressure in the Gas

Once the connected components of gas are numbered, an approximation P_i^{n+1} of the constant pressure in bubble i at time t^{n+1} has to be computed with (13). In the case of a single bubble in the liquid, (13) yields $P_1^{n+1} V_1^{n+1} = P_1^n V_1^n$. In the case when two bubbles merge, this relation becomes $P_1^{n+1} V_1^{n+1} = P_1^n V_1^n + P_2^n V_2^n$. When a bubble B_1^n splits onto two,

each of its parts at time t^n contributes to bubbles B_1^{n+1} and B_2^{n+1}. The volume fraction of bubble B_1^n which contributes to bubble B_j^{n+1} is noted $V_{1,j}^{n+1/2}$, $j = 1, 2$. The pressure in the bubble B_j^{n+1} is computed by taking into account the compression/decompression of the two fractions of bubbles $P_j^{n+1} = P_1^n V_{1,j}^{n+1/2}/V_j^{n+1}$, $j = 1, 2$.

Details of the implementation require to take into account several situations, when two bubbles at time t^n and t^{n+1} do or do not intersect between two time steps, and are detailed in [CPR05]. The value of the pressure can be inserted as a boundary term in (9) for the resolution of the generalized Stokes problem (7), (8).

Remark 3. By using the divergence theorem in the variational formulation (9) and the fact that P^{n+1} is piecewise constant, the integral on the free surface Γ_h^{n+1} is transformed into an integral on Ω_h^{n+1} and, therefore, an approximation of the normal vector **n** is not explicitly needed.

3.3 Numerical Results

Numerical results are presented here for mold filling simulations in order to show the influence of the gas pressure and to compare with results in Section 2.4.

The same S-shaped channel is initially filled with gas at atmospheric pressure $P = 101300$ Pa. A valve is located at the upper extremity of the channel allowing gas to escape. Numerical results (cf. Figure 6) show the persistence of the bubbles. The CPU time for the simulations is approximately 344 min with the bubbles computations (to compare with 319 min in Section 2).
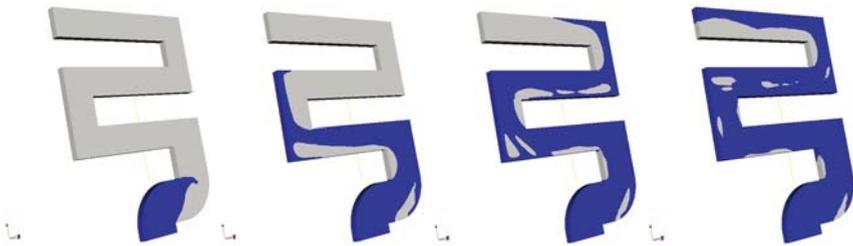

Fig. 6. S-shaped channel: 3D results when the cavity is initially filled with compressible gas at atmospheric pressure. Time equals 8.0 ms, 26.0 ms, 44.0 ms and 53.9 ms.

4 Extension to the Modeling of Incompressible Liquid-Compressible Gas Two-Phases Flows with Surface Tension Effects

4.1 Extension of the Model

Surface tension effects are usually neglected for high Reynolds numbers. However, for creeping flows (with low Reynolds number and high Capillary number), the surface tension effects become relevant. The model presented in Section 3 is extended, so that tangential and capillary forces are still neglected on the free surface, but the normal forces due to the surface tension effects are added. Details can be found in [Cab06]. The relationship (12) is replaced by

$$-p\mathbf{n} + 2\mu\mathbf{D}(\mathbf{v})\mathbf{n} = -P\mathbf{n} + \sigma\kappa\mathbf{n} \quad \text{on } \Gamma_t, \ t \in (0, T), \tag{15}$$

where $\kappa = \kappa(x, t)$ is the mean curvature of the interface Γ_t at point $x \in \Gamma_t$ and σ is a constant surface tension coefficient which depends on both media on each side of the interface (namely the liquid and the gas). The *continuum surface force* (CSF) model, see, e.g., [BKZ92, RK98, WKP99], is considered for the modeling of surface tension effects.

4.2 Modification of the Numerical Method

The relationship (15) on the interface requires the computation of the curvature κ and the normal vector $\mathbf{n}$. An additional step is added in the time splitting scheme to compute these two unknowns before the diffusion part. The approximations κ^{n+1} and $\mathbf{n}^{n+1}$ of κ and $\mathbf{n}$ respectively are computed at time t^{n+1} on the interface Γ^{n+1} as follows.

Since the characteristic function φ^{n+1} is not smooth, it is first mollified, see, e.g., [WKP99], in order to obtain a smoothed approximation $\tilde{\varphi}^{n+1}$, such that the liquid-gas interface Γ^{n+1} is given by the level line $\{x \in \Lambda : \tilde{\varphi}^{n+1}(x) = 1/2\}$, with $\tilde{\varphi}^{n+1} < 1/2$ in the gas domain and $\tilde{\varphi}^{n+1} > 1/2$ in the liquid domain. The smoothed characteristic function $\tilde{\varphi}^{n+1}$ is obtained by convolution of φ^{n+1} with the fourth-order kernel function K_ε described in [WKP99]:

$$\tilde{\varphi}^{n+1}(x) = \int_\Lambda \varphi^{n+1}(y) K_\varepsilon(x - y) \, dy \quad \forall x \in \Lambda. \tag{16}$$

The smoothing of φ^{n+1} is performed only in a layer around the free surface. The parameter ε is the smoothing parameter that describes the size of the support of K_ε, i.e. the size of the smoothing layer around the interface. At each time step, the normal vector $\mathbf{n}^{n+1}$ and the curvature κ^{n+1} on the liquid-gas interface are given respectively by $\mathbf{n}^{n+1} = -\nabla\tilde{\varphi}^{n+1}/\|\nabla\tilde{\varphi}^{n+1}\|$ and $\kappa^{n+1} = -\operatorname{div}(\nabla\tilde{\varphi}^{n+1}/\|\nabla\tilde{\varphi}^{n+1}\|)$, see, e.g., [OF01, Set96].

Instead of using the structured grid of cells to compute the curvature, see, e.g., [AMS04, SZ99], the computation of κ^{n+1} is performed on the finite element mesh, in order to use the variational framework of finite elements.

The normal vector $\mathbf{n}_h^{n+1}$ is given by the normalized gradient of $\tilde{\varphi}_h^{n+1}$ at each grid point P_j, $j = 1, \ldots, M$ where M denotes the number of nodes in the finite element discretization. Details can be found in [Cab06]. The curvature κ^{n+1} is approximated by its L^2-projection on the piecewise linear finite elements space with *mass lumping* and is denoted by κ_h^{n+1}. The basis functions of the piecewise linear finite element space associated to each node P_j in the cavity being denoted by ψ_{P_j}, κ_h^{n+1} is given by the relation

$$\int_\Lambda \kappa_h^{n+1} \psi_{P_j} \, d\mathbf{x} = \int_\Lambda -\operatorname{div} \frac{\nabla \tilde{\varphi}_h^{n+1}}{\|\nabla \tilde{\varphi}_h^{n+1}\|} \psi_{P_j} \, d\mathbf{x}, \quad \text{for all } j = 1, \ldots, M.$$

The left-hand side of this relation is computed with *mass lumping*, while the right-hand side is integrated by parts. Explicit values of the curvature of the level lines of $\tilde{\varphi}_h^{n+1}$ are obtained at the vertices of the finite element mesh being in a layer around the free surface. The restriction of κ_h^{n+1} to the nodes lying on Γ_h^{n+1} is used to compute (15).

4.3 Numerical Results

We consider a bubble of gas at the bottom of a cylinder filled with liquid, under gravity forces. The bubble rises and reaches an upper free surface between water and air, see Figure 7. The physical constants are $\mu = 0.01 \, \text{kg/(ms)}$, $\rho = 1000 \, \text{kg/m}^3$ and $\sigma = 0.0738 \, \text{N/m}$. The mesh made out of 115200 tetrahedra. The size of the cells of the structured mesh used for advection step is approximately 5 to 10 times smaller than the size of the finite elements and
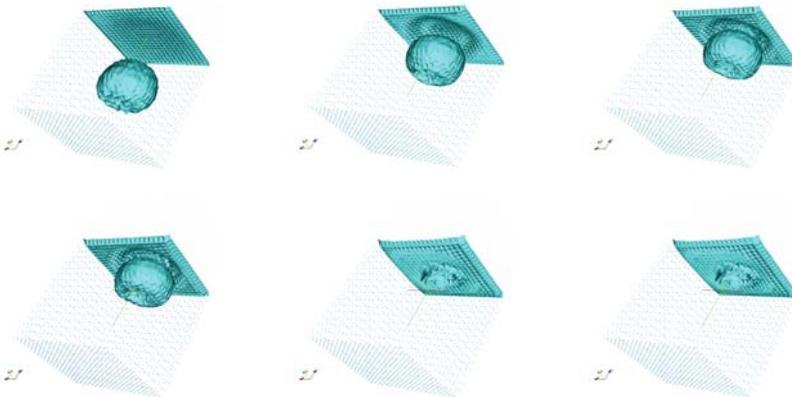

Fig. 7. Three-dimensional rising bubble under a free surface: Representation of the gas domain at times $t = 100.0, 200.0, 230.0, 240.0., 300.0$ and 320.0 ms (left to right, top to bottom).

the time step is chosen such that the CFL number is approximately one. The smoothing parameter is $\varepsilon = 0.005$. The CPU time for this computation is approximately 20 hours to achieve 1000 time steps.

5 Extension to the Modeling of Viscoelastic Flows with a Free Surface

5.1 Extension of the Model

The total stress tensor for incompressible viscoelastic fluids is, by definition, the sum of a Newtonian part $2\mu\mathbf{D}(\mathbf{v}) - p\mathbf{I}$ and a non-Newtonian part denoted by $\boldsymbol{\sigma} : Q_T \to \mathbb{R}^{3\times3}$. Owning this decomposition, the system (1)–(2) becomes

$$\rho\frac{\partial\mathbf{v}}{\partial t} + \rho(\mathbf{v}\cdot\nabla)\mathbf{v} - 2\operatorname{div}\left(\mu\mathbf{D}(\mathbf{v}) + \boldsymbol{\sigma}\right) + \nabla p = \mathbf{f} \quad \text{in } Q_T, \qquad (17)$$

$$\operatorname{div}\mathbf{v} = 0 \quad \text{in } Q_T. \qquad (18)$$

The simplest constitutive (or closure) equation for the extra-stress $\boldsymbol{\sigma}$, namely the Oldroyd-B model [Old50], is chosen to supplement the above system

$$\boldsymbol{\sigma} + \lambda\left(\frac{\partial\boldsymbol{\sigma}}{\partial t} + (\mathbf{v}\cdot\nabla)\boldsymbol{\sigma} - (\nabla\mathbf{v})\boldsymbol{\sigma} - \boldsymbol{\sigma}(\nabla\mathbf{v})^T\right) = 2\eta\mathbf{D}(\mathbf{v}) \quad \text{in } Q_T. \qquad (19)$$

Here $\lambda > 0$ is the relaxation time (the time for the stress to return to zero under constant-strain condition) and $\eta > 0$ is the polymer viscosity. The extra-stress $\boldsymbol{\sigma}$ has to be imposed only at the inflow. For more details, we refer to [BPL06].

Remark 4. The numerical procedures described in this section can be extended to more general deterministic models such as Phan-Thien Tanner [PTT77], Giesekus [Gie82] and stochastic models such as, e.g., FENE [War72], FENE-P [BDJ80]. Two-dimensional computations of free surface flows with FENE dumbbells have been performed in [GLP03].

5.2 Modification of the Numerical Procedure

The convective term in (19) is treated in the same fashion as (5). Continuous, piecewise linear finite elements are considered to approximate the extra-stress tensor $\boldsymbol{\sigma}$ and an EVSS (*Elastic Viscous Split Stress*) procedure [FGP97, BPS01, PR01] is used in order to obtain a stable algorithm even if the solvent viscosity μ vanishes.

Advection Step. Together with (5), solve between the times t^n and t^{n+1}

$$\frac{\partial\boldsymbol{\sigma}}{\partial t} + (\mathbf{u}\cdot\nabla)\boldsymbol{\sigma} = 0 \qquad (20)$$

with initial conditions given by the value of the tensor $\boldsymbol{\sigma}$ at time t^n. This step is also solved using the characteristics method on the structured grid, see Figure 2, using the relation $\boldsymbol{\sigma}^{n+1/2}(x+\delta t^n\mathbf{v}^n(x)) = \boldsymbol{\sigma}^n(x)$. As for the velocity and volume fraction of liquid, the extra-stress tensor $\boldsymbol{\sigma}^{n+1/2}$ is computed on the structured grid of cells (ijk) leading to values $\boldsymbol{\sigma}^{n+1/2}_{ijk}$. Then, values are interpolated at the nodes of the finite element mesh using the same kind of formula as in (10), which yields the continuous, piecewise linear extra-stress $\boldsymbol{\sigma}^{n+1/2}_h$.

Diffusion Step. The diffusion step consists in solving the so-called three-fields Stokes problem on the finite element mesh. Following the EVSS method, we define a new extra-tensor $\mathbf{B}^{n+1/2}_h : \Omega^{n+1}_h \to \mathbb{R}^{3\times3}$ as the L^2-projection into the finite element space of the predicted deformation tensor $\mathbf{D}(\mathbf{v}^{n+1/2}_h)$, i.e.

$$\int_{\Omega^{n+1}_h} \mathbf{B}^{n+1/2}_h : \mathbf{E}_h \, dx = \int_{\Omega^{n+1}_h} \mathbf{D}(\mathbf{v}^{n+1/2}_h) : \mathbf{E}_h \, dx,$$

for all test functions $\mathbf{E}_h$. Then (9) is modified to take explicitly into account the term coming from the extra-stress tensor. The extra term

$$2\int_{\Omega^{n+1}_h} \eta\mathbf{D}(\mathbf{v}^{n+1}_h) : \mathbf{D}(\mathbf{w}_h)\,dx - 2\int_{\Omega^{n+1}_h} \eta\mathbf{B}^{n+1/2}_h : \mathbf{D}(\mathbf{w}_h)\,dx,$$

which vanishes at continuous level, is also added. Thus, the weak formulation (9) becomes, find the piecewise linear finite element approximations $\mathbf{v}^{n+1}_h$ and p^{n+1}_h such that $\mathbf{v}^{n+1}_h$ satisfies the essential boundary conditions on the boundary of the cavity Λ and such that

$$\rho \int_{\Omega^{n+1}_h} \frac{\mathbf{v}^{n+1}_h - \mathbf{v}^{n+1/2}_h}{\delta t^n} \cdot \mathbf{w}_h \, dx + 2\int_{\Omega^{n+1}_h} (\mu+\eta)\mathbf{D}(\mathbf{v}^{n+1}_h) : \mathbf{D}(\mathbf{w}_h)\,dx$$

$$-\int_{\Omega^{n+1}_h} p^{n+1}_h \operatorname{div}\mathbf{w}_h\,dx + \int_{\Omega^{n+1}_h} \boldsymbol{\sigma}^{n+1/2}_h : \mathbf{D}(\mathbf{w}_h)\,dx$$

$$-2\int_{\Omega^{n+1}_h} \eta\mathbf{B}^{n+1/2}_h : \mathbf{D}(\mathbf{w}_h)\,dx - \int_{\Omega^{n+1}_h} \mathbf{f}\cdot\mathbf{w}_h\,dx - \int_{\Omega^{n+1}_h} q_h \operatorname{div}\mathbf{v}^{n+1}_h\,dx = 0,$$

$$(21)$$

for all test functions $\mathbf{w}_h, q_h$. Once the velocity $\mathbf{v}^{n+1}_h$ is computed, the extra-stress is recovered using (19). More precisely the continuous, piecewise linear extra-stress $\boldsymbol{\sigma}^{n+1}_h$ satisfies the prescribed boundary conditions at inflow and

$$\int_{\Omega_h^{n+1}} \boldsymbol{\sigma}_h^{n+1} : \boldsymbol{\tau}_h \, d\mathbf{x} + \lambda \int_{\Omega_h^{n+1}} \frac{\boldsymbol{\sigma}_h^{n+1} - \boldsymbol{\sigma}_h^{n+1/2}}{\delta t^n} : \boldsymbol{\tau}_h \, d\mathbf{x}$$

$$= 2\eta \int_{\Omega_h^{n+1}} \mathbf{D}(\mathbf{v}_h^{n+1}) : \boldsymbol{\tau}_h \, d\mathbf{x}$$

$$+ \lambda \int_{\Omega_h^{n+1}} \left((\nabla \mathbf{v}_h^{n+1}) \boldsymbol{\sigma}_h^{n+1/2} + \boldsymbol{\sigma}_h^{n+1/2} (\nabla \mathbf{v}_h^{n+1})^T \right) : \boldsymbol{\tau}_h \, d\mathbf{x}, \quad (22)$$

for all test functions $\boldsymbol{\tau}_h$. Finally, the fields $\mathbf{u}_h^{n+1}$ and $\boldsymbol{\sigma}_h^{n+1}$ are interpolated at the center of the cells C_{ijk}.

Theoretical investigations for a simplified problem without advection and free surface have been performed in [BCP07]. Using an implicit function theorem, existence of a solution and convergence of the finite element scheme have been obtained. We refer to [BCP06b, BCP06a] for an extension to the stochastic Hookean dumbbells model.

5.3 Numerical Results

Two different simulations are provided here, the *buckling* of a jet and the stretching of a filament. In the first simulation, different behaviors between Newtonian and viscoelastic fluids are observed and the elastic effect of the relaxation time λ is pointed out. In the second simulation, *fingering instabilities* can be observed, which corresponds to experiments. More details and test cases can be found in [BPL06].

Jet buckling

The transient flow of a jet of diameter $d = 0.005$ m, injected into a parallelepiped cavity of width 0.05 m, depth 0.05 m and height 0.1 m, is reproduced. Liquid enters from the top of the cavity with vertical velocity $U = 0.5$ m/s. The fluids parameters are given in Table 1, the effects of surface tension being not considered.

The finite element mesh has 503171 vertexes and 2918760 tetrahedra. The cells size is 0.0002 m and the time step is 0.001 s thus the CFL number of the cells is 2.5. A comparison of the shape of the jet with Newtonian flow is shown in Figure 8. This computation takes 64 hours on a AMD Opteron CPU with 8Gb memory. The elastic effects in the liquid are clearly observed: when the viscoelastic jet starts to buckle, the Newtonian jet has already produced many

Table 1. Jet buckling. Liquid parameters.

	ρ [kg/m^3]	μ [Pa·s]	η [Pa·s]	λ [s]	De $= \lambda U/d$
Newtonian	1030	10.3	0	0	0
Viscoelastic	1030	1.03	9.27	1	100

Fig. 8. Jet buckling in a cavity. Shape of the jet at time $t = 0.125$ s (col. 1), $t = 0.45$ s (col. 2), $t = 0.6$ s (col. 3), $t = 0.9$ s (col. 4), $t = 1.15$ s (col. 5), $t = 1.6$ s (col. 6), Newtonian fluid (row 1), viscoelastic fluid (row 2).

folds. For a discussion on the condition for a jet to buckle and comparison with results obtained in [TMC$^+$02], we refer to [BPL06].

Fingering instabilities

The numerical model is capable to reproduce fingering instabilities, as reported in [RH99, BRLH02, MS02, DLCB03] for non-Newtonian flows. The flow of an Oldroyd-B fluid contained between two parallel coaxial circular disks with radius $R_0 = 0.003$ m is considered. At the initial time, the distance between the two end-plates is $L_0 = 0.00015$ m and the liquid is at rest. Then, the top end-plate is moved vertically with velocity $L_0\dot{\varepsilon}_0 e^{\dot{\varepsilon}_0 t}$ where $\dot{\varepsilon}_0 = 4.68\,\mathrm{s}^{-1}$. The liquid parameters are $\rho = 1030\,\mathrm{kg/m}^3$, $\mu = 9.15\,\mathrm{Pa \cdot s}$, $\eta = 25.8\,\mathrm{Pa \cdot s}$, $\lambda = 0.421\,\mathrm{s}$. Following [MS02, Section 4.4], since the aspect ratio R_0/L_0 is equal to 20, the Weissenberg number $\mathrm{We} = \mathrm{De}R_0^2/L_0^2$ is large.

The finite element mesh has 50 vertexes along the radius and 25 vertexes along the height, thus the mesh size is 0.00006 m. The cells size is 0.00001 m and the initial time step is $\delta t = 0.01$ s thus the CFL number of the cells is initially close to one. The shape of the filament is reported in Figure 9 and 2D cuts in the middle of the height are reported in Figure 10. Fingering instabilities can be observed from the very beginning of the stretching, leading to branched structures, as described in [MS02, BRLH02, DLCB03]. These instabilities are essentially elastic, without surface tension effects [RH99]. Clearly, such complex shapes cannot be obtained using Lagrangian models, the mesh distortion being too large.
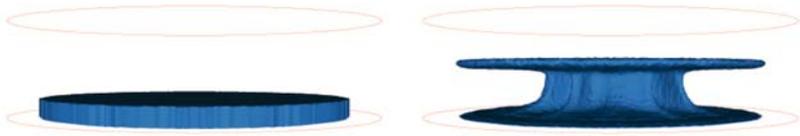
Fig. 9. Fingering instabilities. Shape of the liquid region at times $t = 0$ s (left) and $t = 0.745$ s (right).
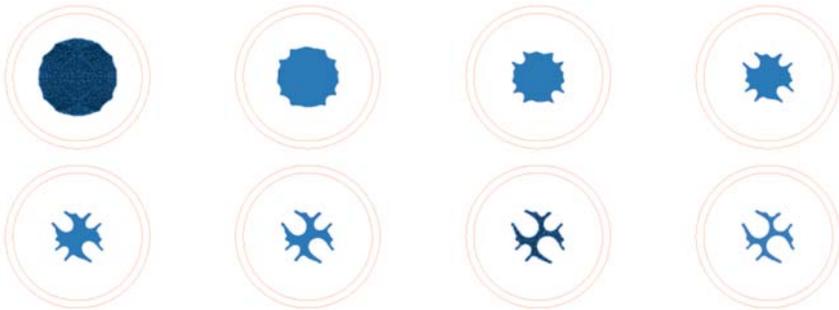

Fig. 10. Fingering instabilities. Horizontal cuts through the middle of the liquid region at times $t = 0.119$ s, $t = 0.245$ s, $t = 0.364$ s, $t = 0.49$ s (first row) and times $t = 0.609$ s, $t = 0.735$ s, $t = 0.854$ s, $t = 0.98$ s (second row).

6 Conclusions

An efficient computational model for the simulation of two-phases flows has been presented. It allows to consider both Newtonian and non-Newtonian flows. It relies on an Eulerian framework and couples finite element techniques with a forward characteristics method. Numerical results illustrate the large range of applications covered by the model. Extensions are being investigated (1) to couple viscoelastic and surface tension effects, (2) to reduce the CPU time required to solve Stokes problems, and (3) to improve the reconstruction of the interface and the computation of surface tension effects.

References

[AMS04]   E. Aulisa, S. Manservisi, and R. Scardovelli. A surface marker algorithm coupled to an area-preserving marker redistribution method for three-dimensional interface tracking. *J. Comput. Phys.*, 197(2):555–584, 2004.

[BCP06a] A. Bonito, Ph. Clément, and M. Picasso. Finite element analysis of a simplified stochastic Hookean dumbbells model arising from viscoelastic flows. *M2AN Math. Model. Numer. Anal.*, 40(4):785–814, 2006.

[BCP06b] A. Bonito, Ph. Clément, and M. Picasso. Mathematical analysis of a simplified Hookean dumbbells model arising from viscoelastic flows. *J. Evol. Equ.*, 6(3):381–398, 2006.

[BCP07]   A. Bonito, Ph. Clément, and M. Picasso. Mathematical and numerical analysis of a simplified time-dependent viscoelastic flow. *Numer. Math.*, 107(2):213–255, 2007.

[BDJ80]   R. B. Bird, N. L. Dotson, and N. L. Johnson. Polymer solution rheology based on a finitely extensible bead-spring chain model. *J. Non-Newtonian Fluid Mech.*, 7:213–235, 1980.

[BKZ92]   J. U. Brackbill, D. B. Kothe, and C. Zemach. A continuum method for modeling surface tension. *J. Comput. Phys.*, 100:335–354, 1992.

[BPL06]   A. Bonito, M. Picasso, and M. Laso. Numerical simulation of 3D viscoelastic flows with free surfaces. *J. Comput. Phys.*, 215(2):691–716, 2006.

[BPS01]   J. Bonvin, M. Picasso, and R. Stenberg. GLS and EVSS methods for a three-field Stokes problem arising from viscoelastic flows. *Comput. Methods Appl. Mech. Engrg.*, 190(29–30):3893–3914, 2001.

[BRLH02] A. Bach, H. K. Rasmussen, P.-Y. Longin, and O. Hassager. Growth of non-axisymmetric disturbances of the free surface in the filament stretching rheometer: experiments and simulation. *J. Non-Newtonian Fluid Mech.*, 108:163–186, 2002.

[Cab05]   A. Caboussat. Numerical simulation of two-phase free surface flows. *Arch. Comput. Methods Engrg.*, 12(2):165–210, 2005.

[Cab06]   A. Caboussat. A numerical method for the simulation of free surface flows with surface tension. *Comput. & Fluids*, 35(10):1205–1216, 2006.

[CPR05]   A. Caboussat, M. Picasso, and J. Rappaz. Numerical simulation of free surface incompressible liquid flows surrounded by compressible gas. *J. Comput. Phys.*, 203(2):626–649, 2005.

[CR05]   A. Caboussat and J. Rappaz. Analysis of a one-dimensional free surface flow problem. *Numer. Math.*, 101(1):67–86, 2005.

[DLCB03] D. Derks, A. Lindner, C. Creton, and D. Bonn. Cohesive failure of thin layers of soft model adhesives under tension. *J. Appl. Phys.*, 93(3):1557–1566, 2003.

[FCD⁺06] M. M. Francois, S. J. Cummins, E. D. Dendy, D. B. Kothe, J. M. Sicilian, and M. W. Williams. A balanced-force algorithm for continuous and sharp interfacial surface tension models within a volume tracking framework. *J. Comput. Phys.*, 213(1):141–173, 2006.

[FF92]   L. P. Franca and S. L. Frey. Stabilized finite element method: II. The incompressible Navier–Stokes equations. *Comput. Methods Appl. Mech. Engrg.*, 99:209–233, 1992.

[FGP97]   M. Fortin, R. Guénette, and R. Pierre. Numerical analysis of the modi-
           fied EVSS method. *Comput. Methods Appl. Mech. Engrg.*, 143(1–2):79–
           95, 1997.
[Gie82]   H. Giesekus. A simple constitutive equation for polymer fluids based
           on the concept of deformation-dependent tensorial mobility. *J. Non-
           Newtonian Fluid Mech.*, 11(1–2):69–109, 1982.
[Glo03]   R. Glowinski. Finite element methods for incompressible viscous flow. In
           P. G. Ciarlet and J.-L. Lions, editors, *Handbook of Numerical Analysis,
           Vol. IX*, pages 3–1176. North-Holland, Amsterdam, 2003.
[GLP03]   E. Grande, M. Laso, and M. Picasso. Calculation of variable-topology
           free surface flows using CONNFFESSIT. *J. Non-Newtonian Fluid
           Mech.*, 113(2):123–145, 2003.
[Mar90]   G. I. Marchuk. Splitting and alternating direction methods. In P. G.
           Ciarlet and J.-L. Lions, editors, *Handbook of Numerical Analysis, Vol.
           I*, pages 197–462. North-Holland, Amsterdam, 1990.
[Mau96]   B. Maury. Characteristics ALE method for the 3D Navier-Stokes equa-
           tions with a free surface. *Int. J. Comput. Fluid Dyn.*, 6:175–188, 1996.
[MPR99]   V. Maronnier, M. Picasso, and J. Rappaz. Numerical simulation of free
           surface flows. *J. Comput. Phys.*, 155:439–455, 1999.
[MPR03]   V. Maronnier, M. Picasso, and J. Rappaz. Numerical simulation of three
           dimensional free surface flows. *Internat. J. Numer. Methods Fluids*,
           42(7):697–716, 2003.
[MS02]   G. H. McKinley and T. Sridhar. Filament-stretching rheometry of com-
           plex fluids. *Ann. Rev. Fluid Mech.*, 34:375–415, 2002.
[NW76]   W. F. Noh and P. Woodward. SLIC (Simple Line Interface Calcula-
           tion). In A. I. van de Vooren and P. J. Zandbergen, editors, *Proc. of the
           5th International Conference on Numerical Methods in Fluid Dynam-
           ics (Enschede, 1976)*, volume 59 of *Lectures Notes in Physics*, pages
           330–340, Springer-Verlag, Berlin, 1976.
[OF01]   S. Osher and R. P. Fedkiw. Level set methods: An overview and some
           recent results. *J. Comput. Phys.*, 169:463–502, 2001.
[Old50]   J. G. Oldroyd. On the formulation of rheological equations of state.
           Proc. Roy. Soc. London. Ser. A., 200(1063):523–541, 1950.
[Pir89]   O. Pironneau. *Finite Element Methods for Fluids.* Wiley, Chichester,
           1989.
[PR01]   M. Picasso and J. Rappaz. Existence, a priori and a posteriori error
           estimates for a nonlinear three-field problem arising from Oldroyd-B
           viscoelastic flows. *M2AN Math. Model. Numer. Anal.*, 35(5):879–897,
           2001.
[PTT77]   N. Phan-Thien and R.I. Tanner. A new constitutive equation derived
           from network theory. *J. Non-Newtonian Fluid Mech.*, 2(4):353–365,
           1977.
[RDG+00]   M. Rappaz, J. L. Desbiolles, C. A. Gandin, S. Henry, A. Semoroz, and
           P. Thevoz. Modelling of solidification microstructures. *Mater. Sci. Fo-
           rum*, 329(3):389–396, 2000.
[RH99]   H. K. Rasmussen and O. Hassager. Three-dimensional simulations of
           viscoelastic instability in polymeric filaments. *J. Non-Newtonian Fluid
           Mech.*, 82:189–202, 1999.
[RK98]   W. J. Rider and D. B. Kothe. Reconstructing volume tracking. *J.
           Comput. Phys.*, 141:112–152, 1998.

[Set96]   J. A. Sethian. *Level Set Methods, Evolving Interfaces in Geometry, Fluid Mechanics, Computer Vision, and Material Science.* Monographs on Applied and Computational Mathematics. Cambridge University Press, 1996.

[SW04]   M. Shashkov and B. Wendroff. The repair paradigm and application to conservation laws. *J. Comput. Phys.*, 198(1):265–277, 2004.

[SZ99]   R. Scardovelli and S. Zaleski. Direct numerical simulation of free surface and interfacial flows. *Ann. Rev. Fluid Mech.*, 31:567–603, 1999.

[TMC$^+$02] M. F. Tomé, N. Mangiavacchi, J. A. Cuminato, A. Castelo, and S. McKee. A finite difference technique for simulating unsteady viscoelastic free surface flows. *J. Non-Newtonian Fluid Mech.*, 106:61–106, 2002.

[War72]   H. R. Warner. Kinetic theory and rheology of dilute suspensions of finitely extendible dumbbells. *Ind. Eng. Chem. Fund.*, 11:379–387, 1972.

[WKP99]   M. W. Williams, D. B. Kothe, and E. G. Puckett. Accuracy and convergence of continuum surface tension models. In *Fluid Dynamics at Interfaces (Gainesville, FL, 1998)*, pages 294–305. Cambridge University Press, 1999.

Modelling and Simulating the Adhesion and Detachment of Chondrocytes in Shear Flow

Jian Hao[1], Tsorng-Whay Pan[1], and Doreen Rosenstrauch[2]

[1] Department of Mathematics, University of Houston, Houston, TX 77204-3008, USA jianh@math.uh.edu, pan@math.uh.edu
[2] The Texas Heart Institute and the University of Texas Health Science Center at Houston, Houston, TX 77030, USA Doreen.Rosenstrauch@uth.tmc.edu

1 Introduction

Chondrocytes are typically studied in the environment where they normally reside such as the joints in hips, intervertebral disks or the ear. For example, in [SKE+99], the effect of seeding duration on the strength of chondrocyte adhesion to articulate cartilage has been studied in shear flow chamber since such adhesion may play an important role in the repair of articular defects by maintaining cells in positions where their biosynthetic products can contribute to the repair process. However, in this investigation, we focus mainly on the use of auricular chondrocytes in cardiovascular implants. They are abundant, easily and efficiently harvested by a minimally invasive technique. Auricular chondrocytes have ability to produce collagen type-II and other important extracellular matrix constituents; this allows them to adhere strongly to the artificial surfaces. They can be genetically engineered to act like endothelial cells so that the biocompatibility of cardiovascular prothesis can be improved. Actually in [SBBR+02], genetically engineered auricular chondrocytes can be used to line blood-contacting luminal surfaces of left ventricular assist device (LVAD) and a chondrocyte-lined LVAD has been planted into the tissue-donor calf and the results *in vivo* have proved the feasibility of using autologous auricular chondrocytes to improve the biocompatibility of the blood-biomaterial interface in LVADs and cardiovascular prothesis. Therefore, cultured chondrocytes may offer a more efficient and less invasive means of covering artificial surface with a viable and adherent cell layer.

In this chapter, we first develop the model of the adhesion of chondrocytes to the artificial surface and then combine the resulting model with a Lagrange multiplier based fictitious domain method to simulate the detachment of chondrocyte cells in shear flow. The chondrocytes in the simulation are treated as neutrally buoyant rigid particles. As argued in [KS06] that the scaling estimates show that for typical parameter values for cell elasticity, deformations

due to shear flow and lubrication forces are small, the cells can be treated as rigid. The Newtonian incompressible viscous flow is modeled by the Navier–Stokes equations since the inertial effect is crucial for the lift-off of the cells; in most studies of cell adhesion, the Stokes flow is considered since the rolling of cells on the surface and then the capture of cells, like white blood cells, are the main interest, e.g., see [KS06, KH01, SZD03].

2 Model for Cell Adhesion

Cell adhesion to the extracellular matrix (ECM) plays key roles in the assembly of cells into functional multicellular organisms. Chondrocytes produce collagen type-II and other important extracellular matrix constituents; this allows them to adhere strongly to the artificial surfaces. Chondrocyte cells are responsible for the synthesis and maintenance of a viable ECM which is suitably adapted to cope with the physical pressures of its environment. On the lined surface of LVAD, a monolayer of cells formed on the surface was reported in [SBBR[+]02]. Adhesive interactions between chondrocytes and ECM occur via a variety of molecular systems (e.g., see discussion for cell-matrix adhesion in [ZBCAG04]). Zaidel et al. have shown in [ZBCAG04] that cell-associated hyaluronan plays a central role in mediating early stages in the attachment of chondrocytes to the surfaces. Their results indicate that chondrocytes establish, initially, "soft contact" to the surface through a hyaluronan-based coat. The surface adhesion, mediated by the hyaluronan coat, occurs within seconds after the cell first encounters the surface. Then within a few tens of seconds-to-minutes, the hyaluronan-mediated adhesion is replaced by integrin-based interactions which is actually a sequential formation starting from dot-shaped focal complexes (FXs), then changing to focal adhesions (FAs) and finally becoming fibrillar adhesions (FBs).

In [ZBCAG04] chondrocytes were allowed to adhere to a serum coated glass coverslip for 10–25 minutes and exposed to shear flow, they drifted under flow for quite a distance (compared to their diameters) before detachment from the surface. In [SKE[+]99] chondrocytes were seeded on the surface of a piece of articular cartilage for specific durations (5–40 minutes) and then were exposed to shear flow in a flow chamber. It was observed that the increase in resistance to shear stress-induced cell detachment with increasing seeding duration. But in [SBBR[+]02], chondrocytes were allowed to have 24 hours for seeding process on the luminal surfaces of LVADs and then 4 days in incubator for promoting ECM synthesis to maximize the adherence of cells. When using flow loop to precondition seeded cells in order to promote good cell adherence, the cell loss during the process did not exceed 12%. The results in [SKE[+]99] suggest that chondrocytes adhere to the surface mainly via hyaluronan gel and the numbers of integrin-based interactions are not high enough since the durations are comparable to the one used in [ZBCAG04]. But in [SBBR[+]02], the results indicate that adhesions are mainly integrin-mediated interactions
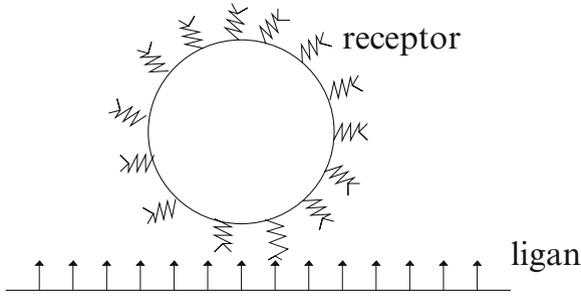
Fig. 1. Model geometry of cell and surface. The surface is covered by ligans and the cell is rigid and covered by receptors distributed randomly.

(FBs) between the members of the integrin family and corresponding ECM proteins, such as collagen type-II and fibronectin [Loe93, GHR04].

To model cell adhesion, Hammer et al. in [KH01, CH96] have developed an adhesive dynamics algorithm, in which adhesion molecules are modeled as linear, Hookean springs, distributed randomly over the particle surface as shown in Figure 1. For chondrocytes, which have microvilli on the cell surface [CKGA03], the randomly distributed receptors as shown in Figure 1 still can be used. The adhesive dynamics algorithm is as follows:

1. All free adhesion molecule receptors in the contact area are tested for formation of binding with the substrate ligand against the probability

$$P_f = 1 - \exp(-k_f n_l \tau),$$

where k_f is the forward reaction rate, n_l is the density of ligans, and the time step is τ. If the generated random number is less than P_f, a bond is established at this time step.

2. All of the currently bound receptors are tested for breakage against the probability

$$P_r = 1 - \exp(-k_r \tau),$$

where k_r is the reverse reaction rate. If the generated random number is less than P_r the bond breaks at this time step.

3. Each existing bond is characterized by the vector $\mathbf{x}_b$ and the force imparted by the spring on the cell is $\mathbf{F}_b = \sigma(|\mathbf{x}_b| - \lambda)\mathbf{u}_b$ with the Hookean spring constant σ, equilibrium length λ and unit directional vector $\mathbf{u}_b = \mathbf{x}_b/|\mathbf{x}_b|$.

4. A summation of the forces from each spring and associated torques is the information that needs to be included in the Newton–Euler equations to study cell interaction with the Navier–Stokes flow discussed in the following section.

The backward reaction rate k_r in [KH01] is given as follows:

$$k_r = k_r^0 \exp\left[\frac{r_0 F}{k_b T}\right],$$

where k_r^0 is the reverse reaction rate when the spring length is at its equilibrium length, r_0 is the reactive compliance, F is the force on the bond and is equal to $\sigma(|\mathbf{x}_b| - \lambda)$, k_b is the Boltzmann constant and T is the temperature. The ratio of the forward reaction rate and the reverse reaction rate at any separation distance is given:

$$\frac{k_f}{k_r} = \frac{k_f^0}{k_r^0} \exp\left[-\frac{\sigma(|\mathbf{x}_b| - \lambda)^2}{2k_b T}\right]$$

where k_f^0 is the forward reaction rate when the spring length is at its equilibrium length. Then the forward reaction rate in [KH01] takes the form

$$k_f = k_f^0 \exp\left[\sigma(|\mathbf{x}_b| - \lambda)(2r_0 - (|\mathbf{x}_b| - \lambda))/(2k_b T)\right].$$

The strength of the adhesion of each cell (or number of bonds formed via the above dynamical process) depends on the densities of ligans and receptors in the contact region between the cell and surface, the area of the contact region, and two reaction rates. For the hyaluronan-mediated adhesion, the above dynamical bonding approach is a good model. But for the integrin-mediated adhesions of chondrocytes reported in [SBBR+02], we can apply the above model to form bonds in a probabilistic way with two different considerations: (1) having larger string constants since focal adhesions and fibrillar adhesions are much stronger than the hyaluronan-mediated adhesions, (2) after the number of bonds reaches its plateau, we switch to the deterministic approach to decide when the bond should be break off by checking whether its length is longer than a chosen one.

3 A Fictitious Domain Formulation for the Fluid/Particle Interaction and Its Discretization

3.1 Fictitious Domain Formulation

In this section we briefly discuss a fictitious formulation for the fluid-particle interaction in shear flow and discretization in space and time developed [PG02]. Let $\Omega \subset \mathbb{R}^2$ be a rectangular region (three-dimensional cases have been discussed in [PG05]). We suppose that Ω is filled with a *Newtonian viscous incompressible* fluid (of *density* ρ_f and *viscosity* μ_f) and contains a moving neutrally buoyant rigid particle B centered at $\mathbf{G} = \{G_1, G_2\}^t$ of *density* ρ_f (see Fig. 2); the flow is modeled by the *Navier–Stokes equations* and the motion of B is described by the *Euler–Newton equations*. We define
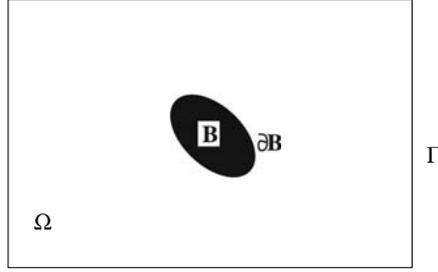
Fig. 2. An example of two-dimensional flow region with one rigid body.

$$W_{\mathbf{g}_0,p} = \{\mathbf{v} \mid \mathbf{v} \in (H^1(\Omega))^2, \ \mathbf{v} = \mathbf{g}_0(t) \text{ on the top and bottom of } \Omega \text{ and}$$
$$\mathbf{v} \text{ is periodic in the } x_1\text{-direction}\},$$

$$W_{0,p} = \{\mathbf{v} \mid \mathbf{v} \in (H^1(\Omega))^2, \ \mathbf{v} = \mathbf{0} \text{ on the top and bottom of } \Omega \text{ and}$$
$$\mathbf{v} \text{ is periodic in the } x_1\text{-direction}\},$$

$$L_0^2 = \left\{ q \mid q \in L^2(\Omega), \ \int_\Omega q \, d\mathbf{x} = 0 \right\},$$

$$\Lambda_0(t) = \{\boldsymbol{\mu} \mid \boldsymbol{\mu} \in (H^1(B(t)))^2, \ \langle\boldsymbol{\mu}, \mathbf{e}_i\rangle_{B(t)} = 0, \ i = 1, 2, \ \langle\boldsymbol{\mu}, \overrightarrow{\mathbf{Gx}^\perp}\rangle_{B(t)} = 0\}$$

with $\mathbf{e}_1 = \{1, 0\}^t$, $\mathbf{e}_2 = \{0, 1\}^t$, $\overrightarrow{\mathbf{Gx}^\perp} = \{-(x_2 - G_2), x_1 - G_1\}^t$ and $\langle\cdot, \cdot\rangle_{B(t)}$ an inner product on $\Lambda_0(t)$ which can be the standard inner product on $(H^1(B(t)))^2$ (see [GPH+01, Section 5] for further information on the choice of $\langle\cdot, \cdot\rangle_{B(t)}$). Then the fictitious domain formulation with distributed Lagrange multipliers for flow around a freely moving neutrally buoyant particle (see [GPHJ99, GPH+01] for detailed discussion of non-neutrally buoyant cases) is as follows:

For a.e. $t > 0$, find $\mathbf{u}(t) \in W_{\mathbf{g}_0,p}$, $p(t) \in L_0^2$, $\mathbf{V_G}(t) \in \mathbb{R}^2$, $\mathbf{G}(t) \in \mathbb{R}^2$, $\omega(t) \in \mathbb{R}$, $\boldsymbol{\lambda}(t) \in \Lambda_0(t)$ such that

$$\rho_f \int_\Omega \left[\frac{\partial\mathbf{u}}{\partial t} + (\mathbf{u} \cdot \boldsymbol{\nabla})\mathbf{u}\right] \cdot \mathbf{v} \, d\mathbf{x} + 2\mu_f \int_\Omega \mathbf{D}(\mathbf{u}) : \mathbf{D}(\mathbf{v}) \, d\mathbf{x} - \int_\Omega p\boldsymbol{\nabla} \cdot \mathbf{v} \, d\mathbf{x}$$

$$-\langle\boldsymbol{\lambda}, \mathbf{v}\rangle_{B(t)} = \rho_f \int_\Omega \mathbf{g} \cdot \mathbf{v} \, d\mathbf{x} + \int_\Omega \mathbf{F} \cdot \mathbf{v} \, d\mathbf{x}, \quad \forall\mathbf{v} \in W_{0,p}, \qquad (1)$$

$$\int_\Omega q\boldsymbol{\nabla} \cdot \mathbf{u}(t)d\mathbf{x} = 0, \quad \forall q \in L^2(\Omega), \qquad (2)$$

$$\langle\boldsymbol{\mu}, \mathbf{u}(t)\rangle_{B(t)} = 0, \quad \forall\boldsymbol{\mu} \in \Lambda_0(t), \qquad (3)$$

$$\frac{d\mathbf{G}}{dt} = \mathbf{V_G}, \qquad (4)$$

$$\mathbf{V_G}(0) = \mathbf{V_G^0}, \quad \omega(0) = \omega^0, \quad \mathbf{G}(0) = \mathbf{G}^0 = \{G_1^0, G_2^0\}^t, \qquad (5)$$

$$\mathbf{u}(\mathbf{x}, 0) = \overline{\mathbf{u}}_0(\mathbf{x}) = \begin{cases} \mathbf{u}_0(\mathbf{x}), & \forall\mathbf{x} \in \Omega \setminus \overline{B(0)}, \\ \mathbf{V_G^0} + \omega^0\{-(x_2 - G_2^0), x_1 - G_1^0\}^t, & \forall\mathbf{x} \in \overline{B(0)}, \end{cases} \qquad (6)$$

where $\mathbf{u}$ and p denote velocity and pressure, respectively, the boundary conditions for the velocity field $\mathbf{g}_0(t)$ is $\mathbf{0}$ at the bottom of Ω and $(c,0)^t$ at the top of Ω with a fixed speed c for shear flow, $\boldsymbol{\lambda}$ is a Lagrange multiplier, $\mathbf{D}(\mathbf{v}) = [\boldsymbol{\nabla}\mathbf{v} + (\boldsymbol{\nabla}\mathbf{v})^t]/2$, $\mathbf{g}$ is gravity, $\mathbf{F}$ is the pressure gradient pointing in the x_1-direction, $\mathbf{V_G}$ is the *translation velocity* of the particle B, and ω is the *angular velocity* of B. We suppose that the *no-slip* condition holds on ∂B. We also use, if necessary, the notation $\phi(t)$ for the function $\mathbf{x} \to \phi(\mathbf{x},t)$.

Remark 1. The hydrodynamical forces and torque imposed on the rigid body by the fluid are built in (1)–(6) implicitly (see [GPHJ99, GPH+01] for details), thus we do not need to compute them explicitly in the simulation. Since in (1)–(6) the flow field is defined on the entire domain Ω, it can be computed with a simple structured grid.

   The forces obtained from those Hookean springs in the model for cell adhesion has been splitted from the above equations and will be used when predicting and correcting the motion and positions of cells with the short repulsion force as discussed in the next section. □

Remark 2. In (3), the rigid body motion in the region occupied by the particle is enforced via Lagrange multipliers $\boldsymbol{\lambda}$. To recover the translation velocity $\mathbf{V_G}(t)$ and the angular velocity $\omega(t)$, we solve the following equations:

$$\begin{cases} \langle \mathbf{e}_i, \mathbf{u}(t) - \mathbf{V_G}(t) - \omega(t)\overrightarrow{\mathbf{Gx}}^{\perp}\rangle_{B(t)} = 0, & \text{for } i = 1,2, \\ \langle \overrightarrow{\mathbf{Gx}}^{\perp}, \mathbf{u}(t) - \mathbf{V_G}(t) - \omega(t)\overrightarrow{\mathbf{Gx}}^{\perp}\rangle_{B(t)} = 0. \end{cases} \tag{7}$$

□

Remark 3. In (1), $2\int_\Omega \mathbf{D}(\mathbf{u}) : \mathbf{D}(\mathbf{v})\,d\mathbf{x}$ can be replaced by $\int_\Omega \boldsymbol{\nabla}\mathbf{u} : \boldsymbol{\nabla}\mathbf{v}\,d\mathbf{x}$ since $\mathbf{u}$ is divergence free and in $W_{0,p}$. Also the gravity $\mathbf{g}$ in (1) can be absorbed into the pressure term. □

3.2 Space Approximation and Time Discretization

Concerning the *space approximation* of the problem (1)–(6) by a finite element method, we have chosen P_1-*iso*-P_2 and P_1 finite elements for the velocity field and pressure, respectively (like in [BGP87]). More precisely, with h, a *space discretization step*, we introduce a finite element triangulation $\mathcal{T}_h$ of $\overline{\Omega}$ and then $\mathcal{T}_{2h}$ a triangulation twice coarser. (In practice, we should construct $\mathcal{T}_{2h}$ first and then $\mathcal{T}_h$ by joining the midpoints of the edges of $\mathcal{T}_{2h}$, dividing thus each triangle of $\mathcal{T}_{2h}$ into four similar subtriangles as shown in Figure 3.)

   We approximate then $W_{\mathbf{g}_0,p}$, $W_{0,p}$, L^2 and L_0^2 by the following finite dimensional spaces, respectively:
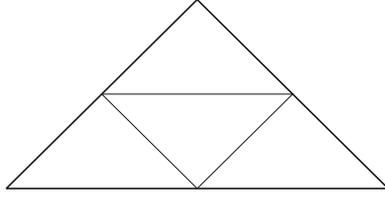
Fig. 3. Subdivision of a triangle of $\mathcal{T}_{2h}$.

$$W_{\mathbf{g}_0,h}(t) = \{\mathbf{v}_h \mid \mathbf{v}_h \in (C^0(\overline{\Omega}))^2, \ \mathbf{v}_h|_T \in P_1 \times P_1, \ \forall T \in \mathcal{T}_h, \ \mathbf{v}_h = \mathbf{g}_0(t)$$
$$\text{on the top and bottom of } \Omega \text{ and } \mathbf{v} \text{ is periodic at } \Gamma$$
$$\text{in the } x_1\text{-direction}\}, \tag{8}$$

$$W_{0,h} = \{\mathbf{v}_h \mid \mathbf{v}_h \in (C^0(\overline{\Omega}))^2, \ \mathbf{v}_h|_T \in P_1 \times P_1, \ \forall T \in \mathcal{T}_h, \ \mathbf{v}_h = \mathbf{0}$$
$$\text{on the top and bottom of } \Omega \text{ and } \mathbf{v} \text{ is periodic at } \Gamma$$
$$\text{in the } x_1\text{-direction}\}, \tag{9}$$

$$L_h^2 = \{q_h \mid q_h \in C^0(\overline{\Omega}), \ q_h|_T \in P_1, \ \forall T \in \mathcal{T}_{2h}, q_h \ \textit{is periodic at } \Gamma$$
$$\text{in the } x_1\text{-direction}\}, \tag{10}$$

$$L_{0,h}^2 = \{q_h \mid q_h \in L_h^2, \ \int_\Omega q_h \, d\mathbf{x} = 0\}. \tag{11}$$

In (8)–(11), P_1 is the space of polynomials in two variables of degree ≤ 1.

Remark 4. A different choice of finite element, the Taylor–Hood finite element, for the velocity field has been considered in [JGP02] for simulating the fluid/particle interaction via distributed Lagrange multiplier based fictitious domain method for non-neutrally buoyant particles. □

A finite dimensional space approximating $\Lambda_0(t)$ is defined as follows: let $\{\mathbf{x}_i\}_{i=1}^N$ be a set of points covering $\overline{B(t)}$ (see Figure 4, for example); we define then

$$\Lambda_h(t) = \left\{ \boldsymbol{\mu}_h \mid \boldsymbol{\mu}_h = \sum_{i=1}^N \boldsymbol{\mu}_i \delta(\mathbf{x} - \mathbf{x}_i), \ \boldsymbol{\mu}_i \in \mathbb{R}^2, \ \forall i = 1, ..., N \right\}, \tag{12}$$

where $\delta(\cdot)$ is the Dirac measure at $\mathbf{x} = \mathbf{0}$. Then, instead of the scalar product of $(H^1(B(t)))^2$ we shall use $\langle \cdot, \cdot \rangle_{B_h(t)}$ defined by

$$\langle \boldsymbol{\mu}_h, \mathbf{v}_h \rangle_{B_h(t)} = \sum_{i=1}^N \boldsymbol{\mu}_i \cdot \mathbf{v}_h(\mathbf{x}_i), \quad \forall \boldsymbol{\mu}_h \in \Lambda_h(t), \ \mathbf{v}_h \in W_{0,h}. \tag{13}$$

Then we approximate $\Lambda_0(t)$ by

$$\Lambda_{0,h}(t) = \left\{ \mu_h \mid \mu_h \in \Lambda_h(t), \ \langle \mu_h, \mathbf{e}_i \rangle_{B_h(t)} = 0, \ i = 1, 2, \ \langle \mu_h, \overrightarrow{\mathbf{G}\mathbf{x}}^{\perp} \rangle_{B_h(t)} = 0 \right\}. \tag{14}$$
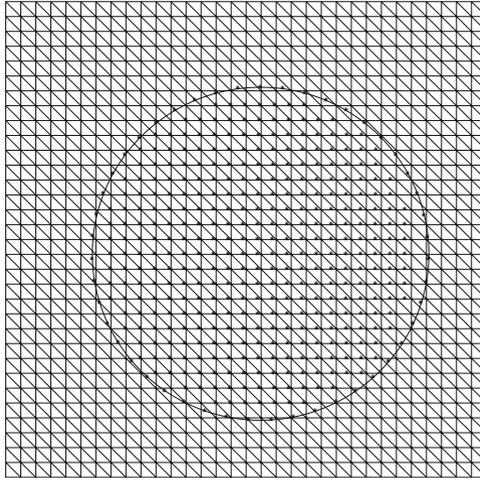
Fig. 4. An example of set of collocation points chosen for enforcing the rigid body motion inside the disk and at its boundary.

Using the above finite dimensional spaces leads to the following approximation of the problem (1)–(6):

For a.e. $t > 0$, find $\mathbf{u}(t) \in W_{\mathbf{g}_0,h}(t)$, $p(t) \in L^2_{0,h}$, $\mathbf{V_G}(t) \in \mathbb{R}^2$, $\mathbf{G}(t) \in \mathbb{R}^2$, $\omega(t) \in \mathbb{R}$, $\boldsymbol{\lambda}_h(t) \in \Lambda_{0,h}(t)$ such that

$$\rho_f \int_\Omega \left[\frac{\partial \mathbf{u}_h}{\partial t} + (\mathbf{u}_h \cdot \boldsymbol{\nabla}) \mathbf{u}_h \right] \cdot \mathbf{v} \, d\mathbf{x} + \mu_f \int_\Omega \boldsymbol{\nabla} \mathbf{u}_h : \boldsymbol{\nabla} \mathbf{v} \, d\mathbf{x}$$

$$- \int_\Omega p_h \boldsymbol{\nabla} \cdot \mathbf{v} \, d\mathbf{x} - \langle \boldsymbol{\lambda}_h, \mathbf{v} \rangle_{B_h(t)} = \int_\Omega \mathbf{F} \cdot \mathbf{v} \, d\mathbf{x}, \quad \forall \mathbf{v} \in W_{0,h}, \qquad (15)$$

$$\int_\Omega q \boldsymbol{\nabla} \cdot \mathbf{u}_h(t) d\mathbf{x} = 0, \quad \forall q \in L^2_h, \qquad (16)$$

$$\langle \boldsymbol{\mu}, \mathbf{u}_h(t) \rangle_{B_h(t)} = 0, \quad \forall \boldsymbol{\mu} \in \Lambda_{0,h}(t), \qquad (17)$$

$$\frac{d\mathbf{G}}{dt} = \mathbf{V_G}, \qquad (18)$$

$$\mathbf{V_G}(0) = \mathbf{V}^0_\mathbf{G}, \quad \omega(0) = \omega^0, \quad \mathbf{G}(0) = \mathbf{G}^0 = \{G^0_1, G^0_2\}^t, \qquad (19)$$

$$\mathbf{u}_h(\mathbf{x}, 0) = \bar{\mathbf{u}}_{0,h}(\mathbf{x}) \quad (\text{with } \boldsymbol{\nabla} \cdot \bar{\mathbf{u}}_{0,h} = 0). \qquad (20)$$

Applying a first order operator splitting scheme, Lie's scheme [CHMM78] and backward Euler scheme at some fractional steps, to discretize the equations (15)–(20) in time, we obtain (after dropping some of the subscripts h):

Algorithm 1

Step 1. $\mathbf{u}^0 = \bar{\mathbf{u}}_{0,h}$, $\mathbf{V}^0_\mathbf{G}$, ω^0, and $\mathbf{G}^0$ are given;

Step 2. For $n \geq 0$, knowing $\mathbf{u}^n$, $\mathbf{V}^n_\mathbf{G}$, ω^n and $\mathbf{G}^n$, compute $\mathbf{u}^{n+1/6}$ and $p^{n+1/6}$ via the solution of

$$\begin{cases} \rho_f \int_\Omega \dfrac{\mathbf{u}^{n+1/6} - \mathbf{u}^n}{\triangle t} \cdot \mathbf{v}\, d\mathbf{x} - \int_\Omega p^{n+1/6} \boldsymbol{\nabla} \cdot \mathbf{v}\, d\mathbf{x} = 0, \quad \forall \mathbf{v} \in W_{0,h}, \\ \int_\Omega q \boldsymbol{\nabla} \cdot \mathbf{u}^{n+1/6}\, d\mathbf{x} = 0, \quad \forall q \in L_h^2; \ \mathbf{u}^{n+1/6} \in W_{\mathbf{g}_0,h}^{n+1}, \ p^{n+1/6} \in L_{0,h}^2. \end{cases}$$

(21)

Step 3. Compute $\mathbf{u}^{n+2/6}$ via the solution of

$$\begin{cases} \int_\Omega \dfrac{\partial \mathbf{u}}{\partial t} \cdot \mathbf{v}\, d\mathbf{x} + \int_\Omega (\mathbf{u}^{n+1/6} \cdot \boldsymbol{\nabla}) \mathbf{u} \cdot \mathbf{v}\, d\mathbf{x} = 0, \\ \qquad \forall \mathbf{v} \in W_{0,h}, \ \text{on } (t^n, t^{n+1}), \\ \mathbf{u}(t^n) = \mathbf{u}^{n+1/6}; \quad \mathbf{u}(t) \in W_{\mathbf{g}_0,h}^{n+1}, \end{cases}$$

(22)

$$\mathbf{u}^{n+2/6} = \mathbf{u}(t^{n+1}).$$

(23)

Step 4. Compute $\mathbf{u}^{n+3/6}$ via the solution of

$$\begin{cases} \rho_f \int_\Omega \dfrac{\mathbf{u}^{n+3/6} - \mathbf{u}^{n+2/6}}{\triangle t} \cdot \mathbf{v}\, d\mathbf{x} + \alpha \mu_f \int_\Omega \boldsymbol{\nabla} \mathbf{u}^{n+3/6} \cdot \boldsymbol{\nabla} \mathbf{v}\, d\mathbf{x} = 0, \\ \forall \mathbf{v} \in W_{0,h}; \quad \mathbf{u}^{n+3/6} \in W_{\mathbf{g}_0,h}^{n+1}. \end{cases}$$

(24)

Step 5. Predict the position and the translation velocity of the center of mass of the particles as follows: Take $\mathbf{V_G}^{n+\frac{4}{6},0} = \mathbf{V_G}^n$ and $\mathbf{G}^{n+\frac{4}{6},0} = \mathbf{G}^n$. Then predict the new position of the particle via the following sub-cycling and predicting-correcting technique:

For $k = 1, \ldots, N$,

Call Adhesive Dynamics Algorithm,

$$\widehat{\mathbf{V}}_{\mathbf{G}}^{n+\frac{4}{6},k} = \mathbf{V}_{\mathbf{G}}^{n+\frac{4}{6},k-1} + \mathbf{F}^r(\mathbf{G}^{n+\frac{4}{6},k-1}) \triangle t/2N,$$

(25)

$$\widehat{\mathbf{G}}^{n+\frac{4}{6},k} = \mathbf{G}^{n+\frac{4}{6},k-1} + (\widehat{\mathbf{V}}_{\mathbf{G}}^{n+\frac{4}{6},k} + \mathbf{V}_{\mathbf{G}}^{n+\frac{4}{6},k-1}) \triangle t/4N,$$

(26)

$$\mathbf{V}_{\mathbf{G}}^{n+\frac{4}{6},k} = \mathbf{V}_{\mathbf{G}}^{n+\frac{4}{6},k-1} + (\mathbf{F}^r(\widehat{\mathbf{G}}^{n+\frac{4}{6},k})$$
$$+ \mathbf{F}^r(\mathbf{G}^{n+\frac{4}{6},k-1})) \triangle t/4N,$$

(27)

$$\mathbf{G}^{n+\frac{4}{6},k} = \mathbf{G}^{n+\frac{4}{6},k-1} + (\mathbf{V}_{\mathbf{G}}^{n+\frac{4}{6},k} + \mathbf{V}_{\mathbf{G}}^{n+\frac{4}{6},k-1}) \triangle t/4N,$$

(28)

enddo;

and let $\mathbf{V}_{\mathbf{G}}^{n+\frac{4}{6}} = \mathbf{V}_{\mathbf{G}}^{n+\frac{4}{6},N}$, $\mathbf{G}^{n+\frac{4}{6}} = \mathbf{G}^{n+\frac{4}{6},N}$.

Step 6. Now, compute $\mathbf{u}^{n+5/6}$, $\boldsymbol{\lambda}^{n+5/6}$, $\mathbf{V}_{\mathbf{G}}^{n+5/6}$, and $\omega^{n+5/6}$ via the solution of

$$\begin{cases} \rho_f \int_\Omega \dfrac{\mathbf{u}^{n+5/6} - \mathbf{u}^{n+3/6}}{\triangle t} \cdot \mathbf{v}\, d\mathbf{x} + \beta \mu_f \int_\Omega \boldsymbol{\nabla} \mathbf{u}^{n+5/6} \cdot \boldsymbol{\nabla} \mathbf{v}\, d\mathbf{x} \\ = \langle \boldsymbol{\lambda}, \mathbf{v} \rangle_{B_h^{n+4/6}}, \ \forall \mathbf{v} \in W_{0,h}, \\ \langle \boldsymbol{\mu}, \mathbf{u}^{n+5/6} \rangle_{B_h^{n+4/6}} = 0, \\ \forall \boldsymbol{\mu} \in \Lambda_{0,h}^{n+4/6}; \ \mathbf{u}^{n+5/6} \in W_{\mathbf{g}_0,h}^{n+1}, \boldsymbol{\lambda}^{n+5/6} \in \Lambda_{0,h}^{n+4/6}, \end{cases}$$

(29)

and solve for $\mathbf{V}_{\mathbf{G}}^{n+5/6}$ and $\omega^{n+5/6}$ from

$$
\begin{cases}
\langle \mathbf{e}_i, \mathbf{u}^{n+5/6} - \mathbf{V}_{\mathbf{G}}^{n+5/6} - \omega^{n+5/6}\, \overrightarrow{G^{n+4/6}x}^{\perp}\rangle_{B_h^{n+4/6}} = 0, \quad \text{for } i = 1, 2,\\
\langle \overrightarrow{G^{n+4/6}x}^{\perp}, \mathbf{u}^{n+5/6} - \mathbf{V}_{\mathbf{G}}^{n+5/6} - \omega^{n+5/6}\, \overrightarrow{G^{n+4/6}x}^{\perp}\rangle_{B_h^{n+4/6}} = 0.
\end{cases}
\tag{30}
$$

Step 7. Finally, take $\mathbf{V}_{\mathbf{G}}^{n+1,0} = \mathbf{V}_{\mathbf{G}}^{n+5/6}$ and $\mathbf{G}^{n+1,0} = \mathbf{G}^{n+4/6}$. Then predict the final position and translation velocity as follows:

For $k = 1, \ldots, N$,

Call Adhesive Dynamics Algorithm,

$$
\widehat{\mathbf{V}}_{\mathbf{G}}^{n+1,k} = \mathbf{V}_{\mathbf{G}}^{n+1,k-1} + \mathbf{F}^r(\mathbf{G}^{n+1,k-1})\triangle t/2N,
\tag{31}
$$

$$
\widehat{\mathbf{G}}^{n+1,k} = \mathbf{G}^{n+1,k-1} + (\widehat{\mathbf{V}}_{\mathbf{G}}^{n+1,k} + \mathbf{V}_{\mathbf{G}}^{n+1,k-1})\triangle t/4N,
\tag{32}
$$

$$
\mathbf{V}_{\mathbf{G}}^{n+1,k} = \mathbf{V}_{\mathbf{G}}^{n+1,k-1} + (\mathbf{F}^r(\widehat{\mathbf{G}}^{n+1,k}) + \mathbf{F}^r(\mathbf{G}^{n+1,k-1}))\triangle t/4N,
\tag{33}
$$

$$
\mathbf{G}^{n+1,k} = \mathbf{G}^{n+1,k-1} + (\mathbf{V}_{\mathbf{G}}^{n+1,k} + \mathbf{V}_{\mathbf{G}}^{n+1,k-1})\triangle t/4N,
\tag{34}
$$

enddo;
and let $\mathbf{V}_{\mathbf{G}}^{n+1} = \mathbf{V}_{\mathbf{G}}^{n+1,N}$, $\mathbf{G}^{n+1} = \mathbf{G}^{n+1,N}$; and set $\mathbf{u}^{n+1} = \mathbf{u}^{n+5/6}$, $\omega^{n+1} = \omega^{n+5/6}$.

In Algorithm 1, we have $t^{n+s} = (n+s)\triangle t$, $W_{\mathbf{g}0,h}^{n+1} = W_{\mathbf{g}0,h}(t^{n+1})$, $\Lambda_{0,h}^{n+s} = \Lambda_{0,h}(t^{n+s})$, B_h^{n+s} is the region occupied by the particle centered at $\mathbf{G}^{n+s}$, and $\mathbf{F}^r$ is the combination of a short range repulsion force which prevents the particle/particle and particle/wall penetration (see, e.g., [GPHJ99, GPH+01]) and the force obtained from the adhesive dynamics algorithm for the cell adhesion. Finally, α and β verify $\alpha + \beta = 1$; we have chosen $\alpha = 1$ and $\beta = 0$ in the numerical simulations discussed later.

The degenerated quasi-Stokes problem (21) is solved by a preconditioned conjugate gradient method introduced in [GPP98], in which discrete elliptic problems from the preconditioning are solved by a matrix-free fast solver from FISHPAK by Adams et al. in [ASS80]. The advection problem (22) for the velocity field is solved by a wave-like equation method as in [DG97]. The problem (24) is a classical discrete elliptic problem which can be solved by the same matrix-free fast solver. To enforce the rigid body motion inside the region occupied by the particles, we have applied the conjugate gradient method discussed in [PG02, PG05].

4 Numerical Results and Discussion

We consider the detachment of 20 cells in shear flow as the test problem for cell adhesion model at the initial stage of the adhesion. The computational domain is $\Omega = (0, 23) \times (0, 10)$ (unit: $10\,\mu m$). Cells have the shape of an ellipse,

Table 1. Simulation parameters.

| Parameters	Definition	simulation value
R	cell radius	4.0–5.0 μm
N_r	receptor number	780
N_L	ligand density	10^6–10^8/cm
λ	equilibrium bond length	0.2 μm
σ	spring constant	0.016 dyne/cm
μ	viscosity	0.01–0.014 g/cm·s
ρ	fluid density	1.0 g/cm^2
$U_{\max}$	shear rate	20–80/s
H_c	cut-off length	0.4 μm
T	temperature	310 K
k_f^0	forward reaction rate	100.0/s
k_r^0	reverse reaction rate	10.0/s
r_0	reactive compliance	0.02 μm

with the long semi-axis r_a equal to 0.5 and the short semi-axis r_b equal to 0.4. The velocity boundary conditions are as follows: a given constant on the top boundary, zero on the bottom boundary, and periodicity in the horizontal direction. The fluid and cells are at rest and the cells are in the contact region initially (see Fig. 6(a)). We assume that the densities of fluid and cells are $1\,\text{g/cm}^3$. The mesh size h for the flow field is 1/48 and the time step $\triangle t$ is 0.001 (unit: 0.1 second). The parameters used in the simulations are given in Table 1.

We observed the simulations up to $t = 100$ (10 s), long enough for the flow to be fully developed. The simulations were conducted at different shear rates and dynamical viscosities, and the results are summarized in Table 2. From the table, we can see, no cells were detached from the wall by the observed time when the shear rate is 20/s for the dynamical viscosity of 0.01 g/cm-s; while the detachment percentage increases from 10% to 40% when the shear rate increases from 30/s to 40/s. All the 20 cells were detached from the wall when the shear rate is greater than 80/s. Figure 5 shows the effect of shear rate on cell detachment. This observation qualitatively agrees with the *in vitro* experiment [SKE$^+$99]. We also observed that the detachment percentage increases from 10% to 35% when the dynamical viscosity is increased from 0.01 to 0.014 (g/cm-s).

Figure 6 shows the snapshots of positions of 20 cells at $t = 0$, 5, 5.35, 6.06, 9.49, and 10 (s), for the simulation with the dynamical viscosity equal to 0.01 (g/cm-s) and the shear rate of 30 (/s). The snapshots quite clearly depict the process of cell detachment from the wall. All the cells adhered to the wall at $t = 5$ s; one cell was about to be detached at $t = 5.35$ s; one cell was completely detached from the layer at $t = 6.06$ s. We found that during the early stage of detachment the percentage of the detached cells is highly

Table 2. The calculated detachment percentages at $t = 10$ s.

Dynamical viscosity (g/cm-s)	Shear rate (/s)	Detachment (%)
0.01	20	0
0.01	30	10
0.014	30	35
0.01	40	40
0.01	80	100
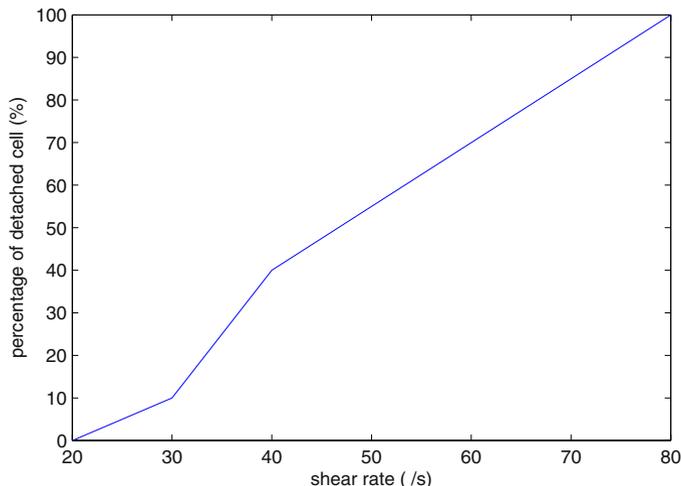
Fig. 5. The effect of shear rate on cell detachment (viscosity$= 0.01$ g/cm-s).

linearly correlated with the observed time. This observation was also found in *in vitro* experiments [SBBR$^+$02].

We have used our models and algorithms to simulate adhesion and detachment of chondrocytes. The simulations successfully depicted the process of cell detachment from the wall. The numerical results qualitatively agree with the experiments in the literature. Since there are few publications on modeling chondrocytes for this problem, our modeling and simulation are quite preliminary. More work is needed in modeling and in investigating parameters for the cell adhesion at different stages as discussed in [ZBCAG04].
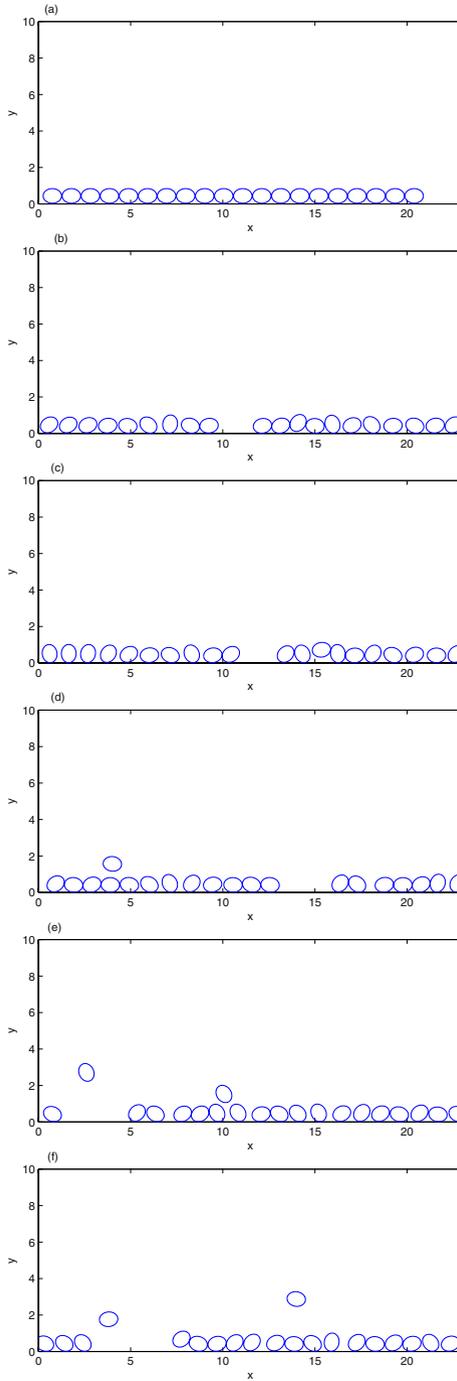
Fig. 6. Snapshots of 20 cells at $t = 0.0\,\mathrm{s}$ (a), $5.0\,\mathrm{s}$ (b), $5.35\,\mathrm{s}$ (c), $6.06\,\mathrm{s}$ (d), $9.49\,\mathrm{s}$ (e), and $10.0\,\mathrm{s}$ (f) (viscosity $= 0.01\,\mathrm{g/cm\text{-}s}$, shear rate $= 30/\mathrm{s}$). The percentage of detached cells is 10% at $t = 10.0\,\mathrm{s}$.

References

[ASS80]   J. Adams, P. Swarztrauber, and R. Sweet. *FISHPAK: A package of Fortran subprograms for the solution of separable elliptic partial differential equations.* The National Center for Atmospheric Research, Boulder, CO, 1980.

[BGP87]   M. O. Bristeau, R. Glowinski, and J. Periaux. Numerical methods for the Navier–Stokes equations. Applications to the simulation of compressible and incompressible viscous flow. *Comput. Phys. Reports*, 6:73–187, 1987.

[CH96]   K. Chang and D. Hammer. Influence of direction and type of applied force on the detachment of macromolecularly-bound particles from surfaces. *Langmuir*, 12:2271–2282, 1996.

[CHMM78]   A. J. Chorin, T. J. R. Hughes, J. E. Marsden, and M. McCracken. Product formulas and numerical algorithms. *Comm. Pure Appl. Math.*, 31:205–256, 1978.

[CKGA03]   M. Cohen, E. Klein, B. Geiger, and L. Addadi. Organization and adhesive properties of the hyaluronan pericellular coat of chondrocytes and epithelial cells. *Biophys. J.*, 85:1996–2005, 2003.

[DG97]   E. J. Dean and R. Glowinski. A wave equation approach to the numerical solution of the Navier–Stokes equations for incompressible viscous flow. *C. R. Acad. Sci. Paris Sér. I Math.*, 325(7):783–791, 1997.

[GHR04]   U. R. Goessler, K. Hörmann, and F. Riedel. Tissue engineering with chondrocytes and function of the extracellular matrix (review). *Int. J. Mol. Med.*, 13:505–513, 2004.

[GPH⁺01]   R. Glowinski, T.-W. Pan, T. I. Hesla, D. D. Joseph, and J. Périaux. A fictitious domain approach to the direct numerical simulation of incompressible viscous flow past moving rigid bodies: Application to particulate flow. *J. Comput. Phys.*, 169(2):363–426, 2001.

[GPHJ99]   R. Glowinski, T.-W. Pan, T. Hesla, and D. D. Joseph. A distributed Lagrange multiplier/fictitious domain method for particulate flows. *Int. J. Multiph. Flow*, 25(5):755–794, 1999.

[GPP98]   R. Glowinski, T.-W. Pan, and J. Périaux. Distributed Lagrange multiplier methods for incompressible flow around moving rigid bodies. *Comput. Methods Appl. Mech. Engrg.*, 151(1–2):181–194, 1998.

[JGP02]   L. H. Juarez, R. Glowinski, and T.-W. Pan. Numerical simulation of the sedimentation of rigid bodies in an incompressible viscous fluid by Lagrange multiplier/fictitious domain methods combined with the Taylor–Hood finite element approximation. *J. Sci. Comput.*, 17:683–694, 2002.

[KH01]   M. R. King and D. A. Hammer. Multiparticle adhesive dynamics. interactions between stably rolling cells. *Biophys. J.*, 81:799–813, 2001.

[KS06]   C. Korn and U. S. Schwarz. Efficiency of initiating cell adhesion in hydrodynamic flow. *Phys. Rev. Lett.*, 97, 2006. 138103.

[Loe93]   R. F. Loeser. Integrin-mediated attachment of articular chondrocytes to extracellular matrix proteins. *Arthritis Rheum.*, 36:1103–1110, 1993.

[PG02]   T.-W. Pan and R. Glowinski. Direct simulation of the motion of neutrally buoyant circular cylinders in plane Poiseuille flow. *J. Comput. Phys.*, 181:260–279, 2002.

[PG05]   T.-W. Pan and R. Glowinski. Direct simulation of the motion of neu-
            trally buoyant balls in a three-dimensional Poiseuille flow. *C. R.
            Mécanique*, 333:884–895, 2005.
[SBBR+02]   T. Scott-Burden, J. P. Bosley, D. Rosenstrauch, K. D. Henderson,
            F. J. Clubb, H. C. Eichstaedt, K. Eya, I. Gregoric, T. J. Myers,
            B. Radovancevic, and O. H. Frazier. Use of autologous auricular chon-
            drocytes for lining artificial surfaces: a feasibility study. *Ann. Thorac.
            Surg.*, 73:1528–1533, 2002.
[SKE+99]   R. M. Schinagl, M. S. Kurtis, K. D. Ellis, S. Chien, and R. L. Sah.
            Effect of seeding duration on the strength of chondrocyte adhesion to
            articular cartilage. *J. Orthopaedic Research*, 17:121–129, 1999.
[SZD03]   M. E. Staben, A. Z. Zinchenko, and R. H. Davis. Motion of a particle
            between two parallel plane walls in low-Reynolds-number Poiseuille
            flow. *Phys. Fluid*, 15:1711–1733, 2003.
[ZBCAG04]   R. Zaidel-Bar, M. Cohen, L. Addadi, and B. Geiger. Hierarchical
            assembly of cell-matrix adhesion complexes. *Biochem. Soc. Trans.*,
            32(3):416–420, 2004.

Computing the Eigenvalues
of the Laplace–Beltrami Operator
on the Surface of a Torus: A Numerical
Approach

Roland Glowinski[1] and Danny C. Sorensen[2]

[1] University of Houston, Department of Mathematics, Houston, TX, 77004, USA
`roland@math.uh.edu`
[2] Rice University, Department of Computational and Applied Mathematics,
Houston, TX, 77251-1892, USA `sorensen@rice.edu`

Summary. In this chapter, we present a methodology for numerically computing
the eigenvalues and eigenfunctions of the Laplace–Beltrami operator on the surface
of a torus. Beginning with a variational formulation, we derive an equivalent PDE
formulation and then discretize the PDE using finite differences to obtain an alge-
braic generalized eigenvalue problem. This finite dimensional eigenvalue problem is
solved numerically using the `eigs` function in Matlab which is based upon ARPACK.
We show results for problems of order 16K variables where we computed lowest 15
modes. We also show a bifurcation study of eigenvalue trajectories as functions of
aspect ration of the major to minor axis of the torus.

1 Introduction

A large number of physical phenomena take place on surfaces. Many of these
are modeled by partial differential equations, a typical example being pro-
vided by elastic shells. It is not surprising, therefore, that many questions
have arisen concerning the spectrum of some partial differential operators de-
fined on surfaces. This area of investigation is known as spectral geometry.
Among these operators defined on surfaces, a most important one is the so
called Beltrami Laplacian, also known as the Laplace–Beltrami operator. The
main goal of this chapter is to discuss the computation of the lowest eigenval-
ues of the Laplace–Beltrami operator associated with the boundary of a torus
of $\mathbb{R}^3$. After a description of our methodology for the computation of these
eigenvalues and their corresponding eigenfunctions, we present selected re-
sults from our numerical experiments. The methodology consists of obtaining
a finite difference discretization of a PDE that is equivalent to a more stan-
dard variational formulation; then the resulting finite dimensional generalized

eigenvalue problem is solved to obtain the approximations. A visualization of
our results show the expected Sturm–Liousville behavior of the eigenfunctions
according to wave number. Eigenvalues are typically multiplicity one or two.
However, we show that for certain ratios of the minor to major radii, it is
possible to create eigenvalues of multiplicity three or four. This indicates an
interesting bifurcation structure is associated with this ratio.

A thorough discussion of the approximate solution of eigenvalue problems
for elliptic operators is given by Babushka and Osborn [BO91].

2 Variational Formulation of the Eigenvalue Problem

Let Σ be the boundary of a three-dimensional torus defined by a great circle
of radius R and a small circle of radius ρ (see Figure 1).

Our goal here is to numerically approximate the eigenvalues and corre-
sponding eigenfunctions of the Laplace–Beltrami operator associated with Σ.
A variational formulation of this problem reads as follows:

Find $\lambda \in \mathbb{R}$, $u \in \mathcal{H}^1(\Sigma)$ such that

$$\int_\Sigma \nabla_\Sigma u \cdot \nabla_\Sigma v \, \mathrm{d}\Sigma = \lambda \int_\Sigma u v \, \mathrm{d}\Sigma, \quad \forall v \in \mathcal{H}^1(\Sigma). \tag{1}$$

In the equation (1):

(i) ∇_Σ is the tangential gradient on Σ,
(ii) $\mathrm{d}\Sigma$ is the infinitesimal superficial (surfacic) measure,
(iii) $\mathcal{H}^1(\Sigma) = \{v | v \in \mathcal{L}^2(\Sigma), \int_\Sigma |\nabla_\Sigma v|^2 \mathrm{d}\Sigma < +\infty\}$.

Any function constant over Σ is an eigenfunction of the Laplace–Beltrami
operator, the corresponding eigenvalue being 0 (of multiplicity 1). Our interest
is in the non-trivial solutions of (1). To compute them (at least some of the
smallest ones), we shall use the (θ, ϕ) coordinates shown in Figure 1. The
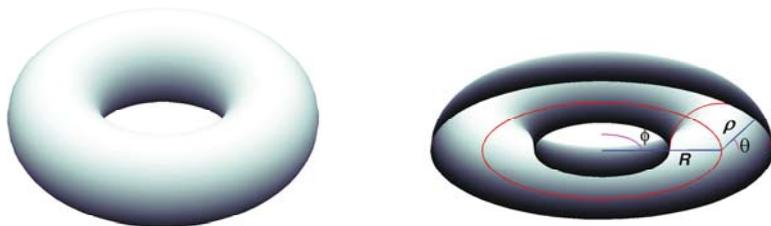problem (1) takes the following form:

Fig. 1. Torus surface Σ (left) and a view from under the top half (right) showing
the major radius R and angle ϕ and the minor radius ρ and angle θ.

Find $u \in \mathcal{H}_p^1(\Omega_0)$ and λ, such that

$$\int_{\Omega_0} \left[\frac{\rho}{R + \rho \cos \theta} \frac{\partial u}{\partial \phi} \frac{\partial v}{\partial \phi} + \frac{R + \rho \cos \theta}{\rho} \frac{\partial u}{\partial \theta} \frac{\partial v}{\partial \theta} \right] \mathrm{d}\phi \mathrm{d}\theta$$

$$= \lambda \int_{\Omega_0} \rho (R + \rho \cos \theta) uv \mathrm{d}\phi \mathrm{d}\theta, \quad (2)$$

for all $v \in \mathcal{H}_p^1(\Omega_0)$, with $\Omega_0 = (0, 2\pi) \times (0, 2\pi)$ and with

$$\mathcal{H}_p^1(\Omega_0) = \{v \mid v \in \mathcal{H}^1(\Omega_0), \; v(0, \theta) = v(2\pi, \theta), \; \text{for a.e. } \theta \in (0, 2\pi),$$
$$v(\phi, 0) = v(\phi, 2\pi), \; \text{for a.e. } \phi \in (0, 2\pi)\},$$

i.e., $\mathcal{H}_p^1(\Omega_0)$ is a space of doubly periodic functions. In the following, keep in mind that $0 < \rho < R$.

3 An Equivalent PDE Formulation

It follows from the theory of uniformly elliptic operators with smooth coefficients that solving (2) is equivalent to finding $u \in \mathcal{C}^\infty(\overline{\Omega}_0)$, such that

$$- (R\rho^{-1} + \cos \theta)^{-1} \frac{\partial^2 u}{\partial \phi^2} - \frac{\partial u}{\partial \theta} \left[(R\rho^{-1} + \cos \theta) \frac{\partial u}{\partial \theta} \right]$$

$$= \lambda \rho^2 (R\rho^{-1} + \cos \theta) u \quad \text{in } \Omega_0, \quad (3)$$

$$u(0, \theta) = u(2\pi, \theta), \quad \forall \theta \in [0, 2\pi], \qquad u(\phi, 0) = u(\phi, 2\pi), \qquad \forall \phi \in [0, 2\pi],$$

$$\frac{\partial u}{\partial \phi}(0, \theta) = \frac{\partial u}{\partial \phi}(2\pi, \theta), \quad \forall \theta \in [0, 2\pi], \qquad \frac{\partial u}{\partial \theta}(\phi, 0) = \frac{\partial u}{\partial \theta}(\phi, 2\pi), \quad \forall \phi \in [0, 2\pi].$$

4 Finite Difference Discretization

Let I be a positive integer ($I \gg 1$ in practice). From I, we define the spatial discretization step h as $h = \frac{2\pi}{I}$ and then $\phi_i = ih$ and $\theta_j = jh$ for $i = 0, 1, \ldots, I$ and $j = 0, 1, \ldots, I$. We denote the point (ϕ_i, θ_j) by M_{ij}. Taking advantage of the periodic boundary conditions, we discretize the elliptic equation in (3) at those points M_{ij} such that $1 \le i \le I$ and $1 \le j \le I$.

With the usual notation $(u_{ij} = u(\phi_i, \theta_j))$ we obtain for all $1 \le i, j \le I$

$$(R\rho^{-1} + \cos \theta_j)^{-1} (2u_{ij} - u_{i+1j} - u_{i-1j}) + (R\rho^{-1} + \cos(\theta_j + h/2))(u_{ij} - u_{ij+1})$$
$$+ (R\rho^{-1} + \cos(\theta_j - h/2))(u_{ij} - u_{ij-1}) = \lambda \rho^2 (R\rho^{-1} + \cos \theta_j) h^2 u_{ij}, \quad (4)$$

with $u_{I+1j} = u_{1j}$ and $u_{0j} = u_{Ij}$, for $j = 1, 2, \ldots, I$, and with $u_{iI+1} = u_{i1}$ and $u_{i0} = u_{iI}$, for $i = 1, 2, \ldots, I$.

If these discrete boundary conditions are used to eliminate the unknowns $u_{I+1j}, u_{0j}, u_{iI+1}$ and u_{i0}, we obtain the following discrete eigenproblem (in $\mathbb{R}^N$, $N = I^2$):

If $2 \le i, j \le I - 1$,

$$
2[(R\rho^{-1} + \cos\theta_j)^{-1} + R\rho^{-1} + \cos\theta_j \cos(h/2)]u_{ij}
$$
$$
- (R\rho^{-1} + \cos\theta_j)^{-1}(u_{i+1j} + u_{i-1j}) - (R\rho^{-1} + \cos(\theta_j + h/2))u_{ij+1}
$$
$$
- (R\rho^{-1} + \cos(\theta_j - h/2))u_{ij-1} = \lambda\rho^2(R\rho^{-1} + \cos\theta_j)h^2 u_{ij}. \quad (5)
$$

If $i = 1$ and $2 \le j \le I - 1$,

$$
2[(R\rho^{-1} + \cos\theta_j)^{-1} + R\rho^{-1} + \cos\theta_j \cos(h/2)]u_{1j}
$$
$$
- (R\rho^{-1} + \cos\theta_j)^{-1}(u_{2j} + u_{Ij}) - (R\rho^{-1} + \cos(\theta_j + h/2))u_{1j+1}
$$
$$
- (R\rho^{-1} + \cos(\theta_j - h/2))u_{1j-1} = \lambda\rho^2(R\rho^{-1} + \cos\theta_j)h^2 u_{1j}. \quad (6)
$$

If $i = j = 1$,

$$
2[(R\rho^{-1} + \cos h)^{-1} + R\rho^{-1} + \cos h \cos(h/2)]u_{11}
$$
$$
- (R\rho^{-1} + \cos h)^{-1}(u_{21} + u_{I1}) - (R\rho^{-1} + \cos(3h/2))u_{12}
$$
$$
- (R\rho^{-1} + \cos(h/2))u_{1I} = \lambda\rho^2(R\rho^{-1} + \cos h)h^2 u_{11}. \quad (7)
$$

If $i = 1$ and $j = I$,

$$
2[(R\rho^{-1} + 1)^{-1} + R\rho^{-1} + \cos(h/2)]u_{1I}
$$
$$
- (R\rho^{-1} + 1)^{-1}(u_{2I} + u_{II}) - (R\rho^{-1} + \cos(h/2))u_{11}
$$
$$
- (R\rho^{-1} + \cos(h/2))u_{1I-1} = \lambda\rho^2(R\rho^{-1} + 1)h^2 u_{1I}. \quad (8)
$$

If $i = I$ and $2 \le j \le I - 1$,

$$
2[(R\rho^{-1} + \cos\theta_j)^{-1} + R\rho^{-1} + \cos\theta_j \cos(h/2)]u_{Ij}
$$
$$
- (R\rho^{-1} + \cos\theta_j)^{-1}(u_{1j} + u_{I-1j}) - (R\rho^{-1} + \cos(\theta_j + h/2))u_{Ij+1}
$$
$$
- (R\rho^{-1} + \cos(\theta_j - h/2))u_{Ij-1} = \lambda\rho^2(R\rho^{-1} + \cos\theta_j)h^2 u_{Ij}. \quad (9)
$$

If $i = I$ and $j = 1$,

$$
2[(R\rho^{-1} + \cos h)^{-1} + R\rho^{-1} + \cos h \cos(h/2)]u_{I1}
$$
$$
- (R\rho^{-1} + \cos h)^{-1}(u_{11} + u_{I-11}) - (R\rho^{-1} + \cos(3h/2))u_{I2}
$$
$$
- (R\rho^{-1} + \cos(h/2))u_{II} = \lambda\rho^2(R\rho^{-1} + \cos h)h^2 u_{I1}. \quad (10)
$$

If $i = I$ and $j = I$,

$$2[(R\rho^{-1} + 1)^{-1} + R\rho^{-1} + \cos(h/2)]u_{II}$$
$$- (R\rho^{-1} + 1)^{-1}(u_{1I} + u_{I-1I}) - (R\rho^{-1} + \cos(h/2))u_{I1}$$
$$- (R\rho^{-1} + \cos(h/2))u_{II-1} = \lambda\rho^2(R\rho^{-1} + 1)h^2 u_{II}. \quad (11)$$

If $2 \leq i \leq I - 1$ and $j = 1$,

$$2[(R\rho^{-1} + \cos(h))^{-1} + R\rho^{-1} + \cos(h)\cos(h/2)]u_{i1}$$
$$- (R\rho^{-1} + \cos(h))^{-1}(u_{i+11} + u_{i-11}) - (R\rho^{-1} + \cos(3h/2))u_{i2}$$
$$- (R\rho^{-1} + \cos(h/2))u_{iI} = \lambda\rho^2(R\rho^{-1} + \cos(h))h^2 u_{i1}. \quad (12)$$

If $2 \leq i \leq I - 1$ and $j = I$,

$$2[(R\rho^{-1} + 1)^{-1} + R\rho^{-1} + \cos(h/2)]u_{iI}$$
$$- (R\rho^{-1} + 1)^{-1}(u_{i+1I} + u_{i-1I}) - (R\rho^{-1} + \cos(h/2))u_{i1}$$
$$- (R\rho^{-1} + \cos(h/2))u_{iI-1} = \lambda\rho^2(R\rho^{-1} + 1)h^2 u_{iI}. \quad (13)$$

If $2 \leq i, j \leq I - 1$,

$$2[(R\rho^{-1} + \cos\theta_j)^{-1} + R\rho^{-1} + \cos\theta_j\cos(h/2)]u_{ij}$$
$$- (R\rho^{-1} + \cos\theta_j)^{-1}(u_{i+1j} + u_{i-1j}) - (R\rho^{-1} + \cos(\theta_j + h/2))u_{ij+1}$$
$$- (R\rho^{-1} + \cos(\theta_j - h/2))u_{ij-1} = \lambda\rho^2(R\rho^{-1} + \cos\theta_j)h^2 u_{ij}. \quad (14)$$

These finite difference formulas generate an approximation to the problem (3) in the form of a symmetric generalized eigenvalue problem

$$\mathbf{Ax} = \lambda\mathbf{Dx}, \quad (15)$$

with $\mathbf{A}$ sparse and symmetric positive semi-definite and with $\mathbf{D}$ positive definite and diagonal (independent of the ordering of the variables). We used Matlab to solve the problem (15) to obtain approximations to eigenvalues and corresponding eigenfunctions of (2).

5 Numerical Experiments

The Matlab function `eigs` which is based upon ARPACK [Sor92, LSY98] was used to perform the numerical calculation of eigenvalues and corresponding eigenvectors. In all cases, we computed the 15 lowest (algebraically smallest) eigenvalues of the generalized eigenvalue problem (15) using the shift-invert option with shift $\sigma = -.0001$. Since the eigenvalues are real and non-negative, the eigenvalues closest to the origin are enhanced with this transformation and thus easily computed with a Krylov method.
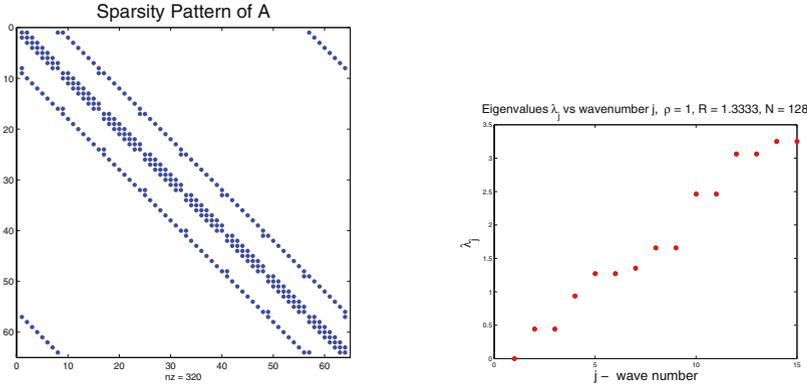
Fig. 2. Sparsity pattern (left) of the matrix **A** and eigenvalue distribution (right) of the lowest 15 modes plotted as a function of index.
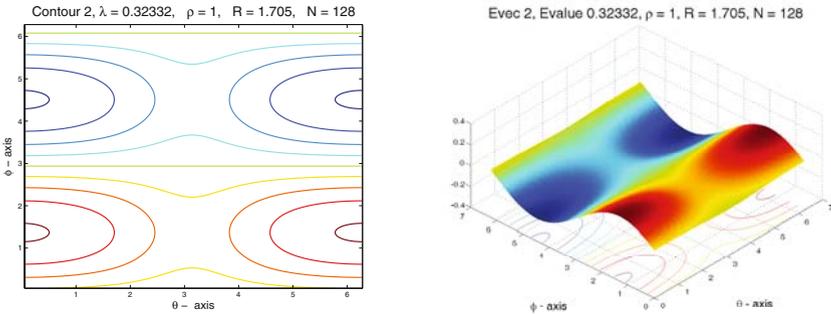

Fig. 3. Contour (left) and surface (right) plots of an eigenfunction corresponding to the lowest nontrivial eigenvalue λ_2 which is a double eigenvalue.

The Matlab command used to accomplish this was

```
[V,Lambda] = eigs(A,D,15,-.0001);
```

which calculates the $k = 15$ eigenvalues closest to the shift $\sigma = -.0001$. The computed eigenvalues are returned as a diagonal matrix `Lambda` and the corresponding eigenvectors are returned as the corresponding columns of the $N \times k$ matrix `V`. Figure 2 shows the sparsity pattern of the matrix **A**.

Figure 3 shows the eigenfunction surface and its contours of the eigenfunction corresponding to the smallest nonzero eigenvalue λ_2. This is a double eigenvalue so $\lambda_3 = \lambda_2$ and the eigenfunction for λ_3 is not shown here. Below this (Fig. 4) are the surface plots of the eigenfunctions of modes 4 to 15. Surfaces 4 and 7 (the simple sheets) correspond to single eigenvalues. The remaining eigenfunction surfaces correspond to double eigenvalues. In all of these plots, $R = 4/3$ and $\rho = 1$. The dimension of the matrix is $N = 16,384$ corresponding to $I = 128$ resulting from a grid stepsize of $h = 2\pi/128$.
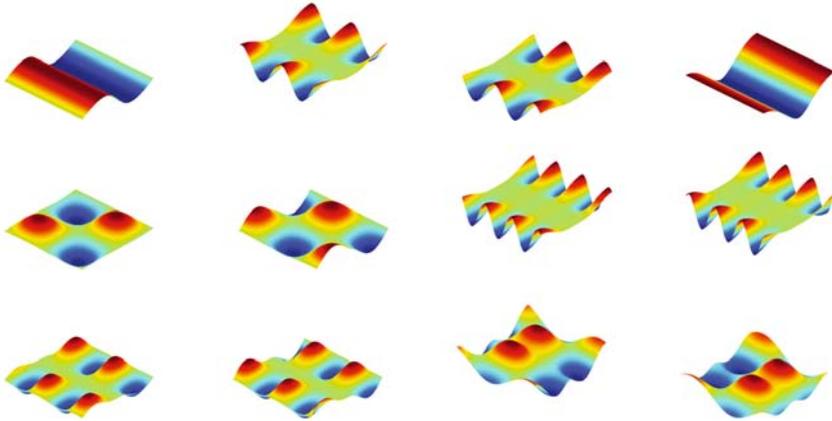
Fig. 4. Eigenfunctions corresponding to eigenvalues λ_4 to λ_{15} (in order left to right, top to bottom).
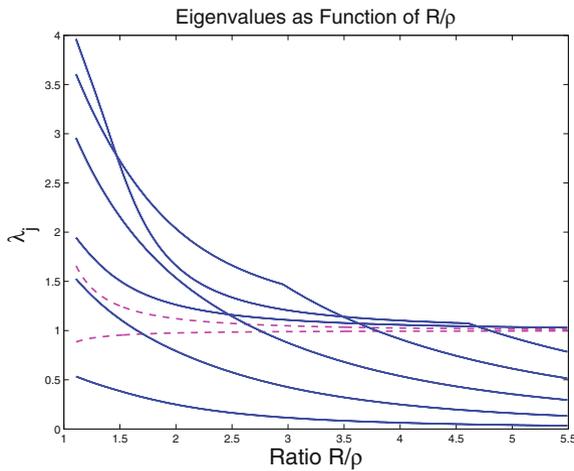
Fig. 5. Bifurcation diagram of 14 leading nontrivial eigenvalues as functions of the ratio R/ρ. Solid curves are double eigenvalues and dashed curves are singletons.

We note that eigenfunctions associated with single eigenvalues are sheets that only change sign in the θ direction. Eigenfunctions corresponding to double eigenvalues change sign in both the θ and ϕ directions. We studied the eigenvalue trajectories plotted as functions of the aspect ratio R/ρ and noted that crossings of these curves provided instances of quadruple eigenvalues and also of triple eigenvalues. Results of this study are shown graphically in Figure 5.

6 Conclusions

We have addressed the numerical solution of a problem from spectral geometry, namely the computation of the lowest eigenvalues of the Laplace–Beltrami operator on the surface of a torus in $\mathbb{R}^3$. The methodology developed here is expected to apply to a number of other surfaces. If combined with appropriate continuation techniques, this approach should enable the numerical solution of certain nonlinear eigenvalue such as those encountered in [FGH07a, FGH07b, ETFS94, SSS]. We also briefly studied the bifurcations of the eigenvalue trajectories as functions of the aspect ration R/ρ. An interesting observation was that trajectories of double eigenvalues could cross other trajectories of double eigenvalues to provide quadruple eigenvalues to appear at certain ratios. The significance of this will be a subject of future study.

References

[BO91]   I. Babuska and J. E. Osborn. Eigenvalue problems. In P. G. Ciarlet and J.-L. Lions, editors, *Handbook of Numerical Analysis. Vol. II, Finite Element Methods (Part 1)*, pages 641–787. North-Holland Publishing Company, Amsterdam, 1991.

[ETFS94]   W. S. Edwards, L. S. Tuckerman, R. A. Friesner, and D. C. Sorensen. Krylov methods for the incompressible Navier–Stokes equations. *Journal of Computational Physics*, 110:82–102, 1994.

[FGH07a]   F. Foss, R. Glowinski, and R. H. W. Hoppe. On the numerical solution of a semilinear elliptic eigenproblem of Lane–Emden type. (I): Problem formulation and description of the algorithms. *Journal of Numerical Mathematics*, 15:181–208, 2007.

[FGH07b]   F. Foss, R. Glowinski, and R. H. W. Hoppe. On the numerical solution of a semilinear elliptic eigenproblem of Lane–Emden type. (II): Numerical experiments. *Journal of Numerical Mathematics*, 15:277–298, 2007.

[LSY98]   R. Lehoucq, D. C. Sorensen, and C. Yang. *ARPACK Users Guide: Solution of Large Scale Eigenvalue Problems with Implicitly Restarted Arnoldi methods.* SIAM Publications, Philadelphia, PA, 1998.

[Sor92]   D. C. Sorensen. Implicit application of polynomial filters in a k-step Arnoldi method. *SIAM Journal on Matrix Analysis and Applications*, 13:357–385, 1992.

[SSS]   H. A. Smith, R. K. Singh, and D. C. Sorensen. A Lanczos-based eigensolution technique for exact vibration analysis. *International Journal for Numerical Methods in Engineering*, 36:1987–2000.

Part V

Inverse Problems

A Fixed Domain Approach in Shape Optimization Problems with Neumann Boundary Conditions

Pekka Neittaanmäki[1] and Dan Tiba[2]

[1] University of Jyväskylä, Department of Mathematical Information Technology, P.O. Box 35 (Agora), FI-40014 University of Jyväskylä, Finland `pn@mit.jyu.fi`
[2] Institute of Mathematics, Romanian Academy, P.O. Box 1-764, RO-014700 Bucharest, Romania `dan.tiba@imar.ro`

Summary. Fixed domain methods have well-known advantages in the solution of variable domain problems, but are mainly applied in the case of Dirichlet boundary conditions. This paper examines a way to extend this class of methods to the more difficult case of Neumann boundary conditions.

1 Introduction

Starting with the well-known monograph of Pironneau [Pir84], shape optimization problems are subject to very intensive research investigations. They concentrate several major mathematical difficulties: unknown and possibly non-smooth character of optimal geometries, lack of convexity of the functional to be minimized, high complexity and stiff character of the equations to be solved numerically, etc. Accordingly, the relevant scientific literature is huge and we quote here just the books of Mohammadi and Pironneau [MP01] and of Neittaanmäki, Sprekels and Tiba [NST06] for an introduction to this domain of mathematics.

In this paper, we study the model optimal design problem

$$\text{Min} \int_{\Omega} j(x, y(x)) \, dx \tag{1}$$

subject to the Neumann boundary value problem

$$\int_{\Omega} \left[\sum_{i,j=1}^{d} a_{ij} \frac{\partial y}{\partial x_i} \frac{\partial v}{\partial x_j} + a_0 y v \right] \, dx = \int_{\Omega} f v \tag{2}$$

for any $v \in H^1(\Omega)$.

Here, $\Omega \subset D \subset \mathbb{R}^d$ is an unknown domain (the minimization parameter), while D is a fixed smooth open set in the Euclidean space $\mathbb{R}^d$. The functions a_0 and a_{ij} are in $L^\infty(D)$ and $f \in L^2(D)$, that is (2) makes sense for any Ω admissible and defines, as it is well known, the unique weak solution $y = y_\Omega \in H^1(\Omega)$ of the second order elliptic equation

$$-\sum_{i,j=1}^{d} \frac{\partial}{\partial x_j}\left(a_{ij}\frac{\partial y}{\partial x_i}\right) + a_0 y = f \quad \text{in } \Omega \tag{3}$$

with Neumann boundary conditions for the conormal derivative

$$\frac{\partial y}{\partial n_A} = \sum_{i,j=1}^{d} a_{ij}\frac{\partial y}{\partial x_j}\cos(\bar{n}, x_i) = 0 \quad \text{on } \partial\Omega. \tag{4}$$

In the classical formulation (3), (4), $\partial\Omega$ has to be assumed smooth and $\bar{n}$ is the (outward) normal to $\partial\Omega$ in the considered points $x = (x_1, x_2, ..., x_d)$. Non-homogeneous Neumann problems (i.e. with the right-hand side non-zero in (4)) may be considered as well by a simple translation argument reducing everything to the homogeneous case.

The functional $j : D \times \mathbb{R} \to \mathbb{R}$ is a general convex integrand in the sense of Rockafellar [Roc70] – more assumptions will be added when necessary.

The open set Ω will be "parametrized" by some continuous function $g : D \to \mathbb{R}$ by

$$\Omega = \Omega_g = \text{int}\{x \in D \mid g(x) \geq 0\} \tag{5}$$

and $g \in C(\bar{D})$ will be the true unknown of the optimization problem (1), (2). The parametrization is, of course, non-unique, but this does not affect the argument. Arbitrary Caratheodory open sets $\Omega \subset D$ may be expressed in the form Ω_g if g is the signed distance function (at some power). Further constraints on $\Omega = \Omega_g$ (beside $\Omega \subset D$) may be imposed in the abstract form

$$g \in C, \tag{6}$$

where $C \subset C(\bar{D})$ is some convex closed subset. For instance, if $E \subset D$ is a given subset and $C = \{g \in C(\bar{D}) \mid g(x) \geq 0, \ x \in E\}$, then the constraint $g \in C$ is equivalent with the condition $E \subset \Omega$. Other cost functionals may be studied as well:

$$\int_E j(x, y(x))\, dx$$

(if the constraint $E \subset \Omega$ is imposed) or

$$\int_\Gamma j(x, y(x))\, dx,$$

where $\Gamma \subset D$ is a smooth given manifold and $\Omega \supset \Gamma$ for all admissible Ω. Robin boundary conditions (instead of (4)) may be also discussed by our

method. In the case of Dirichlet boundary conditions other approaches may be used [NPT07, NT95, Tib92].

In Section 2 we recall some geometric controllability properties that are at the core of our approach, while Section 3 contains the basic arguments. The paper ends with some brief Conclusions.

2 A Controllability-Like Result

In the classical book of Lions [Lio68], it is shown that, when $u \in L^2(\Gamma_1)$ is arbitrary and y_u is the unique solution (in the transposition sense) of

$$-\Delta y = 0 \quad \text{in } G,$$

$$y = u \quad \text{on } \Gamma_1, \quad y = 0 \quad \text{on } \Gamma_2,$$

then the set of normal traces $\{\frac{\partial y_u}{\partial n} \mid u \in L^2(\Gamma_1)\}$ is linear and dense in the space $H^{-1}(\Gamma_2)$. Notice that $\frac{\partial y_u}{\partial n} \in H^{-1}(\Gamma_2)$ due to some special regularity results, Lions [Lio68]. Here $G \subset \mathbb{R}^d$ is an open connected set such that its boundary $\partial G = \Gamma_1 \cup \Gamma_2$ and $\bar{\Gamma}_1 \cap \bar{\Gamma}_2 = \emptyset$. This density result may be interpreted as an approximate controllability property in the sense that the "attainable" set of normal derivatives $\frac{\partial y_u}{\partial n}$ (when u ranges in $L^2(\Gamma_1)$) may approximate any element in the "image" space $H^{-1}(\Gamma_2)$. Constructive approaches, results involving constraints on the boundary control u are reported in [NST06, Ch. 5.2].

We continue with a distributed approximate controllability property, which is a constructive variant of Theorem 5.2.21 in [NST06]. We consider the equation (2) in D and with a modified right-hand side:

$$\int_D \left[\sum_{i,j=1}^{d} a_{ij} \frac{\partial \tilde{y}}{\partial x_i} \frac{\partial \tilde{v}}{\partial x_j} + a_0 \tilde{y}\tilde{v} \right] dx = \int_D \chi_0 u\tilde{v}\, dx \quad \forall \tilde{v} \in H^1(D), \qquad (7)$$

where $u \in L^2(D)$ is a distributed control and χ_0 is the characteristic function of some smooth open set $\Omega_0 \subset D$ such that $\partial D \subset \bar{\Omega}_0$. That is, Ω_0 is a relative neighborhood of ∂D and we denote $\Gamma = \partial \Omega_0 \setminus \partial D$. Clearly, $\bar{\Gamma} \cap \partial D = \emptyset$.

Theorem 1. *Let $w \in H^{1/2}(\Gamma)$ be given and let $[u_\varepsilon, y_\varepsilon]$ be the unique optimal pair of the control problem:*

$$\operatorname*{Min}_{u \in L^2(\Omega_0)} \left\{ \frac{1}{2}|y - w|_{H^{1/2}(\Gamma)} + \frac{\varepsilon}{2}|u|^2_{L^2(\Omega_0)} \right\}, \quad \varepsilon > 0, \qquad (8)$$

$$\int_\Omega \left[\sum_{i,j=1}^{d} a_{ij} \frac{\partial y}{\partial x_i} \frac{\partial z}{\partial x_j} + a_0 yz \right] dx = \int_{\Omega_0} uz\, dx \quad \forall z \in H^1(\Omega_0). \qquad (9)$$

Then, we have

$$y_\varepsilon|_\Gamma \xrightarrow[\varepsilon \to 0]{} w \quad \text{strongly in } H^{1/2}(\Gamma). \qquad (10)$$

Proof. The existence and the uniqueness of the optimal pair $[u_\varepsilon, y_\varepsilon] \in L^2(\Omega_0) \times H^1(\Omega_0)$ of the control problem (8), (9) is obvious. The pair $[0,0]$ is clearly admissible and, for any $\varepsilon > 0$, we obtain

$$\frac{1}{2}|y_\varepsilon - w|^2_{H^{1/2}(\Gamma)} + \frac{\varepsilon}{2}|u_\varepsilon|^2_{L^2(\Omega_0)} \le \frac{1}{2}|w|^2_{H^{1/2}(\Gamma)}.$$

Therefore, $\{y_\varepsilon\}$ and $\{\varepsilon^{1/2}u_\varepsilon\}$ are bounded respectively in $H^{1/2}(\Gamma)$, $L^2(\Omega_0)$. We denote by $l \in H^{1/2}(\Gamma)$ the weak limit (on a subsequence) of $\{y_\varepsilon - w\}$.

Let us define the adjoint system by:

$$\int_{\Omega_0} \left[\sum_{i,j=1}^d a_{ij} \frac{\partial z}{\partial x_i} \frac{\partial p_\varepsilon}{\partial x_j} + a_0 z p_\varepsilon \right] dx = \int_\Gamma (y_\varepsilon - w) z \, d\sigma \quad \forall z \in H^1(\Omega_0), \quad (11)$$

which is a non-homogeneous Neumann problem and $p_\varepsilon \in H^1(\Omega_0)$. We also introduce the equation in variations

$$\int_{\Omega_0} \left[\sum_{i,j=1}^d a_{ij} \frac{\partial \mu}{\partial x_i} \frac{\partial z}{\partial x_j} + a_0 \mu z \right] dx = \int_{\Omega_0} \nu z \, dx \quad \forall z \in H^1(\Omega_0), \quad (12)$$

which defines the variations $y_\varepsilon + \lambda\mu$, $u_\varepsilon + \lambda\nu$ for any $\nu \in L^2(\Omega_0)$ and $\lambda \in \mathbb{R}$.

A standard computation using (11), (12) and the optimality of $[u_\varepsilon, y_\varepsilon]$ gives

$$0 = \varepsilon(u_\varepsilon, \nu)_{L^2(\Omega_0)} + (y_\varepsilon - w, \mu)_{H^{1/2}(\Gamma)}$$

$$= \varepsilon(u_\varepsilon, \nu)_{L^2(\Omega_0)} + \int_{\Omega_0} \left[\sum_{i,j=1}^d a_{ij} \frac{\partial \mu}{\partial x_i} \frac{\partial p_\varepsilon}{\partial x_j} + a_0 \mu p_\varepsilon \right] dx$$

$$= \varepsilon(u_\varepsilon, \nu)_{L^2(\Omega_0)} + (p_\varepsilon, \nu)_{L^2(\Omega_0)}. \quad (13)$$

Due to the convergence properties of the right-hand side in (11), $\{p_\varepsilon\}$ is bounded in $H^1(\Omega_0)$ and we can pass to the limit (on a subsequence) $p_\varepsilon \to p$ weakly in $H^1(\Omega_0)$, to obtain

$$\int_{\Omega_0} \left[\sum_{i,j=1}^d a_{ij} \frac{\partial z}{\partial x_i} \frac{\partial p}{\partial x_j} + a_0 z p \right] dx = \int_\Gamma l z \, d\sigma \quad \forall z \in H^1(\Omega_0). \quad (14)$$

The passage to the limit in (13), as $\{\varepsilon^{1/2}u_\varepsilon\}$ is bounded, gives that $p \equiv 0$ in Ω_0 and (14) shows that $l = 0$ in Γ.

We have proved (10) in the weak topology of $H^{1/2}(\Gamma)$. The strong convergence is a consequence of the Mazur theorem [Yos80] and of a variational argument.

Remark 1. The Mazur theorem alone and the linearity of (9) produces a sequence $\tilde{u}_\varepsilon$ (of convex combinations of u_ε) such that the corresponding sequence of states $\tilde{y}_\varepsilon$ satisfies (10). Theorem 1 gives a constructive answer to the approximate controllability property.

If Ω_0 is smooth enough and $w \in H^{3/2}(\Gamma)$, then the trace theorem ensures the existence of $\hat{y} \in H^2(\Omega_0)$ such that $\frac{\partial \hat{y}}{\partial n_A} = 0$ (null conormal derivative) and $\hat{y}|_\Gamma = w$. That is, the control

$$\hat{u} = -\sum_{i,j=1}^d \frac{\partial}{\partial x_j}\left(a_{ij}\frac{\partial \hat{y}}{\partial x_i}\right) + a_0 \hat{y}$$

ensures the exact controllability property. Notice that $\hat{u}$ is not unique since any element in $H_0^2(\Omega_0)$ may be added to $\hat{y}$ with all the properties being preserved.

3 A Variational Fixed Domain Formulation

We assume that $\Omega = \Omega_g$, where $g \in C(\bar{D})$, is as in (5). Motivated by the result in the previous section, we consider the following homogeneous Neumann problem in D:

$$-\sum_{i,j=1}^d \frac{\partial}{\partial x_j}\left(a_{ij}\frac{\partial \tilde{y}}{\partial x_i}\right) + a_0 \tilde{y} = f + (1 - H(g))u \quad \text{in } D, \qquad (15)$$

$$\frac{\partial y}{\partial n_A} = 0 \qquad\qquad\qquad \text{on } \partial D. \qquad (16)$$

Here $H(\cdot)$ is the Heaviside function in $\mathbb{R}$ and $H(g)$ is, consequently, the characteristic function of Ω_g. Under conditions of Theorem 1, the restriction $y = \tilde{y}|_{\Omega_g}$ is the solution of (2) in $\Omega = \Omega_g$. Moreover, since $g = 0$ on $\partial\Omega_g$, under smoothness conditions, ∇g is parallel to $\bar{n}$, the normal to $\partial\Omega_g$. Then, we can rewrite (4) as

$$\sum_{i,j=1}^d a_{ij}\frac{\partial y}{\partial x_j}\nabla g \cdot e_i = 0 \quad \text{on } \partial\Omega_g, \qquad (17)$$

where we use that $\cos(\bar{n}, x_i) = \cos(\nabla g, x_i)$ and e_i is the vector of the axis x_i.

If the elliptic operator is the Laplace operator, then (17) becomes simply

$$\nabla g \cdot \nabla y = 0 \quad \text{on } \partial\Omega_g.$$

In order to fix a unique $u \in L^2(D)$ satisfying to (15), (16), (17), we define the following optimal control problem with state constraints:

$$\operatorname*{Min}_{u \in L^2(D)}\left\{\frac{1}{2}\int_D u^2\, dx\right\}, \qquad (18)$$

governed by the state system (15), (16) and subject to the state constraint (17).

The discussion in Section 2 shows the existence of infinitely many admissible pairs $[u, y]$ for the constrained control problem (15)–(18). (Here g is fixed satisfying the necessary smoothness properties.)

In case g and $\Omega_g \subset D$ are variable and unknown, we say that (15)–(18) is the variational fixed domain (in D!) formulation of the Neumann boundary value problem. One can write the optimality conditions that give a system of equations equivalent with (15)–(18) and extend the Neumann problem from Ω_g to D.

We introduce the penalized control problem, for $\varepsilon > 0$, as follows (here $[g \equiv 0]$ denotes $\partial \Omega_g$):

$$\underset{u \in L^2(D)}{\text{Min}} \left\{ \frac{1}{2} \int_D u^2 \, dx + \frac{1}{2\varepsilon} \int_{[g \equiv 0]} F(y_\varepsilon)^2 \, d\sigma \right\} \tag{19}$$

subject to

$$-\sum_{i,j=1}^{d} \frac{\partial}{\partial x_j} \left(a_{ij} \frac{\partial y_\varepsilon}{\partial x_i} \right) + a_0 y_\varepsilon = f + (1 - H(g))u \quad \text{in } D, \tag{20}$$

$$\frac{\partial y_\varepsilon}{\partial n_A} = 0 \qquad\qquad\qquad \text{on } \partial D. \tag{21}$$

Above,

$$F(y) = \sum_{i,j=1}^{d} a_{ij} \frac{\partial y}{\partial x_j} \nabla g \cdot e_i$$

and the problem (19)–(21), which is unconstrained, remains a coercive and strictly convex control problem. That is, we have the existence and the uniqueness of the approximating optimal pair $[u_\varepsilon, y_\varepsilon] \in L^2(D) \times H^2(D)$ (if ∂D is smooth enough).

Proposition 1. *We have*

$$|F(y_\varepsilon)|_{L^2(\partial \Omega_g)} \leq C \varepsilon^{\frac{1}{2}}, \tag{22}$$

$$u_\varepsilon \to \hat{u} \quad \text{strongly in } L^2(D), \tag{23}$$

$$y_\varepsilon \to \hat{y} \quad \text{strongly in } H^2(D), \tag{24}$$

where C is a constant independent of $\varepsilon > 0$ and $[\hat{u}, \hat{y}] \in L^2(D) \times H^2(D)$ is the unique optimal pair of (15)–(18).

Proof. As in Section 2, by the trace theorem, we may choose $\tilde{y} \in H^2(D \setminus \Omega_g)$ with the property that $\frac{\partial \tilde{y}}{\partial n_A} = 0$ in $\partial(D \setminus \Omega_g)$ and $\tilde{y}$ may be extended to the solution of (2) inside Ω_g. We can compute $\tilde{u} \in L^2(D \setminus \Omega_g)$ by (20) and extend it by 0 inside Ω_g. Then $[\tilde{u}, \tilde{y}]$ is an admissible pair for the control problem (19)–(21) and, by the optimality of $[u_\varepsilon, y_\varepsilon]$, we get

$$\frac{1}{2}\int_D u_\varepsilon^2\,dx + \frac{1}{2\varepsilon}\int_{[g\equiv 0]} F(y_\varepsilon)^2\,d\sigma \le \frac{1}{2}\int_D \tilde u^2\,dx \qquad (25)$$

since $F(\tilde y) = 0$ in $\partial\Omega_g$.

The inequality (25) gives (22) and $\{u_\varepsilon\}$ bounded in $L^2(D)$. By (20), (21), $\{y_\varepsilon\}$ is bounded in $H^2(D)$ and, on a subsequence, we have $y_\varepsilon \to \hat y$, $u_\varepsilon \to \hat u$ weakly in $H^2(D)$, respectively in $L^2(D)$, where $[\hat u, \hat y]$ again satisfy (20), (21). Moreover, one can pass to the limit in (22) with $\varepsilon \to 0$, to see that $F(\hat y) = 0$ in $\partial\Omega_g$. This shows that $[\hat u, \hat y]$ is an admissible pair for the original state constrained control problem (15)–(18). For any admissible pair $[\mu, z] \in L^2(D) \times H^2(D)$ of (15)–(18), we have $F(z) = 0$ on $\partial\Omega_g$ and the inequality (25) is valid with $\tilde u$ replaced by μ and we infer

$$\frac{1}{2}\int_D u_\varepsilon^2\,dx \le \frac{1}{2}\int_D \mu^2\,dx.$$

The weak lower semicontinuity of the norm gives

$$\frac{1}{2}\int_D (\hat u)^2\,dx \le \frac{1}{2}\int_D \mu^2\,dx,$$

that is, the pair $[\hat u, \hat y]$ is, in fact, the unique optimal pair of (15)–(18) and we also have

$$\lim_{\varepsilon \to 0}\int_D u_\varepsilon^2\,dx = \int_D (\hat u)^2\,dx.$$

Then $u_\varepsilon \to \hat u$ strongly in $L^2(D)$ and $y_\varepsilon \to \hat y$ strongly in $H^2(D)$ by the strong convergence criterion in uniformly convex spaces. The convergence is valid without taking subsequences due to the uniqueness of $[\hat u, \hat y]$.

Remark 2. One can further regularize H in (20), by replacing it with a mollification H^ε of the Yosida approximation H_ε of the maximal monotone extension of H.

Remark 3. One may take in D even null Dirichlet boundary conditions instead of (16). Similar distributed controllability properties (approximate or exact) may be established in very much the same way.

To write shortly, we consider the case of the Laplace operator. The penalized and regularized problem is the following:

$$\underset{u \in L^2(D)}{\mathrm{Min}}\left\{ \frac{1}{2}\int_D u^2\,dx + \frac{1}{2\varepsilon}\int_{[g\equiv 0]} [\nabla y \cdot \nabla g]^2\,d\sigma \right\},$$

$$-\Delta y + y = f + (1 - H^\varepsilon(g))u \qquad \text{in } D,$$

$$y = 0 \qquad \text{on } \partial D.$$

Here, the control u ensures the "transfer" from Dirichlet to Neumann (null) conditions on $\partial\Omega_g$ and all the results are similar as for the Neumann–Neumann case.

Theorem 2. *The gradient of the cost functional (19) with respect to* $u \in L^2(D)$ *is given by*

$$\nabla J(u_\varepsilon) = u_\varepsilon + (1 - H(g))p_\varepsilon \quad in \ D, \tag{26}$$

where $p_\varepsilon \in L^2(D)$ *is the unique solution of the adjoint equation*

$$\int_D p_\varepsilon \left[-\sum_{i,j=1}^d \frac{\partial}{\partial x_j}\left(a_{ij}\frac{\partial z}{\partial x_i}\right) + a_0 z \right] dx = \frac{1}{\varepsilon} \int_{[g\equiv 0]} F(y_\varepsilon)F(z)\, d\sigma$$

$$\forall z \in H^2(D), \ \frac{\partial z}{\partial n_A} = 0 \ on \ \partial D, \tag{27}$$

in the sense of transpositions.

Proof. We discuss first the existence of the unique transposition solution to (27).

The equation in variations corresponding to (20), (21) is

$$-\sum_{i,j=1}^d \frac{\partial}{\partial x_j}\left(a_{ij}\frac{\partial z}{\partial x_i}\right) + a_0 z = (1 - H(g))v \quad in \ D, \tag{28}$$

$$\frac{\partial z}{\partial n_A} = 0 \qquad\qquad on \ \partial D, \tag{29}$$

for any $v \in L^2(D)$. By regularity theory for differential equations, the unique solution of (28), (29) satisfies $z \in H^2(D)$.

We perturb this equation by adding δv, $\delta > 0$, in the right-hand side and we denote by z_δ the corresponding solution, $z_\delta \in H^2(D)$. The mapping $v \to z_\delta$, as constructed above, is an isomorphism $T_\delta : L^2(D) \to W = \{z \in H^2(D) \mid \frac{\partial z}{\partial n_A} = 0 \text{ on } \partial D\}$.

We define the linear continuous functional on $L^2(D)$ by

$$v \longrightarrow \frac{1}{\varepsilon} \int_{[g\equiv 0]} F(y_\varepsilon)F(T_\delta v)\, d\sigma \quad \forall v \in L^2(D). \tag{30}$$

The Riesz representation theorem applied to (30) ensures the existence of a unique $\tilde{p}_\delta \in L^2(D)$ such that

$$\int_D \tilde{p}_\delta v = \frac{1}{\varepsilon} \int_{[g\equiv 0]} F(y_\varepsilon)F(T_\delta v)\, d\sigma \quad \forall v \in L^2(D). \tag{31}$$

Choosing $v = T_\delta^{-1}z$, $z \in W$ arbitrary, the relation (31) gives

$$\int_D \tilde{p}_\delta (1 - H(g) + \delta)^{-1}\left(-\sum_{i,j=1}^d \frac{\partial}{\partial x_j}\left(a_{ij}\frac{\partial z}{\partial x_i}\right) + a_0 z\right) dx$$

$$= \frac{1}{\varepsilon} \int_{[g\equiv 0]} F(y_\varepsilon)F(z)\, d\sigma \quad \forall z \in W. \tag{32}$$

By redenoting $p_\varepsilon = \tilde{p}_\delta (1 - H(g) + \delta)^{-1} \in L^2(D)$ (which conceptually may depend on $\delta > 0$) in (32) we have proved the existence for (27). The uniqueness of p_ε may be shown by contradiction, directly in (27), as the factor multiplying p_ε in the left-hand side of (27) "generates" the whole $L^2(D)$ when $z \in W$ is arbitrary.

Coming back to the equation in variations (28), (29) and to the definition of the control problem (19)–(21), the directional derivative of the cost functional (19) is given by

$$\lim_{\lambda \to 0} \frac{1}{\lambda}[J(u_\varepsilon + \lambda v) - J(u_\varepsilon)] = \int_D u_\varepsilon v \, dx + \frac{1}{\varepsilon} \int_{[g \equiv 0]} F(y_\varepsilon) F(z) \, d\sigma \qquad (33)$$

and the Euler equation is

$$0 = \int_D u_\varepsilon v \, dx + \frac{1}{\varepsilon} \int_{[g \equiv 0]} F(y_\varepsilon) F(z) \, d\sigma \quad \forall v \in L^2(D) \qquad (34)$$

with z defined by (28), (29). By using (27) in (34), since z given by (28), (29) is an admissible test function, we get

$$0 = \int_D u_\varepsilon v \, dx + \int_D p_\varepsilon \left[-\sum_{i,j=1}^{d} \frac{\partial}{\partial x_j} \left(a_{ij} \frac{\partial z}{\partial x_i} \right) + a_0 z \right] dx$$

$$= \int_D u_\varepsilon v \, dx + \int_D p_\varepsilon (1 - H(g)) v \, dx. \qquad (35)$$

This proves (26) and ends the argument.

Remark 4. Theorem 2 may be applied for any control $u \in L^2(D)$. For the optimal control u_ε, the directional derivative (and the gradient) is null and we obtain $u_\varepsilon = -p_\varepsilon (1 - H(g))$, that is, u_ε has support in $D \setminus \Omega_g$. This relation is the maximum (Pontryagin) principle applied to the control problem (19)–(21). Moreover, one can eliminate u_ε and write the following system of two elliptic equations:

$$-\sum_{i,j=1}^{d} \frac{\partial}{\partial x_j} \left(a_{ij} \frac{\partial y_\varepsilon}{\partial x_i} \right) + a_0 y_\varepsilon = f - (1 - H(g))^2 p_\varepsilon \quad \text{in } D, \qquad (36)$$

$$\frac{\partial y_\varepsilon}{\partial n_A} = 0 \qquad \qquad \text{on } \partial D,$$

$$\int_D p_\varepsilon \left[-\sum_{i,j=1}^{d} \frac{\partial}{\partial x_j} \left(a_{ij} \frac{\partial z}{\partial x_i} \right) + a_0 z \right] = \frac{1}{\varepsilon} \int_{[g \equiv 0]} F(y_\varepsilon) F(z) \, d\sigma \quad \forall z \in W,$$

$$(37)$$

which constructs in an explicit manner the extension of the Neumann boundary value problem from Ω_g to D, modulo the approximation discussed in Proposition 1.

4 Conclusions

The shape optimization problem (1), (2) is transformed in this way into the optimal control problem

$$\underset{g \in C}{\text{Min}} \int_D H(g) j(x, y(x)) \, dx \qquad (38)$$

subject to (15)–(17) which, in turn, may be approximated by (19)–(21) or, equivalently, by (36)–(37). To obtain good differentiability properties with respect to g in the optimization problem (38), one should replace H by H^ε, some regularization of H, as previously mentioned. Analyzing further approximation properties and the gradient for (38) is a nontrivial task. However, the application of evolutionary algorithms is possible since it involves just the values of the cost (38) and no computation of the gradient with respect to g.

As initial population of controls g for the genetic algorithm, corresponding to the finite element mesh in D, one may use the basis functions for the piecewise linear and continuous finite element basis. In case some supplementary information is available on the desired shape (for instance, coming from the constraints), this should be imposed on the initial population. Then, standard procedures specific to evolutionary algorithms [Hol75] are to be applied.

References

[Hol75]   J. R. Holland. *Adaptation in natural and artificial systems.* The University of Michigan Press, Ann Arbor, MI, 1975.

[Lio68]   J.-L. Lions. *Contrôle optimal des systèmes gouvernées par des equations aux dérivées partielles.* Dunod, Paris, 1968.

[MP01]   B. Mohammadi and O. Pironneau. *Applied shape optimization for fluids.* The Clarendon Press, Oxford University Press, New York, 2001.

[NPT07]   P. Neittaanmäki, A. Pennanen, and D. Tiba. Fixed domain approaches in shape optimization problems with Dirichlet boundary conditions. Reports of the Department of Mathematical Information Technology, Series B, Scientific Computing B16/2007, University of Jyväskylä, Jyväskylä, 2007.

[NST06]   P. Neittaanmäki, J. Sprekels, and D. Tiba. *Optimization of elliptic systems.* Springer-Verlag, Berlin, 2006.

[NT95]   P. Neittaanmäki and D. Tiba. An embedding of domains approach in free boundary problems and optimal design. *SIAM J. Control Optim.*, 33(5):1587–1602, 1995.

[Pir84]   O. Pironneau. *Optimal shape design for elliptic systems.* Springer-Verlag, Berlin, 1984.

[Roc70]   R. T. Rockafellar. *Convex analysis.* Princeton University Press, Princeton, NJ, 1970.

[Tib92]   D. Tiba. Controllability properties for elliptic systems, the fictitious domain method and optimal shape design problems. In *Optimization, optimal control and partial differential equations (Iaşi, 1992)*, number 107 in Internat. Ser. Numer. Math., pages 251–261, Basel, 1992. Birkhäuser.

[Yos80]   K. Yosida. *Functional analysis.* Springer-Verlag, Berlin, 1980.

Reduced-Order Modelling of Dispersion

Jean-Marc Brun[1] and Bijan Mohammadi[2]

[1] CEMAGREF/ITAP, FR-34095 Montpellier, France
   `jean-marc.brun@cemagref.fr`
[2] I3M-Univ. Montpellier II, CC051, FR-34095 Montpellier, France
   `bijan.mohammadi@univ-montp2.fr`

Summary. We present low complexity models for the transport of passive scalars for environmental applications. Multi-level analysis has been used with a reduction in dimension of the solution space at each level. Similitude solutions are used in a non-symmetric metric for the transport over long distances. Model parameters identification is based on data assimilation. The approach does not require the solution of any PDE and, therefore, is mesh free. The model also permits to access the solution in one point without computing the solution over the whole domain. Sensitivity analysis is used for risk analysis and also for the identification of the sources of an observed pollution.

Key words: Reduced order modelling, source identification, risk analysis by sensitivity, non-symmetric geometry.

1 Introduction

Air and water contamination by pesticides is a major preoccupation for health and environment. One aims to model pesticide transport in atmospheric flows with very low calculation cost making assimilation-simulation and statistic risk analysis by Monte Carlo simulations realistic. In this problem available data is incomplete with large variability and the number of parameters involved large. Solution space reduction and reduced order modelling appear, therefore, as natural way to proceed.

Our contribution is to build a multi-level approach where a given level provides the inlet condition for the level above. In each level one aims to use a priori information in the definition of the search space for the solution and avoid the solution of partial differential equations.

More precisely, a near field (to the injection device) search space is build using experimental observations. Once this local solution known, the amount of specie leaving the atmospheric sub-layer is evaluated. This quantity is candidate for long distance transport using similitude solutions for mixing layers and plumes [Sim97]. These are known in Cartesian metrics. An original

contribution here is the generalization of these solutions in a non-symmetric travel-time based metric to account for non-uniform winds. We add constraint such that solutions built with this approach to be solution of the direct model (i.e. flow equations and transport model for a passive scalar). In particular, the divergence free condition for the generated winds, conservation, positivity and linearity of the solution of transport equations are requested.

Numerical examples show a comparison of this approach with a PDE based simulation. Examples also show multi-source configurations as well as sensitivity analysis of detected pollution. This is useful for both source identification and risk analysis.

2 Reduced-Order Modelling

One aims to model very large multi-scale phenomenon present in agricultural phyto treatment of cultures. The different entities to account for range from rows of plants to water attraction basins and one should also consider local topography and atmospheric conditions. It is, therefore, obvious that modelling phenomenon falling in length scales below a few meters becomes inevitable.

Consider the calculation of a state variable $V(p)$, function of independent variables p. Our aim is to define a suitable search space for the solution $V(p)$ instead of considering a general function space. This former approach is what one does in finite element methods, for instance, where the solution is expressed in some subspace $S(\{W_N\})$ described by the functional basis chosen $\{W_N\}$, with the quality of the solution being monitored either through the mesh quality and/or increasing the order of the finite element [Cia78]. In all cases, the size of the problem is large $1 \ll N < \infty$ and if the approach is consistent, the projected solution tends to the exact solution when $N \to \infty$.

In a low-complexity approach, one replaces the calculation of $V(p)$ by a projection over a subspace $S(\{w_n\})$ generated, for instance, by $\{w_n\}$, a family of solutions ('snapshots') of the initial full model ($p \to V(p)$). In particular, one aims $n \ll N$ [VP05].

In our approach, we aim to remove the calculation of these snapshots as this is not always an easy task. We take advantage of what we know on the physic of the problem and replace the direct model $p \to V(p)$ by an approximate model $p \to v(p)$ easier to evaluate. This is a very natural way to proceed, as often one does not need all the details on a given state. Also it is sufficient for the low-complexity model to have a local validation domain: one does not necessarily use the same low-complexity model over the whole range of the parameters. We have used this approach in the incomplete sensitivity concept where the linearization is performed not for the direct model but for an approximate state equation [MP01].

2.1 Near-Field Solution

The first step is to model the solution at the outlet of the injection device used to expand the phyto treatment in between rows. One important hypothesis is to assume two different time scales based on the injection velocity and the velocity at which the injection source moves. The injection velocity being much higher, one assumes the local concentration at the outlet of the injection device to be established instantaneously. This instantaneous local flow field is devoted to vanish immediately and not to affect the overall atmospheric circulation. This injection velocity is only designed to determine the part of the pollutant leaving near-ground area and being candidate for transport over large distances (see Section 2.2). These are strong hypotheses which seriously reduce the search space for the solution.

One considers a cylindrical local reference frame where z indicates the motion direction for the vehicle in the field. One looks for local injection solutions of the form:

$$\mathbf{u}_l \sim f_1(r)g_1(\theta)(\mathbf{z}h_1(z) + (1 - h_1(z))\mathbf{r}) \quad \text{and} \quad c_l \sim f_2(r)g_2(\theta)h_2(z), \quad (1)$$

where the subscript l reads for local. $\mathbf{r}$ is a unit vector having its origin at the injection point and visiting the unit circle around this point in the plan perpendicular to $\mathbf{z}$. This defines an instantaneous flow field around the injection point. c_l denotes the local distribution of a passive scalar. $f_i(r)$, $i = 1, 2$, are solutions of a control problem for the assimilation of experimental data by a PDE based model obtained by dimension reduction of the Navier–Stokes and transport equations [Fin00, RT81, Sum71, Bru06]. These experimental data show that after injection both the flow velocity and phyto products concentration drop to nearly zero after three rows of vegetation. $g_i(\theta)$, $i = 1, 2$, are Gauss distributions describing the characteristics of the injection device and are provided by the manufacturer. $h_i(z)$, $i = 1, 2$, include the characteristics of the vegetation by assimilation of experimental data and inform on how the density of the vegetation deviates the flow horizontally. $h_1(z) \in [-1, 1]$ is an *erf* function, odd and monotonic increasing, and $h_2(z) \in [0, 1]$ is a Gauss distribution.

At this level, one includes compatibility conditions coming from the governing equations. In particular, one aims for the conservation condition to hold for the concentration of the passive scalar, the flow field to be divergence free and both variables to verify an advection equation:

$$\nabla \cdot \mathbf{u}_l = \mathbf{u}_l \cdot \nabla c_l = 0 \quad \int_{\mathbb{R}^3} c_l \, dv = \text{given.} \quad (2)$$

To summarize, the coefficients in functions f_i, g_i, h_i, $i = 1, 2$, are a solution of an assimilation problem for experimental data under the constraint (2) [Bru06].

From now, one expresses the variables in a global Cartesian reference frame where z denotes the vertical axis.

2.2 Long Range Transport and Non-Symmetric Geometry

The modelling above gives a local distribution for the advected quantities. We are now interested by the quantities candidate for a transport over large distances. We suppose that those are given by

$$c^+(x,y) = \int_{z>H} c_l dz \quad \text{or} \quad c^+(x,y) = u_l^+ c_l,$$

where $H \sim 2 - 3\,\mathrm{m}$ and $u_l^+ = \max(0, (\mathbf{u} \cdot \mathbf{z})/\|\mathbf{u}\|)$. The total quantity being transported is given by

$$C = \int_{\mathbb{R}^2} c^+(x,y) d\sigma,$$

which should be conserved by the reduced-order transport model we would like to build and for which c^+ is the input condition.

One aims now to again reduce the search space for the solution. The primary factors influencing the dispersion of a neutral plume are advection by the wind and turbulent mixing. The simplest model of this process is to assume that the plume advects downwind and spreads out in the horizontal and vertical directions. Hence, the distribution of a passive scalar c, emitted from a given point and transported by a uniform plane flow filed U along x-coordinate, is given by

$$c(x,y,z) = c_c(x) f(\sqrt{y^2 + z^2}, \delta(x)), \tag{3}$$

where

$$c_c(x) \sim \exp(-a(U)x) \quad \text{and} \quad f(\sqrt{y^2 + z^2}, \delta(x)) \sim \exp(-b(U, \delta(x))\sqrt{y^2 + z^2}).$$

c_c is the behavior along the central axis of the distribution and $\delta(x)$ characterizes the thickness of the distribution at a given x-coordinate. An analogy exists with plane or axisymmetric mixing layers and neutral plumes where δ is parabolic for a laminar jet and linear in turbulent cases [Cou89, Sim97]. $a(\cdot)$ is a positive monotonic decreasing function and $b(\cdot, \cdot)$ is positive, monotonic increasing in U and decreasing in δ. In a uniform atmospheric flow field, this solution can be used for the transport of c^+ above.

We would like to generalize this solution in a non-symmetric metric defined by migration times based on the flow field and hence treat the case of variable flow fields.

Nonsymmetric Geometry

In a symmetric geometry the distance function between two points A and B verifies

$$d(A,B) = 0 \Rightarrow A = B, \quad d(A,B) = d(B,A), \quad d(A,B) \leq d(A,C) + d(C,B).$$

But the distance function can be non-uniform with anisotropy (the unit spheres being ellipsoids). In a chosen metric $\mathcal{M}$ the distance between A and B is given by

$$d_{\mathcal{M}}(AB) = \int\limits_0^1 \left({}^t\overrightarrow{AB}\mathcal{M}(A + t\overrightarrow{AB})\overrightarrow{AB} \right)^{1/2} dt,$$

where $\mathcal{M}$ is positive definite and symmetric in symmetric geometries. With $\mathcal{M} = I$, one recovers the Euclidean geometry and variable $\mathcal{M}$ permits to account for anisotropy and non-uniformity of the distance function. We have widely used this approach for mesh adaptation for steady and unsteady phenomenon [AGFM02, HM97, BGM97] linking the metric to the Hessian of the solution. This definition of the metric permits to equi-distribute the interpolation error over a given mesh and, therefore, monitor the quality of the solution.

Consider now the following distance function definition:

Definition 1. *If A is upwind with respect to B then*

$$d(B, A) = \infty \quad and \quad d(A, B) = \int_A^{B^\perp} ds/u = T,$$

where T is the migration time from A to $B^\perp$ along the characteristic passing by A.

u is the local velocity along this characteristic and is, by definition, tangent to the characteristic. $B^\perp$ denotes the projection of B over this characteristic in the Euclidean metric. One supposes that this characteristic is unique, hence avoiding sources and attraction points in the flow field. In case of non-uniqueness of this projection, one chooses the direction of the projection which satisfies best the constraint $\mathbf{u} \cdot \nabla c_g = 0$ in B.

Generalized Plume Solution

Once this distance built, we assume the distribution of a passive scalar transported by a flow $\mathbf{u}$ can be written as:

$$c_g = c_c(d)f(d_E^\perp, \delta(d)). \tag{4}$$

Here the subscript g reads for global and mentions long distance transport. $d_E^\perp$ is the Euclidean distance in the normal direction local to the characteristic at $B^\perp$ (i.e. along direction $BB^\perp$).

Flow Field

One should keep in mind that in realistic configurations, one has very little information on the details of the atmospheric flow compared to the accuracy

one would like for the transport. As an example, the flow will be described probably by less than one point by several square kilometers. We consider the near to ground flow field built from observation data as solution of the following system:

$$\mathbf{u} = \nabla\phi, \quad -\Delta\phi = \sum_{i=1,..,n_{\mathrm{obs}}} \|\nabla\phi(x_i) - \mathbf{u}_{\mathrm{obs}}(x_i)\|, \tag{5}$$

where ϕ is a scalar potential and n_{obs} the number of observation points. The observations are close to the ground at $z = H$ and this construction gives a map of the flow near the ground. This is completed in the vertical direction using generalized wall functions for turbulent flows [MP94, MP06]:

$$(\mathbf{u} \cdot \boldsymbol{\tau})^+ = (\mathbf{u} \cdot \boldsymbol{\tau})/u_\tau = f(z^+) = f(zu_\tau/\nu),$$

where $\boldsymbol{\tau} = \mathbf{u}_H/\|\mathbf{u}_H\|$ is the local tangent unit vector to the ground in the direction of the flow and we assume $(\mathbf{u} \cdot \mathbf{n}(z = H) = 0)$ if n is the normal to the ground. This is a non-linear equation giving u_τ, the friction velocity, knowing $(\mathbf{u} \cdot \boldsymbol{\tau})_H$ and is used, in turn, to define the horizontal velocity $\mathbf{u} \cdot \boldsymbol{\tau} = u_\tau f(z^+)$ for $z > H$. This construction gives two components of the flow and the divergence free condition implies the third component is constant and, therefore, it vanishes as it is supposed zero at $z = H$. This construction can be improved but we find it sufficient for the level of accuracy required. In presence of ground variations, the flow is locally rotated to remain parallel to the ground (see also Section 2.2 for ground variation modelling).

Calculation of Migration Times

As we said, our approach aims to provide the solution at a given point without calculating the whole solution. Being in point B, one needs an estimation of the migration time from the source in A to B using the construction in Section 2.2.

We avoid the construction of characteristics using an iterative polynomial definition for a characteristic $s(t) = (x(t), y(t), z(t))$, $t \in [0, 1]$, starting from a third-order polynomial function verifying for each coordinate:

$$P_n(0) = x_A, \quad P_n(1) = x_B, \quad P'_n(0) = u_A^1, \quad P'_n(1) = u_B^1 \text{ (same for } y \text{ and } z).$$

If $P'_n(\zeta) \neq u^1(x = P_n(\zeta))$ this new point should be assimilated by the construction increasing by one the polynomial order. $\zeta \in]0, 1[$ is chosen randomly.

The migration time is computed over this polynomial approximation of the characteristic. Here we make the approximation $B^\perp = B$ which means the characteristic passing by A passes exactly by B which is unlikely. In a uniform flow, this means we suppose the angle between the central axis and $\mathbf{AB}$ is small (cosine near 1). One introduces, therefore, a correction factor of $2/3 = 0.636$ on the calculated times. This is the stochastic averaged cosine

value for a white noise for angles between 0 and π. Once d is calculated by this procedure one needs to define $d_E^\perp$ which is unknown as $B^\perp$ is unknown. We make the approximation $d_E^\perp \sim d_E(B, B^*)$ where B^* is the projection of B over the vector $\overline{\mathbf{u}}$, the averaged velocity along the polynomial characteristic. This approach gives satisfactory results for smooth atmospheric flow fields which is our domain of interest as no phyto treatments is, in principle, applied when the wind is too strong or if the temperature is too high (e.g., for winds stronger than $20\,\mathrm{km/h}$ and air temperature more than $30°$ C). This also makes that the polynomial construction above gives satisfaction with low order polynomials.

Ground Variations

At this point one accounts for the topography or ground variations $((x, y) \to \psi(x, y))$ in the prediction model above. These are available from digital terrain models (DTM) [Arc06]. Despite this plays an important role in the dispersion process, it is obviously hopeless to target direct simulation based on a detailed ground description. One should mention that ground variations effects are implicitly present in observation data for wind and transport as mentioned in Section 2.2. However, as we said, observations are quite incomplete and to improve the predictive capacity of the model one needs to model the dependency between ground variations and migration time. Therefore, in addition to the mentioned assimilation problem, one scales the migration times used for transport over large distances by a positive monotonic decreasing function $f(\phi)$ with $f(0) = 1$ where

$$\phi = (\nabla_{x,y}\psi \cdot \mathbf{u}_H)/\|\mathbf{u}_H\|.$$

Here $\mathbf{u}_H$ is the 'close to ground' constructed flow field based on the assimilated observations.

3 Parameter and Source Identification

Two types of inverse problems have been treated. The first inverse problem is for parameter identification in the model above assimilating either local experimental data (as described in Section 2.1) or partial data available on wind $\mathbf{u}_{\mathrm{obs}}$ and transported species c_{obs} measured by localized apparatus. In particular, the unknown parameters in our global transport model comes from the solution of a minimization problem for:

$$J(p, \mathbf{u}_{\mathrm{obs}}, c_{\mathrm{obs}}) = \|c(p, \mathbf{u}(p, \mathbf{u}_{\mathrm{obs}})) - c_{\mathrm{obs}}\|, \tag{6}$$

where p gathers all unknown independent parameters in Section 2.2. $\|\cdot\|$ is a discrete L^2-norm over the measurement points. $\mathbf{u}(p)$ is the completion of available wind measurements ($\mathbf{u}_{\mathrm{obs}}$) over the domain described in Section 2.2.

Once the model is established, the second inverse problem of interest is the identification of possible sources of an observed pollution. This region is defined where J'_p is large. In this case, the parameter p is the location of the different sources (cultures).

To solve the minimization problems, we use a semi-deterministic global optimization algorithm based on the solution of the following boundary value problem [MS03, Ivo06, IMSH06]:

$$\begin{cases} p_{\zeta\zeta} + p_\zeta = -J'_p(p(\zeta)), \\ p(0) = p_0, \quad J(p(1)) = J_m = 0, \end{cases} \tag{7}$$

where $\zeta \in [0,1]$ is a fictitious parameter. J_m is the infimum of our inverse problems (here taken as 0). This can be solved using solution techniques for BVPs with free surface to find $p(1)$ realizing the infimum (i.e. $J(p(1)) = J_m$). An analogy can be given with the problem of finding the interface between water and ice which is only implicitly known through the iso-value of zero temperature. In case a local minima is enough, the second boundary condition can be replaced by $J'_p(p(1)) = 0$.

This algorithm requires the sensitivity of the functional with respect to independent variables p. An interesting feature of the present low-cost modelling is that gradients are also available at very low calculation cost. Indeed, sensitivity evaluation for large dimension minimization problems is not an easy task. The most efficient approach is to use an adjoint variable with the difficulty that it requires the development of a specific software. Automatic differentiation brings some simplification, but does not avoid the main difficulty of intermediate states storage, even though check-pointing technique brings some relief [Gri01, CFG$^+$01]. By simplifying the solution of the transport problem, the present approach also addresses this issue.

4 Numerical Results

The application of low complexity transport model to several flow condition is shown. Typical fields of $0.01 \sim 0.1\,\mathrm{km}^2$ have been considered in a region of $400\,\mathrm{km}^2$. Rows are spaced by about 1.5 m. The source of the treatment moves at a speed of around 1 m/s and the injection velocity is taken at 7 to 10 m/s for a typical treatment of $100\,\mathrm{kg/km}^2$. Mono and multi sources situations (Figs. 3 and 4) are considered and examples of the constructed flow field are shown together with the wind measurement points assimilated by the model (Figs. 1 and 4). The transport-based and the Euclidean distances have been reported for a given point in Fig. 2. The impact of ground variations on the advected species is shown in Fig. 5. An example of source identification problem is shown in Fig. 6.
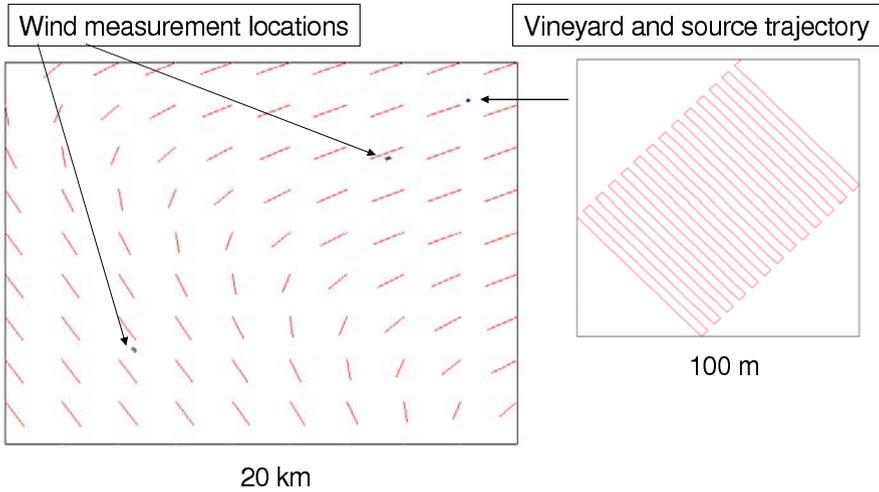
Fig. 1. Typical trajectory of the vehicle in a culture of $10000\,\mathrm{m}^2$ and the location of this field in a calculation domain of $400\,\mathrm{km}^2$. Wind measurements based on two points have been reported together with the constructed divergence free flow field at $z = H \sim 3\,\mathrm{m}$.
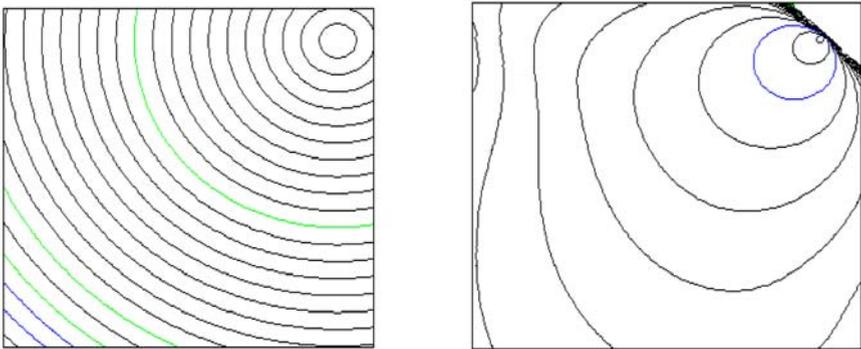
Fig. 2. Examples of symmetric Euclidean and non-symmetric travel time based distances.

5 Concluding Remarks

A low-complexity model has been presented for the prediction of passive scalar dispersion in atmospheric flows for environmental and agricultural applications. The solution search space has been reduced using a priori physical information. A non-symmetric metric based on migration times has been used to generalize injection and plume similitude solutions in the context of variable flow fields. Data assimilation has been used to define the flow field and the parameters in the dispersion model. Sensitivity analysis has been used
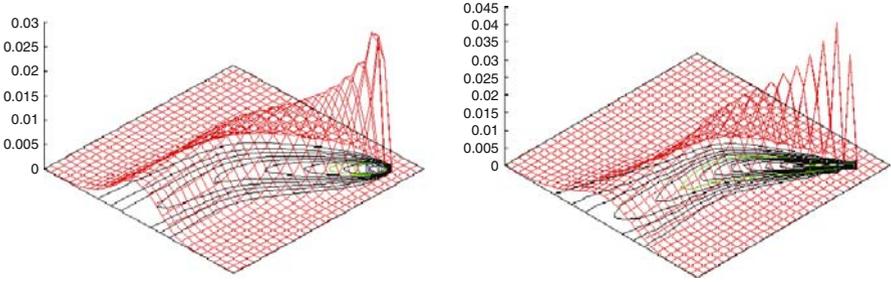
Fig. 3. Generalized similitude solution (right) for a 2-point based wind (similar to Fig. 1) compared to a direct simulation with a PDE based transport-diffusion model for the same wind. The similitude solution has been evaluated on all the nodes of the finite element mesh for comparison.
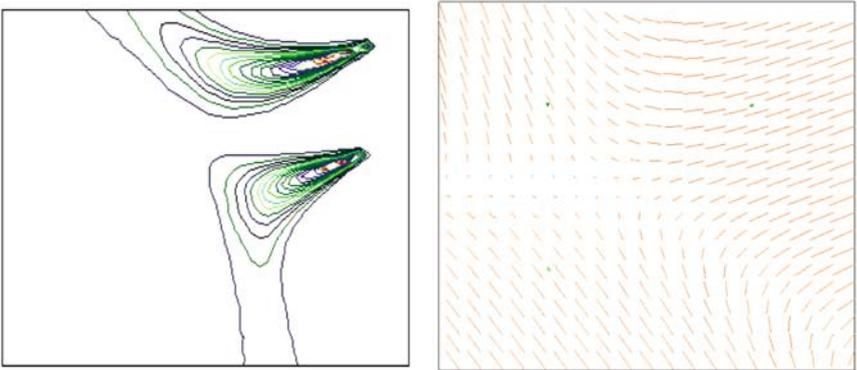
Fig. 4. Regions affected from the treatment of two sources. The flow field has been built from three points of measurement indicated on the picture.

Fig. 5. Left: a typical digital terrain model (x and y coordinates range over 2 km). Dispersion in a uniform north wind with (middle) and without (right) the ground model (Section 2.2).

Fig. 6. Left: constructed flow field. Middle: dispersion from a vineyard. Right: sensitivity analysis for a dispersion detected on the lower left corner. One can, therefore, give possible origins of a pollution.

together with this low-complexity modelling to introduce robustness issues in the prediction. In addition to the data assimilatio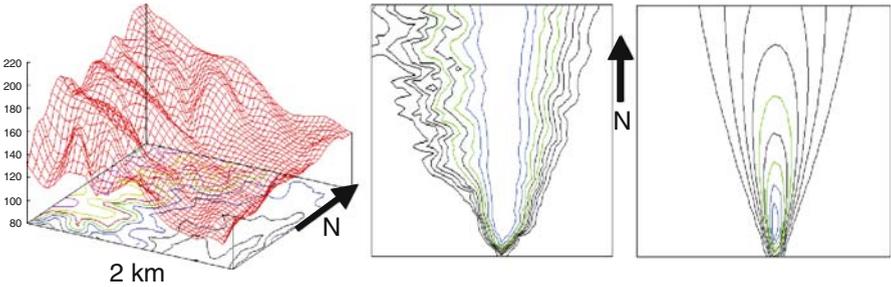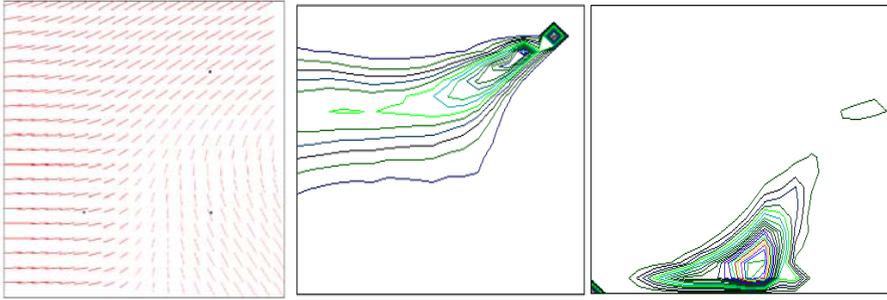n inverse problem, inverse source reconstruction has been considered as a natural demand in environmental survilance. The current work concerns the introduction of stochastic analysis in the present model to produce regional parametric risk maps using Monte Carlo simulations which become achievable thanks to the low calculation cost of the approach.

References

[AGFM02] F. Alauzet, P.-L. George, P. Frey, and B. Mohammadi. Transient fixed point based unstructured mesh adaptation. *Internat. J. Numer. Methods Fluids*, 43(6):729–745, 2002.

[Arc06] ArcGIS. Geographic information system, 2006. http://www.esri.com/software/arcgis.

[BGM97] H. Borouchaki, P.-L. George, and B. Mohammadi. Delaunay mesh generation governed by metric specifications. *Finite Element in Analysis and Design*, 2:85–109, 1997.

[Bru06] J. M. Brun. *Modélisation à complexité réduite de la dérive*. PhD thesis, University of Montpellier, 2006.

[CFG+01] G. Corliss, C. Faure, A. Griewank, L. Hascoet, and U. Naumann, editors. *Automatic differentiation of algorithms: From simulation to optimization*. Number 50 in Lect. Notes Comput. Sci. Eng. Springer, Berlin, 2001. Selected papers from the AD2000 Conference, Nice, France, June 2000.

[Cia78]   Ph. Ciarlet. *The finite element method for elliptic problems.* North-Holland, 1978.

[Cou89]   J. Cousteix. *Turbulence et couche limite.* Cepadues publishers, 1989.

[Fin00]   J. Finnigan. Turbulence in plant canopies. *Annu. Rev. Fluid Mech.*, 32:519–571, 2000.

[Gri01]   A. Griewank. *Computational differentiation.* Springer, New york, 2001.

[HM97]   F. Hecht and B. Mohammadi. Mesh adaptation by metric control for multi-scale phenomena and turbulence. AIAA paper 1997-0859, 1997.

[IMSH06]   B. Ivorra, D. E. Hertzog, B. Mohammadi, and J. G. Santiago. Semi-deterministic and genetic algorithms for global optimization of microfluidic protein-folding devices. *Internat. J. Numer. Methods Engrg.*, 66(2):319–333, 2006.

[Ivo06]   B. Ivorra. *Semi-deterministic global optimization.* PhD thesis, University of Montpellier, 2006.

[MP94]   B. Mohammadi and O. Pironneau. *Analysis of the k-epsilon turbulence model.* Wiley, 1994.

[MP01]   B. Mohammadi and O. Pironneau. *Applied shape optimization for fluids.* Oxford University Press, 2001.

[MP06]   B. Mohammadi and G. Puigt. Wall functions in computational fluid dynamics. *Comput. & Fluids*, 40(3):2101–2124, 2006.

[MS03]   B. Mohammadi and J. H. Saiac. *Pratique de la simulation numérique.* Dunod, Paris, 2003.

[RT81]   M. R. Raupach and A. S. Thom. Turbulence in and above plant canopies. *Annu. Rev. Fluid Mech.*, 13:97–129, 1981.

[Sim97]   J. Simpson. *Gravity currents in the environment and laboratory.* Cambridge University Press, 2nd edition, 1997.

[Sum71]   B. Sumner. A modeling study of several aspects of canopy flow. *Monthly Weather Review*, 99(6):485–493, 1971.

[VP05]   K. Veroy and A. Patera. Certified real-time solution of the parametrized steady incompressible Navier–Stokes equations: Rigorous reduced-basis a posteriori error bounds. *Internat. J. Numer. Methods Fluids*, 47(2):773–788, 2005.

Part VI

Finance (Option Pricing)

Calibration of Lévy Processes with American Options

Yves Achdou[1]

UFR Mathématiques, Université Paris 7, Case 7012, FR-75251 PARIS Cedex 05, France and Laboratoire Jacques-Louis Lions, Université Paris 6, France
`achdou@math.jussieu.fr`

Summary. We study options on financial assets whose discounted prices are exponential of Lévy processes. The price of an American vanilla option as a function of the maturity and the strike satisfies a linear complementarity problem involving a non-local partial integro-differential operator. It leads to a variational inequality in a suitable weighted Sobolev space. Calibrating the Lévy process may be done by solving an inverse least square problem where the state variable satisfies the previously mentioned variational inequality. We first assume that the volatility is positive: after carefully studying the direct problem, we propose necessary optimality conditions for the least square inverse problem. We also consider the direct problem when the volatility is zero.

1 Introduction

Black–Scholes' model [BS73, Mer73] is a continuous time model involving a risky asset (the underlying asset) whose price at time τ is S_τ and a risk-free asset whose price at time τ is $S_\tau^0 = e^{r\tau}$, $r \geq 0$. It assumes that the price of the risky asset satisfies the following stochastic differential equation:

$$dS_\tau = S_\tau(rd\tau + \sigma dW_\tau), \tag{1}$$

where W_τ is a standard Brownian motion on the probability space $(\Omega, \mathcal{A}, \mathbb{P}^*)$ (the probability $\mathbb{P}^*$ is called the risk-neutral probability).

An American vanilla call (resp. put) option on the risky asset is a contract giving its owner the right to buy (resp. sell) a share at a fixed price x at any time before a maturity date t. The price x is called the strike. Exercising the option yields a payoff $\overline{P_\circ}(S) = (S-x)_+$ (resp. $\overline{P_\circ}(S) = (S-x)_-$) for the call (resp. put) option, when the price of the underlying asset is S.

---

[1] I wish to dedicate this work to O. Pironneau with all my friendship. I have been working with Olivier for almost fifteen years now, and for me, it has always been an exciting intellectual and human experience.

European options are similar contracts, except that they can be exercised only at maturity t.

Consider an American option with payoff $\overline{P_\circ}$ and maturity t. Under the assumptions that the market is complete and rules arbitrage out, Black–Scholes' theory predicts that the price of this option at time τ is

$$P_\tau = \sup_{s \in \mathcal{T}_{\tau,t}} \mathbb{E}^* \left(e^{-r(s-\tau)} \overline{P_\circ}(S_s) \middle| F_\tau \right), \tag{2}$$

where $\mathcal{T}_{\tau,t}$ denotes the set of stopping times in $[\tau, t]$ (see [LL97] for the proof of this formula). It can also be proved, see, e.g., [BL84, JLL90] that $P_\tau = P(\tau, S_\tau)$, where the two variables function P is found by solving a parabolic linear complementarity problem

$$\frac{\partial P}{\partial \tau} + \frac{\sigma^2 S^2}{2} \frac{\partial^2 P}{\partial S^2} + rS \frac{\partial P}{\partial S} - rP \leq 0, \quad P(\tau, S) \geq \overline{P_\circ}(S), \quad \tau \in [0, t), \; S > 0,$$

$$\left(\frac{\partial P}{\partial \tau} + \frac{\sigma^2 S^2}{2} \frac{\partial^2 P}{\partial S^2} + rS \frac{\partial P}{\partial S} - rP \right) (P - \overline{P_\circ}(S)) = 0, \quad \tau \in [0, t), \; S > 0,$$

$$P(\tau = t, S) = \overline{P_\circ}(S). \tag{3}$$

The critical parameter in the Black–Scholes model is the volatility σ. Unfortunately, taking σ to be constant and using (2) or (3) often leads to poor predictions of the prices of the options which are available on the markets. One possible fix is to assume that the process driving S_t is a more general Lévy process: Lévy processes are processes with stationary and independent increments which are continuous in probability, see, for example, the book by Cont and Tankov [CT04] and the references therein.

For a Lévy process X_τ on a filtered probability space with probability $\mathbb{P}^*$, the Lévy–Khintchine formula says that there exists a function $\chi : \mathbb{R} \to \mathbb{C}$ such that

$$\mathbb{E}^*(e^{iuX_\tau}) = e^{-\tau\chi(u)}, \tag{4}$$

$$\chi(u) = \frac{\sigma^2 u^2}{2} - i\beta u + \int_{|z|<1} (e^{iuz} - 1 - iuz)\nu(dz) + \int_{|z|>1} (e^{iuz} - 1)\nu(dz), \tag{5}$$

for $\sigma \geq 0$, $\beta \in \mathbb{R}$ and a positive measure ν on $\mathbb{R} \setminus \{0\}$ such that

$$\int_{\mathbb{R}} \min(1, z^2)\nu(dz) < +\infty.$$

The measure ν is called the Lévy measure of X.

We assume that under $\mathbb{P}^*$, the discounted price of the risky asset is a martingale, and that it is represented as the exponential of a Lévy process:

$$e^{-r\tau} S_\tau = S_0 e^{X_\tau}.$$

The fact that the discounted price is a martingale is equivalent to $\mathbb{E}^*(e^{X_\tau}) = 1$, i.e.

$$\int_{|z|>1} e^z \nu(dz) < \infty \quad \text{and} \quad \beta = -\frac{\sigma^2}{2} - \int_{\mathbb{R}} (e^z - 1 - z1_{|z|\leq 1})\nu(dz).$$

We will also assume that $\int_{|z|>1} e^{2z}\nu(dz) < \infty$, so the discounted price is a square integrable martingale.

We note $\overline{B}$ the integral operator:

$$(\overline{B}v)(S) = \int_{\mathbb{R}} \left(v(Se^z) - v(S) - S(e^z - 1)\frac{\partial}{\partial S}v(S) \right) \nu(dz).$$

Consider an American option with payoff $\overline{P_\circ}$ and maturity t: in [BL84], Bensoussan and Lions assumed $\sigma > 0$ and studied the variational inequality stemming from the complementarity problem $P(t, S) = \overline{P_\circ}(S)$, and for $\tau < t$ and $S > 0$,

$$\frac{\partial P}{\partial \tau}(\tau, S) + \frac{\sigma^2 S^2}{2}\frac{\partial^2 P}{\partial S^2}(\tau, S) + rS\frac{\partial P}{\partial S}(\tau, S) - rP(\tau, S) + (\overline{B}P)(\tau, S) \leq 0, \tag{6}$$

$$P(\tau, S) \geq \overline{P_\circ}(S), \tag{7}$$

and

$$\left(\begin{array}{c} \frac{\partial P}{\partial \tau}(\tau, S) + \frac{\sigma^2 S^2}{2}\frac{\partial^2 P}{\partial S^2}(\tau, S) + rS\frac{\partial P}{\partial S}(\tau, S) \\ -rP(\tau, S) + (\overline{B}P)(\tau, S) \end{array} \right) (P(\tau, S) - \overline{P_\circ}(S)) = 0, \tag{8}$$

in suitable Sobolev spaces with decaying weights near $+\infty$ and 0. They proved that the price of the American option is $P_\tau = P(\tau, S_\tau)$. Other approaches with viscosity solutions are possible, see [Pha98], especially in the case $\sigma = 0$. One advantage of the variational methods is that they provide stability estimates. For numerical methods for options on Lévy driven assets, see [MvPS04, MSW04, MNS03, AP05a, CV04, CV03].

In what follows, we assume that the Lévy measure has a density, $\nu(dz) = k(z)dz$. The main goal of the present work is to study a least-square method for calibrating the volatility σ and the jump density k in order to recover the prices of a family of American options available on the market.

We shall focus on a family of vanilla put options indexed by $i \in I$, with maturities t_i and strikes x_i. One observes $S_\circ$ the price of the risky asset and the prices $(\bar{P}_i)_{i\in I}$ of the above-mentioned family of options. We call T the maximal maturity: $T = \max_{i\in I} t_i$.

The first idea is to try to minimize the functional $(\sigma, k) \mapsto \sum_{i\in I} \omega_i|\bar{P}_i - P_i(0, S_\circ)|^2 + J_R(\sigma, k)$ for k and σ in a suitable set, where

- ω_i are positive weights,
- J_R is a suitable regularizing functional,
- the prices $P_i(0, S_\circ)$ are computed by solving problem (6)–(8), with $t = t_i$ and $\overline{P_\circ}(S) = (x_i - S)_+$.

Evaluating the functional requires solving $\#I$ variational inequalities. This approach was chosen in [Ach05, AP05b] for calibrating models of local volatility (i.e. the volatility is a function of t and S) with American options.

In the present case, it is possible to choose a better approach: we call $(\tau, S) \mapsto P(\tau, S, t, x)$ the pricing function for the vanilla American put with maturity t and strike x. Hereafter, we use the notation

$$P_\circ(x) = (x - S)_+. \tag{9}$$

It can be seen that the solution of (6)–(8) is of the form $P(\tau, S, t, x) = xg(\xi, y)$, $y = \frac{S}{x} \in \mathbb{R}_+$, $\xi = t - \tau \in (0, \tau)$, where g is the solution of a complementarity problem independent of x, easily deduced from (6)–(8). For brevity, we do not write this problem. From this observation, easy calculations show that, as a function of t and x, $P(0, S, t, x)$ satisfies the following forward problem: $P(t = 0) = P_\circ$ and for $t \in (0, T]$ and $x > 0$,

$$\left(\frac{\partial P}{\partial t} - \frac{\sigma^2 x^2}{2} \frac{\partial^2 P}{\partial x^2} + rx \frac{\partial P}{\partial x} + BP \right) \geq 0, \tag{10}$$

$$P(t, x) \geq P_\circ(x), \tag{11}$$

$$\left(\frac{\partial P}{\partial t} - \frac{\sigma^2 x^2}{2} \frac{\partial^2 P}{\partial x^2} + rx \frac{\partial P}{\partial x} + BP \right) (P - P_\circ) = 0, \tag{12}$$

where the integral operator B is defined by

$$(Bu)(x) = - \int_{z \in \mathbb{R}} k(z) \left(x(e^z - 1) \frac{\partial u}{\partial x}(x) + e^z (u(xe^{-z}) - u(x)) \right) dz. \tag{13}$$

The problem (10)–(12) can also be obtained by probabilistic arguments. The new approach for calibrating the Lévy process is to minimize the functional $(\sigma, k) \mapsto \sum_{i \in I} \omega_i |\bar{P}_i - P(t_i, x_i)|^2 + J_R(\sigma, k)$ for σ and k in a suitable set, where the prices $P(t_i, x_i)$ are computed by solving (10)–(12), with $P_\circ(x) = (x - S_\circ)_+$. In contrast with the previous approach, evaluating the functional requires solving one variational inequality only.

Such a forward problem is reminiscent of the forward equation which is often used for the calibration of the local volatility with vanilla European options. This equation is known as Dupire's equation in the finance community, see [Dup97, AP05a]. Note that the arguments used to obtain (10)–(12) are easier than those used for getting Dupire's equation, because the operator in (6)–(8) is invariant by any change of variable $S \mapsto \lambda S$, $\lambda > 0$, which is not the case with local volatility. Note also that finding a forward linear complementarity problem in the variables t and x is not possible in the case of American options with local volatility.

Calibration of σ and k is an inverse problem for finding the coefficients of a variational inequality involving a partial integro-differential operator. The main goal of the paper is to study the last least square optimization problem theoretically, for a special parameterization of k, see (25) below, with σ bounded away from 0, and to give necessary optimality conditions. The results presented here have their discrete counterparts when the variational inequalities are discretized with finite elements of finite differences. Numerical results will be presented in a forthcoming paper.

2 Preliminary Results

2.1 Change of Unknown Function in the Forward Problem

It is helpful to change the unknown function: we set

$$u_\circ(x) = (S - x)_+, \quad u(t, x) = P(t, x) - x + S. \tag{14}$$

The function u satisfies: for $t \in (0, T]$ and $x > 0$,

$$\frac{\partial u}{\partial t} - \frac{\sigma^2 x^2}{2} \frac{\partial^2 u}{\partial x^2} + rx \frac{\partial u}{\partial x} + Bu \geq -rx, \tag{15}$$

$$u(t, x) \geq u_\circ(x), \tag{16}$$

$$\left(\frac{\partial u}{\partial t} - \frac{\sigma^2 x^2}{2} \frac{\partial^2 u}{\partial x^2} + rx \frac{\partial u}{\partial x} + Bu + rx \right) (u - u_\circ) = 0. \tag{17}$$

The initial condition for u is

$$u(t = 0, x) = u_\circ(x), \quad x > 0. \tag{18}$$

For writing the variational inequalities stemming from (15)–(18), we need to introduce suitable weighted Sobolev spaces. In particular, fractional order weighted Sobolev spaces will be useful for studying the non-local part of the operator.

2.2 Functional Setting

Sobolev Spaces on $\mathbb{R}$

For a real number s, let the Sobolev space $H^s(\mathbb{R})$ be defined as follows: the distribution w defined on $\mathbb{R}$ belongs to $H^s(\mathbb{R})$ if and only if its Fourier transform $\widehat{w}$ satisfies

$$\int_{\mathbb{R}} (1 + \xi^2)^s |\widehat{w}(\xi)|^2 d\xi < +\infty.$$

The spaces $H^s(\mathbb{R})$ are Hilbert spaces, with the inner product and norm:

$$(w_1, w_2)_{H^s(\mathbb{R})} = \int_{\mathbb{R}} (1 + \xi^2)^s \widehat{w_1}(\xi) \overline{\widehat{w_2}(\xi)} d\xi, \quad \|w\|_{H^s(\mathbb{R})} = \sqrt{(w, w)_{H^s(\mathbb{R})}}.$$

We refer to [Ada75] for the properties of the spaces $H^s(\mathbb{R})$. If s is a non-negative integer, we define the semi-norm

$$|v|_{H^s(\mathbb{R})} = \left(\sum_{\ell=1}^{s} \left\| \frac{d^\ell v}{dy^\ell} \right\|^2_{L^2(\mathbb{R})} \right)^{\frac{1}{2}}.$$

If $s > 0$ is not an integer, we define $|v|_{H^s(\mathbb{R})}$ by

$$|v|^2_{H^s(\mathbb{R})} = \sum_{\ell=1}^{m} \left\| \frac{d^\ell v}{dy^\ell} \right\|^2_{L^2(\mathbb{R})} + \int_{\mathbb{R}} \int_{\mathbb{R}} \frac{\left(\frac{d^m v}{dy^m}(y) - \frac{d^m v}{dy^m}(z) \right)^2}{|y - z|^{1+2s}},$$

where m is the integer part of s.

Some Weighted Sobolev Spaces on $\mathbb{R}_+$

Let $L^2(\mathbb{R}_+)$ be the Hilbert space of square integrable functions on $\mathbb{R}_+$, endowed with the norm $\|v\|_{L^2(\mathbb{R}_+)} = (\int_{\mathbb{R}_+} v(x)^2 dx)^{\frac{1}{2}}$ and the inner product $(v, w)_{L^2(\mathbb{R}_+)} = \int_{\mathbb{R}_+} v(x)w(x)dx$. Let V^1 be the weighted Sobolev space

$$V^1 = \left\{ v \in L^2(\mathbb{R}_+), \ x \frac{\partial v}{\partial x} \in L^2(\mathbb{R}_+) \right\}, \tag{19}$$

which is a Hilbert space with the norm

$$\|v\|_{V^1} = \left(\|v\|^2_{L^2(\mathbb{R}_+)} + \left\| x \frac{\partial v}{\partial x} \right\|^2_{L^2(\mathbb{R}_+)} \right)^{\frac{1}{2}}. \tag{20}$$

It is proved in [AP05a] that $\mathcal{D}(\mathbb{R}_+)$ is a dense subspace of V^1, and that the following Poincaré inequality is true: for all $v \in V^1$,

$$\|v\|_{L^2(\mathbb{R}_+)} \leq 2 \left\| x \frac{dv}{dx} \right\|_{L^2(\mathbb{R}_+)}. \tag{21}$$

Thus the semi-norm $|\cdot|_{V^1} \colon |v|_{V^1} = \|x \frac{dv}{dx}\|_{L^2(\mathbb{R}_+)}$ is a norm equivalent to $\|\cdot\|_{V^1}$.

For a function v defined on $\mathbb{R}_+$, call $\tilde{v}$ the function defined on $\mathbb{R}$ by

$$\tilde{v}(y) = v(\exp(y)) \exp\left(\frac{y}{2} \right). \tag{22}$$

By using the change of variable $y = \log(x)$, it can be seen that the mapping $v \mapsto \tilde{v}$ is a topological isomorphism from $L^2(\mathbb{R}_+)$ onto $L^2(\mathbb{R})$, and from V^1 onto $H^1(\mathbb{R})$. This leads to defining the space V^s, for $s \in \mathbb{R}$, by:

$$V^s = \{v : \tilde{v} \in H^s(\mathbb{R})\}, \tag{23}$$

which is a Hilbert space with the norm $\|v\|_{V^s} = \|\tilde{v}\|_{H^s(\mathbb{R})}$. Using the interpolation theorem given, e.g., in [Ada75, Theorem 7.17], one can prove that if $0 < s < 1$, then V^s can be obtained by real interpolation between the spaces V^1 and $L^2(\mathbb{R}_+)$ (the parameter for the real interpolation is $\nu = \frac{1}{2} - s$), and that the norm obtained by the interpolation process is equivalent to the one defined above. For $s > 0$, the space V^{-s} is the topological dual of V^s. For $s > 0$, we introduce the semi-norm $|v|_{V^s} = |\tilde{v}|_{H^s(\mathbb{R})}$.

Proposition 1. *Let s be a real number such that $\frac{1}{2} < s \leq 1$. Then for all $u \in V^s$, v is continuous on $(0, +\infty)$ and there exists a constant $C > 0$ such that for all $x \in [1, +\infty)$,*

$$\sqrt{x}|v(x)| \leq C\|v\|_{V^s}. \tag{24}$$

2.3 The Integro-Differential Operator

The Integral Operator

We study the integral operator B defined in (13). Let ψ be a measurable, non-negative and essentially bounded function defined on $\mathbb{R}$, and α be a real number, $0 \leq \alpha < 1$. Consider the kernel

$$k(z) = \frac{\psi(z)}{|z|^{1+2\alpha}}. \tag{25}$$

We assume that $z \mapsto \psi(z) \max\left(e^{2z}, 1\right)$ is a bounded function. If $\alpha = 0$ assume, furthermore, that $\int_{-\infty}^{-1} \frac{\psi(z)}{|z|} dz < +\infty$. Note that, for B defined in (13), Bu is well defined if, for example, $u \in \mathcal{D}(\mathbb{R}_+)$.

Remark 1. To avoid ambiguities in the definition of k, we assume in most of what follows that there exists a positive constant $\underline{\psi}$ such that $\psi(z) \geq \underline{\psi} > 0$ in a fixed neighborhood of $z = 0$. This assumption is a little restrictive, since, for example, a logarithmic singularity of k will be ruled out. Most of the results below hold without the last assumption on ψ.

Proposition 2. *Assume that $z \mapsto \psi(z) \max\left(e^{2z}, 1\right)$ is a bounded function. If $\alpha = 0$ assume, furthermore, that $\int_{-\infty}^{-1} \frac{\psi(z)}{|z|} dz < +\infty$. Then, for each $s \in \mathbb{R}$,*

  (i) if $\alpha > \frac{1}{2}$, then the operator B is continuous from V^s to $V^{s-2\alpha}$,
  (ii) if $\alpha < \frac{1}{2}$, then the operator B is continuous from V^s to V^{s-1},
  (iii) if $\alpha = \frac{1}{2}$, then the operator B is continuous from V^s to $V^{s-1-\varepsilon}$, for any $\varepsilon > 0$.

Remark 2. As a consequence of Proposition 2, if $\frac{1}{2} < \alpha < 1$, then the operator B is continuous from V^α to $V^{-\alpha}$.

Proposition 3. *If the assumptions of Proposition 2 are satisfied and if $\frac{1}{2} < \alpha < 1$, then for any $v, w \in V^\alpha$,*

$$
\langle Bu, v \rangle + \langle Bv, u \rangle = \begin{pmatrix} \displaystyle\int_{\mathbb{R}_+} \int_{\mathbb{R}} k(z) e^z (u(x) - u(xe^{-z}))(v(x) - v(xe^{-z})) dx dz \\ + \left(\displaystyle\int_{\mathbb{R}} k(z)(2e^z - e^{2z} - 1) dz \right) \displaystyle\int_{\mathbb{R}_+} u(x) v(x) dx \end{pmatrix},
\tag{26}
$$

where $\langle \cdot, \cdot \rangle$ stands for the duality pairing between $V^{-\alpha}$ and V^α.

If $0 \le \alpha \le \frac{1}{2}$, then (26) is true for $u, v \in V^s$, $s > \frac{1}{2}$, defining $\langle \cdot, \cdot \rangle$ as the duality pairing between V^{-s} and V^s.

Proposition 4 (Gårding inequality). *If the assumptions of Proposition 2 are satisfied and if there exists a constant $\underline{\psi}$ such that $\psi \ge \underline{\psi} > 0$ almost everywhere in a neighborhood of 0, then*

(i) *if $\frac{1}{2} < \alpha < 1$, there exists a positive constant $\underline{C}$ and a non-negative constant λ such that, for all $v \in V^\alpha$,*

$$
\langle Bv, v \rangle \ge \underline{C} |v|_{V^\alpha}^2 - \lambda \|v\|_{L^2(\mathbb{R}_+)}^2;
\tag{27}
$$

(ii) *if $\alpha \le \frac{1}{2}$, then (27) holds for any $v \in V^s$, $s > \frac{1}{2}$ ($\langle \cdot, \cdot \rangle$ standing for the duality pairing between V^{-s} and V^s).*

Consider the two situations:

1. $\frac{1}{2} < \alpha < 1$, ψ satisfies the assumptions of Proposition 2, and $u \in V^\alpha$, then it can be shown (using the interpolation theorem in [Ada75, Theorem 7.17]) that the functions u_+ and u_- belong to V^α;
2. $\alpha \le \frac{1}{2}$, ψ satisfies the assumptions of Proposition 2, and $u \in V^1$.

In both cases, $\int_{\mathbb{R}_+} \int_{\mathbb{R}} k(z) e^z u_-(xe^{-z}) u_+(x) dx dz$ is well defined because

$$
\int_{\mathbb{R}_+} \int_{\mathbb{R}} k(z) e^z u_-(xe^{-z}) u_+(x) dx dz
$$

$$
= \int_{\mathbb{R}_+} \int_{\mathbb{R}} k(z) e^z (u_-(xe^{-z}) - u_-(x)) u_+(x) dx dz
$$

$$
\le \left(\int_{\mathbb{R}_+} \int_{\mathbb{R}} k(z) e^z (u_-(xe^{-z}) - u_-(x))^2 dz dx \right)^{\frac{1}{2}} \|u_+\|_{L^2(\mathbb{R}_+)},
$$

and is non-negative. Therefore,

$$
\langle Bu, u_+ \rangle = \langle Bu_+, u_+ \rangle - \int_{\mathbb{R}_+} \int_{\mathbb{R}} k(z) e^z (u(xe^{-z}) - u_+(xe^{-z})) u_+(x) dx dz
$$

$$
= \langle Bu_+, u_+ \rangle + \int_{\mathbb{R}_+} \int_{\mathbb{R}} k(z) e^z u_-(xe^{-z}) u_+(x) dx dz \ge \langle Bu_+, u_+ \rangle.
$$

We have proved

Proposition 5. *Under the assumptions of Proposition 4, there exist a positive constant $\underline{C}$ and a constant $\lambda \geq 0$ such that, for all $u \in V^{\alpha}$ if $\alpha > 1/2$ or for all $u \in V^1$ if $\alpha \leq 1/2$,*

$$\langle Bu, u_+\rangle \geq \underline{C}|u_+|_{V^{\alpha}}^2 - \lambda\|u_+\|_{L^2(\mathbb{R}_+)}^2. \tag{28}$$

A weak maximum principle for parabolic problems stems from Proposition 5.

The Integro-Differential Operator when the Volatility σ is Positive

When $\sigma > 0$, the space V^1 plays a special role. Thus, we use the shorter notation $V = V^1$.

With B defined in (13), we introduce the integro-differential operator A:

$$Av = -\frac{\sigma^2 x^2}{2}\frac{\partial^2 v}{\partial x^2} + rx\frac{\partial v}{\partial x} + Bv. \tag{29}$$

If $\sigma > 0$, and if (α, ψ) satisfy the assumptions of Proposition 4, then

- A is a continuous operator from V to V^{-1},
- we have the Gårding inequality: there exist $\underline{c} > 0$ and $\lambda \geq 0$ such that

$$\langle Av, v\rangle \geq \underline{c}|v|_V^2 - \lambda\|v\|_{L^2(\mathbb{R}_+)}^2, \quad \forall v \in V, \tag{30}$$

- for any $v \in V$,

$$\langle Av, v_+\rangle \geq \underline{c}|v_+|_V^2 - \lambda\|v_+\|_{L^2(\mathbb{R}_+)}^2, \tag{31}$$

- the operator $A + \lambda I$ is one to one and continuous from V^2 onto $L^2(\mathbb{R}_+)$, with a continuous inverse.

Remark 3. Note that the assumption that $\psi > 0$ near $z = 0$ is not necessary for A to have the above properties: indeed, since $\sigma > 0$, Gårding's inequality holds even if $\psi = 0$ near 0. The main advantage of this assumption is rather that it permits a clear identification of the kernel's singularity at $z = 0$.

3 The Variational Inequality when the Volatility σ is Positive

We are ready to write the variational inequalities corresponding to the linear complementarity problem (15)–(18).

We introduce the closed subspace of V:

$$K = \{v \in V, \; v(x) \geq u_\circ(x) \text{ in } \mathbb{R}_+\}. \tag{32}$$

The variational problem consists of finding $u \in L^2(0,T;V) \cap \mathcal{C}^0([0,T];L^2(\mathbb{R}_+))$, with $\frac{\partial u}{\partial t} \in L^2(0,T;V')$, such that

1. there exists a constant $X_T > S$ such that $u(t,x) = 0$ for any $t \in [0,T]$, $x \geq X_T$;
2. $u(t) \in K$ for almost every $t \in (0,T)$;
3. for any $v \in K$ with bounded support, for almost every $t \in (0,T)$,

$$\left\langle \frac{\partial u}{\partial t} + Au + rx, v - u \right\rangle \geq 0, \qquad (33)$$

here $\langle \cdot, \cdot \rangle$ stands for the duality pairing between V' (the dual of V) and V;
4. $u(t=0) = u_o$.

Hereafter, this problem will be referred to as (VIP).

3.1 Existence and Uniqueness

Theorem 1. *If $\sigma > 0$ and under the assumptions of Proposition 4, there exists a unique u solution of problem (VIP) defined above. Furthermore, $u \in \mathcal{C}^0([0,T];K) \cap L^2(0,T;V^2)$ and $\frac{\partial u}{\partial t} \in L^2((0,T) \times \mathbb{R}_+)$.*

There exists a non-decreasing and lower semi-continuous function $\gamma : (0,T] \to (S,X_T)$, such that for all $t \in (0,T)$, $\{x > 0 \text{ s.t. } u(t,x) = u_o(x)\} = [\gamma(t), +\infty)$.

Calling

$$\mu = \frac{\partial u}{\partial t} + Au + rx, \qquad (34)$$

we have a.e. $0 \leq \mu \leq rx1_{\{u=0\}} = rx1_{\{x \geq \gamma(t)\}}$. The function μ is non-decreasing with respect to x (i.e. the distribution $\frac{\partial \mu}{\partial x}$ is negative) and non-increasing with respect to t, (i.e. the distribution $\frac{\partial \mu}{\partial t}$ is positive). For any $X > X_T$, the total variation of μ in $(0,T) \times (0,X)$ is bounded by $rX(T+X)$.

Almost everywhere in the coincidence set where $u(t,x) = 0$, it holds $\mu(t,x) > 0$.

Proof. The proof is too long to be given here. It is written in [Ach06]. Here, we limit ourselves to list the main steps. The fact that Problem (15)–(18) is posed in an unbounded domain induces technical difficulties for variational methods. This leads us to first consider an approximate problem posed in a bounded domain. Therefore, the program is to

1. approximate (15)–(18) by a similar problem posed in $[0,T] \times [0,X]$, with a homogeneous Dirichlet condition on the boundary $x = X$, for some given positive parameter $X > S$, and write the related variational problem, which will be called (VIP$_X$) below;

2. solve first a penalized version of (VIP_X). For a function $v \in L^2((0, X))$ we call $\mathcal{E}_X(v)$ the function in $L^2(\mathbb{R}_+)$ obtained by extending v by 0 outside $(0, X)$. We introduce the Sobolev space

$$V_X = \{v \in L^2(0, X), \mathcal{E}_X(v) \in V\}, \tag{35}$$

with $\|v\|_{V_X} = \|\mathcal{E}_X(v)\|_V$. We define the operators A_X and $B_X \colon V_X \to V_X'$,

$$\langle A_X v, w \rangle = \langle A\mathcal{E}_X(v), \mathcal{E}_X(w) \rangle \quad \text{and} \quad \langle B_X v, w \rangle = \langle B\mathcal{E}_X(v), \mathcal{E}_X(w) \rangle. \tag{36}$$

The penalized problem is to find $u_{X,\varepsilon}$ such that

$$\frac{\partial u_{X,\varepsilon}}{\partial t} + A_X u_{X,\varepsilon} + rx(1 - 1_{\{x>S\}}\mathcal{V}_\varepsilon(u_{X,\varepsilon})) = 0, \quad t \in (0, T], \ 0 < x < X,$$

$$u_{X,\varepsilon}(t = 0, x) = u_o(x), \quad 0 < x < X,$$

$$u_{X,\varepsilon}(t, X) = 0, \quad t \in (0, T],$$

$$\tag{37}$$

where $\mathcal{V}_\varepsilon(u) = \mathcal{V}(\frac{u}{\varepsilon})$ and $\mathcal{V}$ is a smooth non-increasing convex function such that

$$\mathcal{V}(0) = 1, \quad \mathcal{V}(u) = 0 \quad \text{for } u \geq 1, \quad 0 \geq \mathcal{V}'(u) \geq -2 \quad \text{for } 0 \leq u \leq 1. \tag{38}$$

By using the theory of Lions [Lio69] for parabolic problems with semilinear monotone operators, one can prove that (37) has a unique solution and pass to the limit as the penalty parameter tends to zero; one obtains the existence and uniqueness for (VIP_X).

3. prove that the free boundary of (VIP_X) stays in a bounded domain as X tends to infinity: this will show that for X large enough a solution of (VIP_X) is actually a solution of (VIP).

Remark 4. By using the theory presented in [BL84], it is possible to study the variational inequality in Sobolev spaces with decaying weights as $x \to 0$ and $x \to +\infty$ (actually the variable $\log(x)$ was used instead of x in [BL84]). In Theorem 1, we have avoided these weights.

Remark 5. The last statement of Theorem 1 tells us that there is almost everywhere strict complementarity: the reaction term μ is positive at almost every point where $u = 0$.

3.2 Bounds and Sensitivity

In what follows, we aim at obtaining estimates for the solution of (VIP) independent of the parameters (σ, α, ψ), when these parameters vary in a suitably defined set. Let us introduce $\mathcal{B} = \{f : z \mapsto f(z) \max(1, |z|, e^{2z}) \in L^\infty(\mathbb{R})\}$ endowed with the norm $\|f\|_{\mathcal{B}} = \|f(\cdot) \max(1, |\cdot|, e^{2\cdot})\|_{L^\infty(\mathbb{R})}$. Let us choose some constants $\underline{\sigma}, \bar{\sigma}, \underline{\alpha}, \underline{\psi}, \bar{\psi}$ and $\bar{z}$ such that $0 < \underline{\sigma} \leq \bar{\sigma}$, $0 < \underline{\alpha} < \frac{1}{2}$, $\bar{\psi} \geq \underline{\psi} > 0$ and $\bar{z} > 0$. Let us define the subset $\mathcal{F}$ of $\mathbb{R}_+^2 \times \mathcal{B}$ by

$$\mathcal{F} = [\underline{\sigma}, \bar{\sigma}] \times [0, 1 - \underline{\alpha}] \times \left\{ \psi \in \mathcal{B} : \left| \begin{array}{l} \|\psi\|_{\mathcal{B}} \leq \bar{\psi}; \ \psi \geq 0, \\ \psi \geq \underline{\psi} \ \text{a.e. in } [-\bar{z}, \bar{z}] \end{array} \right. \right\}. \tag{39}$$

We can make the three observations:

1. The norm of A as an operator from V to V' is bounded independently of (σ, α, ψ) in $\mathcal{F}$.
2. The constants in (30) and (31) can be taken independent of (σ, α, ψ) in $\mathcal{F}$.
3. With λ in (30) independent of (σ, α, ψ) in $\mathcal{F}$, the operator $A + \lambda I$ is one to one and continuous from V^2 onto $L^2(\mathbb{R}_+)$ and $(A + \lambda I)^{-1} : L^2(\mathbb{R}_+) \mapsto V^2$ is bounded with constants independent of (σ, α, ψ) in $\mathcal{F}$.

These last points are used for proving the following:

Proposition 6 (Bounds). *The function γ is bounded in $[0, T]$ by some constant $\bar{X}$ independent of (σ, α, ψ) in $\mathcal{F}$. The quantities $\|u\|_{L^\infty(0,T;V)}$, $\|u\|_{L^2(0,T;V^2)}$ and $\|\frac{\partial}{\partial t} u\|_{L^2((0,T)\times\mathbb{R}_+)}$ are bounded independently of (σ, α, ψ) in $\mathcal{F}$.*

Proposition 7 (Sensitivity). *There exists a constant C, such that for all (σ, α, ψ), $(\tilde{\sigma}, \tilde{\alpha}, \tilde{\psi})$ in $\mathcal{F}$,*

$$\|u - \tilde{u}\|_{L^2(0,T;V)} + \|u - \tilde{u}\|_{L^\infty(0,T;L^2(\mathbb{R}_+))} \leq C\left(|\sigma - \tilde{\sigma}| + |\alpha - \tilde{\alpha}| + \|\psi - \tilde{\psi}\|_{\mathcal{B}}\right),$$

$$\int_0^T \int_{\mathbb{R}} \left(\mu(\tilde{u} - u_\circ) + \tilde{\mu}(u - u_\circ)\right) \leq C\left(|\sigma - \tilde{\sigma}| + |\alpha - \tilde{\alpha}| + \|\psi - \tilde{\psi}\|_{\mathcal{B}}\right)^2,$$

calling $u = u(\sigma, \alpha, \psi)$ and $\mu = \mu(\sigma, \alpha, \psi)$ the solution of (VIP) and the parameters (σ, α, ψ) and the corresponding reaction term (see (34)). Furthermore, let $(\sigma_n, \alpha_n, \psi_n)_{n\in\mathbb{N}}$ be a sequence of coefficients in $\mathcal{F}$ such that $\lim_{n\to\infty}(|\sigma - \sigma_n| + |\alpha - \alpha_n| + \|\psi - \psi_n\|_{\mathcal{B}}) = 0$. With the notations $u_n = u(\sigma_n, \alpha_n, \psi_n)$ and $\mu_n = \mu(\sigma_n, \alpha_n, \psi_n)$,

$$\lim_{n\to+\infty} \|u_n - u\|_{L^\infty((0,T)\times\mathbb{R}_+)} = 0, \quad \lim_{n\to+\infty} \|\mu_n - \mu\|_{L^p((0,T)\times\mathbb{R}_+)} = 0,$$

for all p, $1 < p < +\infty$, and

$$\lim_{n\to+\infty} \left(\|u_n - u\|_{L^\infty(0,T;V)} + \|u_n - u\|_{L^2(0,T;V^2)} + \right.$$

$$\left. + \left\|\frac{\partial u_n}{\partial t} - \frac{\partial u}{\partial t}\right\|_{L^2((0,T)\times\mathbb{R}_+)} \right) = 0.$$

4 Calibration by Least Squares

4.1 Orientation

For calibrating the Lévy process, one observes the spot price S and the prices $(\bar{p}_i)_{i\in I}$ of a family of American put options with maturities/strikes given by

(T_i, x_i); we call $\bar{u}_i = \bar{p}_i - x_i + S$, $i \in I$. The parameters of the Lévy process, i.e. the volatility σ, the exponent α and the function ψ will be found as solutions of a least square problem, where the functional to be minimized is the sum of a suitable Tychonoff regularization functional $J_R(\sigma, \alpha, \psi)$ and of

$$J(u) = \sum_{i \in I} \omega_i (u(T_i, x_i) - \bar{u}_i)^2,$$

where ω_i are positive weights, and $u = u(\sigma, \alpha, \psi)$ is a solution of (VIP), with $T = \max_{i \in I} T_i$.

We aim at finding some necessary optimality conditions satisfied by the solutions of the least square problem. The main difficulty comes from the fact that the derivability of the functional $J(u)$ with respect to the parameter (σ, α, ψ) is not guaranteed. To obtain some necessary optimality conditions, we shall consider first a least square problem where u is the solution of the penalized problem (37) rather than (VIP), obtain necessary optimality conditions for this new problem, then have the penalty parameter ε tend to 0 and pass to the limit in the optimality conditions. Such a program has already been applied in [Ach05] for calibrating the local volatility with American options, see also [AP05b, AP05a] for a related numerical method and results. The idea originally comes from Hintermüller [Hin01] and Ito and Künisch [IK00], who applied a similar program for elliptic variational inequalities. At this point, we should also mention Mignot and Puel [MP84] who applied an elegant method for finding optimality conditions for a special control problem for a parabolic variational inequality.

4.2 Preliminary Technical Results

With the aim of finding optimality conditions for the least square problem (not completely defined yet), we first state some results concerning the adjoint of B.

Under the assumptions of Proposition 2, it can be checked that the operator B^T defined by

$$B^T u(x) = \int_{z \in \mathbb{R}} k(z) \left(x(e^z - 1) \frac{\partial u}{\partial x}(x) - e^{2z} u(xe^z) + (2e^z - 1)u(x) \right) dz \quad (40)$$

is a continuous operator

$$\begin{cases} \text{from } V^s \text{ to } V^{s-2\alpha}, & \text{if } \alpha > \frac{1}{2}, \\ \text{from } V^s \text{ to } V^{s-1}, & \text{if } \alpha < \frac{1}{2}, \\ \text{from } V^s \text{ to } V^{s-1-\varepsilon}, \text{ for any } \varepsilon > 0, & \text{if } \alpha = \frac{1}{2}. \end{cases}$$

If $\alpha > \frac{1}{2}$, then for all $u, v \in V^\alpha$, $\langle B^T u, v \rangle = \langle Bv, u \rangle$. This identity holds for all $u, v \in V^s$ with $s > \frac{1}{2}$ if $\alpha \leq \frac{1}{2}$.

Lemma 1. *Under the assumptions of Proposition 2, and if*

(i) *either $\alpha < \frac{1}{2}$,*

(ii) *or ψ is continuous near 0 and there exists a bounded function $\omega : \mathbb{R} \to \mathbb{R}$ and two positive numbers ζ and C such that $\psi(z)e^{\frac{3}{2}z} - \psi(0)e^{-\frac{3}{2}z} = z\omega(z)$, with $|\omega(z)| \leq C|z|e^{-\zeta|z|}$, for all $z \in \mathbb{R}$,*

then for any $s \in \mathbb{R}$, the operator $B - B^T$ is continuous from V^s to V^{s-1}.

4.3 The Least Square Problem and Its Penalized Version

In order to properly define the least square problem, we have to define the set where (σ, α, ψ) may vary and the regularization functional.

Let us introduce an Hilbert space H_ψ endowed with the norm $\| \cdot \|_{H_\psi}$, relatively compact in $\mathcal{B}$. Let J_ψ be a convex, coercive and $\mathcal{C}^1$ function defined on H_ψ. It is well known that J_ψ is also weakly lower semicontinuous in H_ψ.

Consider $\mathcal{H}_\psi$ a closed and convex subset of H_ψ. We assume that $\mathcal{H}_\psi$ is contained in $\{\psi : \|\psi\|_\mathcal{B} \leq \bar{\psi}; \ \psi \geq 0\}$ and that

1. the functions $\psi \in \mathcal{H}_\psi$ are continuous near 0,
2. there exists two positive constants $\underline{\psi}$ and $\bar{z}$ such that $\psi(z) \geq \underline{\psi}$ for all z such that $|z| \leq \bar{z}$,
3. there exist two constants $\zeta > 0$ and $C \geq 0$ such that for all $\psi \in \mathcal{H}_\psi$, $\psi(z)e^{\frac{3}{2}z} - \psi(0)e^{-\frac{3}{2}z} = z\omega(z)$, with $|\omega(z)| \leq C|z|e^{-\zeta|z|}$, for all $z \in \mathbb{R}$. This assumption will allow us to use the results stated in Lemma 1.

Finally, consider the set $\mathcal{H} = [\underline{\sigma}, \bar{\sigma}] \times [0, 1 - \underline{\alpha}] \times \mathcal{H}_\psi$ and define

$$J_R(\sigma, \alpha, \psi) = |\sigma - \sigma_\circ|^2 + |\alpha - \alpha_\circ|^2 + J_\psi(\psi),$$

where $\sigma_\circ$ and $\alpha_\circ$ are suitable prior parameters.

Consider the least square problem:

Minimize $J(u) + J_R(\sigma, \alpha, \psi) \ \big| \ (\sigma, \alpha, \psi) \in \mathcal{H}, \ u = u(\sigma, \alpha, \psi)$ satisfies (VIP).

(41)

We fix $\bar{X}$ (independent of $(\sigma, \alpha, \psi) \in \mathcal{H}$) as in Proposition 6, and assume that $x_i < \bar{X}$, $i \in I$. Taking $X \geq \bar{X}$, it is also possible to consider the least square inverse problem corresponding to the penalized problem

Minimize $J(u_\varepsilon) + J_R(\sigma, \alpha, \psi) \ \big| \ (\sigma, \alpha, \psi) \in \mathcal{H}, \ u_\varepsilon$ satisfies (37).   (42)

Propositions 6 and 7 are useful for proving the following:

Proposition 8 (Approximation of the least square problem). *Let $(\varepsilon_n)_n$ be a sequence of penalty parameters such that $\varepsilon_n \to 0$ as $n \to \infty$, and let $(\sigma^*_{\varepsilon_n}, \alpha^*_{\varepsilon_n}, \psi^*_{\varepsilon_n}), u^*_{\varepsilon_n}$ be a solution of the problem (42), with X fixed as above. Consider a subsequence such that $(\sigma^*_{\varepsilon_n}, \alpha^*_{\varepsilon_n}, \psi^*_{\varepsilon_n})$ converges to $(\sigma^*, \alpha^*, \psi^*)$ in $\mathcal{F}$, $\psi^*_{\varepsilon_n}$ weakly converges to ψ^* in H_ψ and $u^*_{\varepsilon_n} \to u^*$ weakly in $L^2(0, T; V_X)$,*

where V_X is defined in (35). Then $(\sigma^*, \alpha^*, \psi^*), u^*$ is a solution of (41), where we agree to use the notation u^* for the function $\mathcal{E}_X(u^*)$. We have that

(i) $u^*_{\varepsilon_n}$ converges to u^* uniformly in $[0, T] \times [0, X]$, and in $L^2(0, T; V_X)$;

(ii) $1_{\{x > S\}} r x \mathcal{V}_{\varepsilon_n}(u^*_{\varepsilon_n})$ converges to μ^* strongly in $L^2((0, T) \times (0, X))$;

(iii) for all smooth function χ with compact support contained in $[0, X)$, $\chi u^*_{\varepsilon_n}$ converges to χu^* strongly in $L^2(0, T; V^2)$ and in $L^\infty(0, T; V)$.

4.4 The Optimality Conditions

We fix X as above. Let a subsequence $(\sigma^*_{\varepsilon_n}, \alpha^*_{\varepsilon_n}, \psi^*_{\varepsilon_n}, u^*_{\varepsilon_n})$ of solutions of (42) converge to $(\sigma^*, \alpha^*, \psi^*, u^*)$ as in Proposition 8, then $(\sigma^*, \alpha^*, \psi^*, u^*)$ is a solution of (41).

The optimality conditions will involve an adjoint problem. Since the cost functional involves point-wise values of u, the adjoint problem will have a singular data. In that context, the notion of very weak solution of boundary value problems will be relevant: for that, we introduce the spaces $\widetilde{Z}$ and Z,

$$\widetilde{Z} = \left\{ v \in L^2(0, T; V_X); \ \frac{\partial v}{\partial t} + A_X v \in L^2((0, T) \times (0, X)) \right\}, \tag{43}$$

$$Z = \{ v \in \widetilde{Z}; \ v(t = 0) = 0 \},$$

where A_X is the operator given by (36), (29) and (13), with the parameters $(\sigma^*, \alpha^*, \psi^*)$. These spaces endowed with the graph norm are Banach spaces.

We also need to introduce some functionals before stating the optimality conditions. We assume that $u^*(T_i, x_i) > u_\circ(x_i)$, for all $i \in I$. It is clear from the continuity of u^* and from the uniform convergence of $u^*_{\varepsilon_n}$ that there exists a positive real number a and an integer N such that for $n > N$, $u^*_{\varepsilon_n}(t, x) > u_\circ(x) + \varepsilon_n$ for all (t, x) such that $|t - T_i| < a$ and $|x - x_i| < a$ for some $i \in I$. We may fix a smooth function ϕ taking the value 1 for all x such that $|x - x_i| \geq \frac{a}{2}$, $|T_i - t| \geq \frac{a}{2}$ for all $i \in I$, and vanishing in neighborhoods of (T_i, x_i), $i \in I$.

For a function p such that $p \in L^2((0, T) \times \mathbb{R}_+)$ and $\phi p \in L^2(0, T; V_X)$ we introduce the quantities

$$\mathcal{G}^{(\sigma)}(u^*, p) = \int_0^T \left\langle x^2 \frac{\partial^2 u^*}{\partial x^2}, \phi p \right\rangle + \int_0^T \int_0^X \left((1 - \phi) x^2 \frac{\partial^2 u^*}{\partial x^2} \right) p, \tag{44}$$

$$\mathcal{G}^{(\alpha)}(u^*, p) = \int_0^T \left\langle B_X^{(\alpha)} u^*, \phi p \right\rangle + \int_0^T \int_0^X \left((1 - \phi) B_X^{(\alpha)} u^* \right) p, \tag{45}$$

$$\left\langle \mathcal{G}^{(\psi)}(u^*, p), \kappa \right\rangle = \int_0^T \left\langle B_X^{(\psi, \kappa)} u^*, \phi p \right\rangle + \int_0^T \int_0^X \left((1 - \phi) B_X^{(\psi, \kappa)} u^* \right) p, \tag{46}$$

where $\kappa \in H_\psi$, $\langle \cdot, \cdot \rangle$ denotes the duality pairing between $(V_X)'$ and V_X, and where

$$B_X^{(\alpha)}v(x) = -\int_{\mathbb{R}} k^*(z) \log(|z|) \left(x(e^z - 1)\frac{\partial v}{\partial x}(x) \right.$$

$$\left. + e^z (1_{\{z > -\log(\frac{X}{x})\}}v(xe^{-z}) - v(x)) \right),$$

$$B_X^{(\psi,\kappa)}v(x) = \int_{\mathbb{R}} \frac{\kappa(z)}{|z|^{1+2\alpha^*}} \left(x(e^z - 1)\frac{\partial v}{\partial x}(x) \right.$$

$$\left. + e^z (1_{\{z > -\log(\frac{X}{x})\}}v(xe^{-z}) - v(x)) \right) dz.$$

One can check that $\mathcal{G}^{(\sigma)}(u^*, p)$, $\mathcal{G}^{(\alpha)}(u^*, p)$ and $\langle \mathcal{G}^{(\psi)}(u^*, p), \kappa \rangle$ are well defined and do not depend of the particular choice of ϕ.

We are now ready to state some necessary optimality for the least square problem (42):

Theorem 2. *Let a subsequence $(\sigma_{\varepsilon_n}^*, \alpha_{\varepsilon_n}^*, \psi_{\varepsilon_n}^*, u_{\varepsilon_n}^*)$ of solutions of (42) converge to $(\sigma^*, \alpha^*, \psi^*, u^*)$ as in Proposition 8 (we know that $(\sigma^*, \alpha^*, \psi^*, u^*)$ is a solution of (41)). We assume that $u^*(T_i, x_i) > u_\circ(x_i)$, for all $i \in I$.*

There exists a function $p^* \in L^2((0,T) \times (0,X))$ and a Radon measure ξ^* such that for all $v \in Z$ (Z is defined by (43))

$$\int_0^T \int_0^X \left(\frac{\partial v}{\partial t} + A_X v \right) p^* + \langle \xi^*, v \rangle = 2 \sum_{i \in I} \omega_i(u^*(T_i, x_i) - \bar{u}_i)v((T_i, x_i)), \quad (47)$$

and

$$\mu^*|p^*| = 0, \tag{48}$$

$$|u^*|\xi^* = 0. \tag{49}$$

Furthermore, with ϕ defined above, $\phi p^* \in L^2(0, T, V_X)$, and for all $(\sigma, \alpha, \psi) \in \mathcal{H}$,

$$(\sigma - \sigma^*) \left(2(\sigma^* - \sigma_\circ) + \sigma^* \mathcal{G}^{(\sigma)}(u^*, p^*) \right) \geq 0, \tag{50}$$

$$(\alpha - \alpha^*) \left(\alpha^* - \alpha_\circ + \mathcal{G}^{(\alpha)}(u^*, p^*) \right) \geq 0, \tag{51}$$

$$\langle DJ_\psi(\psi^*), \psi - \psi^* \rangle + \left\langle \mathcal{G}^{(\psi)}(u^*, p^*), \psi - \psi^* \right\rangle \geq 0. \tag{52}$$

with $\mathcal{G}^{(\sigma)}$, $\mathcal{G}^{(\alpha)}$ and $\mathcal{G}^{(\psi)}$ defined respectively by (44), (45) and (46).

Proof. The proof consists of first finding the optimality conditions for (42), then passing to the limit as the penalty parameter tends to zero. It is written in [Ach06]. Optimality conditions for (42) can be obtained in a now classical way (see, e.g., the pioneering book of O. Pironneau [Pir84], he was among the first to understand the potentiality of optimal control techniques in relation with partial differential equations and optimum design).

Note that p^* satisfies

$$\frac{\partial p^*}{\partial t} - A_X^T p^* = -2 \sum_{i \in I} \omega_i (u^*(T_i, x_i) - \bar{u}_i) \delta_{t=T_i} \otimes \delta_{x=x_i} \qquad (53)$$

in the sense of distributions in the open set $\{x, t : u^*(t, x) > u_\circ(x)\}$ and that (48) implies that p^* vanishes in the coincidence set.

5 The Variational Inequality when $\sigma = 0$

We focus on the case when $\sigma = 0$ and when $(\alpha, \psi) \in \mathcal{F}_2$ with

$$\mathcal{F}_2 = \left[\frac{1}{2} + \underline{\alpha}, 1 - \underline{\alpha} \right] \times \left\{ \psi \in \mathcal{B} : \begin{array}{l} \|\psi\|_\mathcal{B} \leq \bar{\psi}; \ \psi \geq 0, \\ \psi \geq \underline{\psi} \text{ a.e. in } [-\bar{z}, \bar{z}] \end{array} \right\}. \qquad (54)$$

for three constants $\underline{\alpha}, \underline{\psi} \ \bar{\psi}, \ 0 < \underline{\alpha} < \frac{1}{2}$ and $\bar{\psi} > \underline{\psi} > 0$.

Remark 6. In the case when $\sigma = 0$ and $\alpha < 1/2$, A is a non-local hyperbolic operator, and the present theory does not apply.

We may prove that

- A is a continuous operator from V^α to $V^{-\alpha}$;
- we have the Gårding inequality: there exist $\underline{c} > 0$ and $\lambda \geq 0$ such that

$$\langle Av, v \rangle \geq \underline{c} |v|_{V^\alpha}^2 - \lambda \|v\|_{L^2(\mathbb{R}_+)}^2, \quad \forall v \in V^\alpha \qquad (55)$$

and

$$\langle Av, v_+ \rangle \geq \underline{c} |v_+|_{V^\alpha}^2 - \lambda \|v_+\|_{L^2(\mathbb{R}_+)}^2, \quad \forall v \in V^\alpha; \qquad (56)$$

- the operator $A + \lambda I$ is one to one and continuous from $V^{2\alpha}$ onto $L^2(\mathbb{R}_+)$.

The goal is to obtain the existence of a weak solution to (15), (17), (18) by a singular perturbation argument: we fix $(\alpha, \psi) \in \mathcal{F}_2$ and for $\eta > 0$, we call u_η the solution to (15), (17), (18) corresponding to $\sigma = \eta$, given by Theorem 1. It can be proven that $\|u_\eta\|_{L^\infty(0,T;V^\alpha)}$ and $\|u_\eta\|_{L^2(0,T;V^{2\alpha})}$ are bounded independently of η, and that the free boundary associated to u_η stays in $[0, T] \times [0, \check{X}]$, where $\check{X}$ does not depend on η. By the results contained in [Lio73, in particular, Théorème 4.1, p. 286], one may pass to the limit as η tends to zero, and prove the following result:

Theorem 3. *We choose $\sigma = 0$ and $(\alpha, \psi) \in \mathcal{F}_2$ and we define*

$$K = \{v \in V^\alpha, \ v(x) \geq u_\circ(x) \text{ in } \mathbb{R}_+\}.$$

There exists a unique weak solution of (15), (17) and (18) in $(0, T) \times \mathbb{R}_+$, i.e. a function u which belongs to $\mathcal{C}^0([0, T]; K)$ and to $L^2(0, T; V^{2\alpha})$, and with

$\frac{\partial u}{\partial t} \in L^2((0,T) \times \mathbb{R}_+)$, *such that $u(t = 0) = u_\circ$ and for all $v \in K$ with bounded support in x,*

$$\left\langle \frac{\partial u}{\partial t} + Au + rx, v - u \right\rangle \geq 0, \quad \text{for a.a. } t > 0. \tag{57}$$

There exists $\check{X} > 0$ such that

$$u(t,x) = 0, \quad \forall t \in [0,T], \ x \geq \check{X}, \tag{58}$$

The function u is non-increasing with respect to x and non-decreasing with respect to t and there exists a non-decreasing continuous function $\gamma : (0,T] \to (S, \check{X})$, such that for all $t \in (0,T)$, $\{x > 0 \ s.t. \ u(t,x) = u_\circ(x)\} = [\gamma(t), +\infty)$.

References

[Ach05]   Y. Achdou. An inverse problem for a parabolic variational inequality arising in volatility calibration with American options. *SIAM J. Control Optim.*, 43(5):1583–1615 (electronic), 2005.

[Ach06]   Y. Achdou. An inverse problem for a parabolic variational inequality with an integro-differential operator arising in the calibration of Lévy processes with American options. Submitted, 2006.

[Ada75]   R. A. Adams. *Sobolev spaces*, volume 65 of *Pure and Applied Mathematics*. Academic Press [A subsidiary of Harcourt Brace Jovanovich Publishers], New York, 1975.

[AP05a]   Y. Achdou and O. Pironneau. *Computational methods for option pricing*, volume 30 of *Frontiers in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2005.

[AP05b]   Y. Achdou and O. Pironneau. Numerical procedure for calibration of volatility with American options. *Appl. Math. Finance*, 12(3):201–241, 2005.

[BL84]   A. Bensoussan and J.-L. Lions. *Impulse control and quasivariational inequalities*. μ. Gauthier-Villars, Montrouge, 1984. Translated from the French by J. M. Cole.

[BS73]   F. Black and M. S. Scholes. The pricing of options and corporate liabilities,. *Journal of Political Economy,*, 81:637–654, 1973.

[CT04]   R. Cont and P. Tankov. *Financial modelling with jump processes*. Chapman & Hall/CRC Financial Mathematics Series. Chapman & Hall/CRC, Boca Raton, FL, 2004.

[CV03]   R. Cont and E. Voltchkova. Finite difference methods for option pricing in jump-diffusion and exponential Lévy models. Rapport Interne 513, CMAP, Ecole Polytechnique, 2003.

[CV04]   R. Cont and E. Voltchkova. Integro-differential equations for option prices in exponential Lévy models. Rapport Interne 547, CMAP, Ecole Polytechnique, 2004.

[Dup97]   B. Dupire. Pricing and hedging with smiles. In *Mathematics of derivative securities (Cambridge, 1995)*, pages 103–111. Cambridge Univ. Press, Cambridge, 1997.

[Hin01]   M. Hintermüller. Inverse coefficient problems for variational inequalities: optimality conditions and numerical realization. *M2AN Math. Model. Numer. Anal.*, 35(1):129–152, 2001.

[IK00]   K. Ito and K. Kunisch. Optimal control of elliptic variational inequalities. *Appl. Math. Optim.*, 41(3):343–364, 2000.

[JLL90]   P. Jaillet, D. Lamberton, and B. Lapeyre. Variational inequalities and the pricing of American options. *Acta Appl. Math.*, 21(3):263–289, 1990.

[Lio69]   J.-L. Lions. *Quelques méthodes de résolution des problèmes aux limites non linéaires*. Dunod, 1969.

[Lio73]   J.-L. Lions. *Perturbations singulières dans les problèmes aux limites et en contrôle optimal*, volume 323 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 1973.

[LL97]   D. Lamberton and B. Lapeyre. *Introduction au calcul stochastique appliqué à la finance*. Ellipses, 1997.

[Mer73]   R. C. Merton. Theory of rational option pricing. *Bell J. Econom. and Management Sci.*, 4:141–183, 1973.

[MNS03]   A.-M. Matache, P.-A. Nitsche, and C. Schwab. Wavelet Galerkin pricing of American options on Lévy driven assets. 2003. Research Report SAM 2003-06.

[MP84]   F. Mignot and J.-P. Puel. Contrôle optimal d'un système gouverné par une inéquation variationnelle parabolique. *C. R. Acad. Sci. Paris Sér. I Math.*, 298(12):277–280, 1984.

[MSW04]   A.-M. Matache, C. Schwab, and T. P. Wihler. Fast numerical solution of parabolic integro-differential equations with applications in finance. Technical report, IMA University of Minnesota, 2004. Reseach report No. 1954.

[MvPS04] A.-M. Matache, T. von Petersdoff, and C. Schwab. Fast deterministic pricing of Lévy driven assets. *Mathematical Modelling and Numerical Analysis*, 38(1):37–72, 2004.

[Pha98]   H. Pham. Optimal stopping of controlled jump-diffusion processes: A viscosity solution approach. *Journal of Mathematical Systems*, 8(1):1–27, 1998.

[Pir84]   O. Pironneau. *Optimal shape design for elliptic systems*. Springer Series in Computational Physics. Springer-Verlag, New York, 1984.

An Operator Splitting Method for Pricing American Options

Samuli Ikonen[1] and Jari Toivanen[2]

[1] Nordea Markets, FI-00020 Nordea, Finland `Samuli.Ikonen@nordea.com`
[2] Department of Mathematical Information Technology, P.O. Box 35 (Agora),
    FI-40014 University of Jyväskylä, Finland `Jari.Toivanen@mit.jyu.fi`

Summary. Pricing American options using partial (integro-)differential equation based methods leads to linear complementarity problems (LCPs). The numerical solution of these problems resulting from the Black–Scholes model, Kou's jump-diffusion model, and Heston's stochastic volatility model are considered. The finite difference discretization is described. The solutions of the discrete LCPs are approximated using an operator splitting method which separates the linear problem and the early exercise constraint to two fractional steps. The numerical experiments demonstrate that the prices of options can be computed in a few milliseconds on a PC.

1 Introduction

Since 1973 Black, Scholes, and Merton developed models for pricing options in [BS73, Mer73] and, on the other hand, the Chicago Board Options Exchange started to operate, the trading of options has grown to tremendous scale. Basic options give either the right to sell (put) or buy (call) the underlying asset with the strike price. European options can be exercised only at the expiry time while American options can be exercised anytime before the expiry. The Black–Scholes partial differential equation (PDE) describes the evolution of the option price in time for European options. In order to avoid arbitrage opportunities with an American option, the so-called early exercise constraint has to be posed on its value. Combining this constraint with the PDE leads to a linear complementarity problem (LCP). For European options it is generally possible to derive formulas for their price, but American options usually need to be priced numerically. This paper considers the solution of these pricing problems. For the general discussion on these topics, we refer to the books [AP05, CT04, TR00, Wil98].

   The Black–Scholes model [BS73] assumes a constant volatility for all options with different strike prices and expiry times on the same underlying asset. In practice, this does not hold in the markets. One possibility to make

the prices consistent with the markets is to define the volatility as a function of time and the value of the underlying asset, and then calibrate this function; see [Dup94], for example. In 1976, Merton suggested to add jumps to the model of the underlying asset in [Mer73]. This jump-diffusion model helps to explain a good part of difference between the market prices and the ones given by the Black–Scholes model with a constant volatility. Since then there has been growing activity to incorporate jumps to the model; see [CT04] and references therein. One of the models used in this paper is Kou's jump-diffusion model. Another generalization is to make the volatility a stochastic process. Several such multifactor models have been proposed; see [FPS00], for example. Here Heston's stochastic volatility model [Hes93] is used. One can also combine stochastic volatility and jump models like in [Bat96, DPS00], for example.

Several ways to solve the discretized LCPs resulting from pricing American options have been described in the literature. Maybe the most common method is the project SOR iteration proposed in [Cry71]. This method is fairly generic and easy to implement, but its convergence rate degrades as grids are refined. For one-dimensional PDE models the resulting LCPs can be solved very efficiently using the direct algorithm in [BS77] if the matrix is a tridiagonal M-matrix and the solution has suitable form. The full matrices resulting from jump-diffusion models require special techniques in order to obtain efficient algorithms. The papers [AO05, AA00, CV05, MSW05] study the numerical pricing of European options, and in [dFL04, dFV05, Toi06] the pricing of American options is considered. For higher-dimensional problems like the ones resulting from Heston's model multigrid methods have been considered in [BC83, CP99, Oos03, RW04], for example. An alternative way is to approximate the LCPs using a penalty method [FV02, ZFV98]. This leads to a sequence system of linear equations with varying matrices. With this approach the constraints are always slightly violated. With a fairly similar Lagrange multiplier method [AP05, HIK03, IK06, IT06b] it can be guaranteed that the constraints are satisfied.

This paper considers an operator splitting method proposed for the Black–Scholes model in [IT04a]. The method was applied to Heston's model and analyzed in [IT04b], and for Kou's model it was applied in [Toi06]. The basic idea of this method is to split a time step with a LCP to two fractional time steps. The first fractional step requires a system of linear equations to be solved and the second one enforces the early exercise constraint. The update to satisfy the constraint is simple and, thus, the main computational burden will be the solution linear systems. A similar approach is commonly used to treat the incompressibility condition in the computational fluid dynamics; see [Glo03], for example. The operator splitting method has two obvious benefits. There are several efficient methods available for solving resulting systems of linear equations while only a few methods are available for the original LCPs and they usually cannot compete in the efficiency. Secondly the operator splitting method is easier to implement than an efficient LCP solver. This

paper demonstrates that the operator splitting method is suitable for pricing American options with different models and that the computation of a sufficiently accurate price for most purposes requires only a few milliseconds on a contemporary PC.

Outline of the paper is the following. We begin by describing the three models and the resulting P(I)DEs for European options. After this we formulate linear complementarity problems for the value of American options. Next we sketch finite difference discretizations for the partial differential operators. Then the operator splitting method is described and after this methods for solving the resulting systems of linear equations are discussed. The paper ends with numerical examples with all of the considered models and conclusions.

2 Models

2.1 Black–Scholes Model

By assuming that the value of the underlying asset denoted by x follows a geometric Brownian motion with a drift, the Black–Scholes PDE [BS73]

$$v_t = A_{BS}v = -\frac{1}{2}(\sigma x)^2 v_{xx} - rxv_x + rv \tag{1}$$

can be derived for the value of an option denoted by v, where σ is the volatility of the value of the asset and r is the risk free interest rate. In practice, the market prices of options do not satisfy (1). One possible way to make the model to match the markets is to use a volatility function σ which depends on the value of the underlying asset and time; see [AP05, Dup94], for example. In this case, the volatility function has to be calibrated with the market data.

2.2 Jump-Diffusion Models

When there is a high market stress like the crash of 1987 the value of assets can move faster than a geometric Brownian motion would predict. Partly due to this, models which allow also jumps for the value of asset have become more common; see [CT04] and references therein. Already in 1976 Merton considered such a model in [Mer76]. With independent and identically distributed jumps a partial integro-differential equation (PIDE)

$$v_t = A_{JD}v = -\frac{1}{2}(\sigma x)^2 v_{xx} - (r - \mu\zeta)xv_x + (r+\mu)v - \mu \int_{\mathbb{R}_+} v(t,xy)f(y)\,dy \tag{2}$$

can be derived for the value of an option, where μ is the rate of jumps, the function f defines the distributions of jumps, and ζ is the mean jump amplitude.

Merton used a Gaussian distribution for jumps in [Mer76]. Kou considered in [Kou02] a log-double-exponential distribution for jumps which leads a more flexible and tractable model. In this case, the density is

$$f(y) = \begin{cases} q\alpha_2 y^{\alpha_2 - 1}, & y < 1, \\ p\alpha_1 y^{-\alpha_1 - 1}, & y \geq 1, \end{cases} \tag{3}$$

where p, q, $\alpha_1 > 1$, and α_2 are positive constants such that $p + q = 1$. The mean jump amplitude is $\zeta = \frac{p\alpha_1}{\alpha_1 - 1} + \frac{q\alpha_2}{\alpha_2 + 1} - 1$. We will employ this model in the numerical experiments. Also in this case one possible way to calibrate the model is to let the volatility σ be a function of time and asset value like in [AA00].

2.3 Stochastic Volatility Models

In practice, the volatility of the value of an asset is not a constant over time. Several models have been also developed for the behavior of the volatility. Among several stochastic volatility models probably the one developed by Heston in [Hes93] is the most popular. It assumes the volatility to be a mean-reverting process. Under the assumption that the market price of risk is zero Heston's model leads to the two-dimensional PDE

$$v_t = A_{SV}v = -\frac{1}{2}yx^2 v_{xx} - \rho\gamma yx v_{xy} - \frac{1}{2}\gamma^2 y v_{yy} - rxv_x - \alpha(\beta - y)v_y + rv, \tag{4}$$

where y is the variance, that is, the square of the volatility, β is the mean level of the variance, α is the rate of reversion on the mean level, and γ is the volatility of the variance. The correlation between the price of the underlying asset and its variance is ρ.

3 Linear Complementarity Problems

The value of an option at the expiry time T is given by

$$v(T, x) = g(x), \tag{5}$$

where the payoff function g depends on the type of the option. For example, for a put option with a strike price K it is

$$g(x) = \max\{K - x, 0\}. \tag{6}$$

The value v of an American option satisfies a linear complementarity problem (LCP)

$$\begin{cases} (v_t - Av) \geq 0, & v \geq g, \\ (v_t - Av)(v - g) = 0, \end{cases} \tag{7}$$

where A is one of the operators A_{BS}, A_{JD}, or A_{SV} defined by (1), (2), and (4), respectively.

The operator splitting method is derived from a formulation with a Lagrange multiplier λ after a temporal discretization. In the continuous level, the formulation with the Lagrange multiplier reads

$$\begin{cases} (v_t - Av) = \lambda, & \lambda \geq 0, \ v \geq g, \\ \lambda(v - g) = 0. \end{cases} \tag{8}$$

4 Discretizations

4.1 Spatial Discretizations

The LCPs are posed on infinite domain as there is no upper limit for the value of the asset and also for variance in the case of Heston's stochastic volatility model. In order to use finite difference discretizations for the spatial derivatives, the domain is truncated from sufficiently large values of x and y which are denoted by X and Y, respectively. The choice of X for the Black–Scholes model is considered in [KN00], for example. On the truncation boundaries a suitable boundary condition needs to be posed. For the one-dimensional models for put options, we use homogeneous Dirichlet boundary condition $v = 0$ at $x = X$. For Heston's model homogeneous Neumann boundary conditions are posed. While these are fairly typical choices for boundary conditions there are also other choices.

For the interval $[0, X]$, we define subintervals $[x_{i-1}, x_i]$, $i = 1, 2, \ldots, m$, where x_is satisfy $0 = x_0 < x_1 < \cdots < x_m = X$. For Heston's model, the interval $[0, Y]$ is similarly divided by the points $0 = y_0 < y_1 < \cdots < x_n = Y$. Finite difference discretizations seek approximations for the value of v at the grid points x_is for one-dimensional models and (x_i, y_j) for Heston's model. The spatial partial derivatives appearing in (7) and (8) needs to be approximated using the grid point values. For the second-order derivative with respect to x, we use a finite difference approximation

$$v_{xx}(t, x_i) \approx \frac{2}{\Delta x_{i-1}(\Delta x_{i-1} + \Delta x_i)} v(t, x_{i-1}) - \frac{2}{\Delta x_{i-1}\Delta x_i} v(t, x_i)$$
$$+ \frac{2}{\Delta x_i(\Delta x_{i-1} + \Delta x_i)} v(t, x_{i+1}), \quad (9)$$

where $\Delta x_{i-1} = x_i - x_{i-1}$ and $\Delta x_i = x_{i+1} - x_i$. For the first-order derivative, one possible approximation is

$$v_x(t, x_i) \approx -\frac{\Delta x_i}{\Delta x_{i-1}(\Delta x_{i-1} + \Delta x_i)} v(t, x_{i-1}) + \frac{\Delta x_i - \Delta x_{i-1}}{\Delta x_{i-1}\Delta x_i} v(t, x_i)$$
$$+ \frac{\Delta x_{i-1}}{\Delta x_i(\Delta x_{i-1} + \Delta x_i)} v(t, x_{i+1}). \quad (10)$$

For Heston's model the approximations for the partial derivatives with respect to y can be defined analogously. The approximations (9) and (10) can be shown to be second-order accurate with respect to the grid step size when the step size varies smoothly; see [MW86], for example.

When the coefficient for the first-order derivative is large compared to the coefficient of the second-order derivative, the above discretizations lead to matrices with positive off-diagonal entries. In this case the matrix cannot have the M-matrix property and the resulting numerical solutions can have oscillations. This situation can be avoided by using locally one-sided differences for the first-order derivative. The drawback of this approach is that it reduces the order of accuracy to be first-order with respect to the grid step size. Nevertheless we will use this choice to ensure that the spatial discretizations lead to M-matrices and, thus, stable discretizations.

Special care must be taken when discretizing the cross derivative v_{xy} in Heston's model if M-matrices are sought. In [IT05], a seven-point stencil leading an M-matrix is described. With strong correlation between the value of asset and its volatility there can be severe restrictions on grid step sizes in order to obtain M-matrices and accurate discretizations.

The discretization of the integral term in the jump-diffusion model (2) leads to a full matrix; see [AO05, dFL04, MSW05], for example. Computationally it is expensive to operate with the full matrix and, due to this, different fast ways have been proposed for operating with it in the above mentioned articles. Fortunately, with Kou's log-double-exponential f in (2) is possible to derive recursive formulas with optimal computational complexity for evaluating quadratures for the integrals. This has been described in [Toi06] and we will employ this approach with our numerical experiments.

The grid point values of v are collected to a vector $\mathbf{v}$. Similarly we define a vector $\mathbf{g}$ containing the grid point values of the payoff function g. The spatial discretization leads to a semi-discrete form of the LCP (7) given by

$$\begin{cases} (\mathbf{v}_t - \mathbf{A}\mathbf{v}) \geq \mathbf{0}, & \mathbf{v} \geq \mathbf{g}, \\ (\mathbf{v}_t - \mathbf{A}\mathbf{v})^T(\mathbf{v} - \mathbf{g}) = 0, \end{cases} \tag{11}$$

where the matrix $\mathbf{A}$ is defined by the used finite differences and the inequalities of vectors are componentwise. The semi-discrete form with the Lagrange multiplier λ corresponding to (8) reads

$$\begin{cases} (\mathbf{v}_t - \mathbf{A}\mathbf{v}) = \boldsymbol{\lambda}, & \boldsymbol{\lambda} \geq \mathbf{0}, \ \mathbf{v} \geq \mathbf{g}, \\ \boldsymbol{\lambda}^T(\mathbf{v} - \mathbf{g}) = 0, \end{cases} \tag{12}$$

where the vector $\boldsymbol{\lambda}$ contains the grid point values of the Lagrange multiplier.

4.2 Temporal Discretization

For the temporal discretization the time interval $[0, T]$ is divided into subintervals which are defined by the times $0 = t_0 < t_1 < \cdots < t_l = T$. The vector containing the grid point values of v at t_k is denoted by $\mathbf{v}^{(k)}$.

Usually in option pricing problems the backward time stepping is started from a non-smooth final value. Due to this, the time stepping scheme should have good damping properties in order to avoid oscillations. For example, the popular Crank–Nicolson method does not have good damping properties and it can lead to approximations with excessive oscillations. Instead of it we employ the Rannacher time-stepping scheme [Ran84]. In the option pricing context it has been analyzed recently in [GC06].

In the Rannacher time-stepping scheme a few first time steps are performed with the implicit Euler method and then the Crank–Nicolson method is used. This leads to second-order accuracy and good damping properties. For the semi-discrete LCP (11) the scheme reads

$$
\begin{cases}
\mathbf{B}^{(k)}\mathbf{v}^{(k)} - \mathbf{C}^{(k)}\mathbf{v}^{(k+1)} - \mathbf{f}^{(k)} \geq \mathbf{0}, \qquad \mathbf{v}^{(k)} \geq \mathbf{g}, \\
\left(\mathbf{B}^{(k)}\mathbf{v}^{(k)} - \mathbf{C}^{(k)}\mathbf{v}^{(k+1)} - \mathbf{f}^{(k)}\right)^{T} \left(\mathbf{v}^{(k)} - \mathbf{g}\right) = 0,
\end{cases}
\tag{13}
$$

for $k = l - 1, \ldots, 0$, where

$$
\mathbf{B}^{(k)} = \mathbf{I} + \theta_k \Delta t_k \mathbf{A}, \qquad \mathbf{C}^{(k)} = \mathbf{I} - (1 - \theta_k)\Delta t_k \mathbf{A},
\tag{14}
$$

and $\mathbf{f}^{(k)}$ is due to possible non-homogeneous Dirichlet boundary conditions. When the first four time steps are performed with the implicit Euler method the parameter θ_k is defined by

$$
\theta_k =
\begin{cases}
1, & k = l - 1, \ldots, l - 4, \\
\frac{1}{2}, & k = l - 5, \ldots, 0.
\end{cases}
\tag{15}
$$

The temporal discretization of the semi-discrete form with the Lagrange multiplier (12) leads to

$$
\begin{cases}
\mathbf{B}^{(k)}\mathbf{v}^{(k)} - \mathbf{C}^{(k)}\mathbf{v}^{(k+1)} - \mathbf{f}^{(k)} = \Delta t_k \boldsymbol{\lambda}^{(k)}, \qquad \boldsymbol{\lambda}^{(k)} \geq \mathbf{0}, \ \mathbf{v}^{(k)} \geq \mathbf{g}, \\
\left(\boldsymbol{\lambda}^{(k)}\right)^{T} \left(\mathbf{v}^{(k)} - \mathbf{g}\right) = 0,
\end{cases}
\tag{16}
$$

for $k = l - 1, \ldots, 0$.

5 Operator Splitting Method

Here we describe an operator splitting method [IT04a] which approximates the solution of the LCP in (16) by two fractional time steps. The first step requires the solution of a system of linear equations and the second step updates the solution and Lagrange multiplier to satisfy the linear complementarity conditions. The advantage of this approach is that it simplifies the solution procedure and allows to use any efficient method for solving linear systems. More precisely, the steps in the operator splitting method are

$$
\mathbf{B}^{(k)}\tilde{\mathbf{v}}^{(k)} = \mathbf{C}^{(k)}\mathbf{v}^{(k+1)} + \mathbf{f}^{(k)} + \Delta t_k \boldsymbol{\lambda}^{(k+1)}
\tag{17}
$$

and

$$\begin{cases} \mathbf{v}^{(k)} - \tilde{\mathbf{v}}^{(k)} - \Delta t_k(\boldsymbol{\lambda}^{(k)} - \boldsymbol{\lambda}^{(k+1)}) = \mathbf{0}, \qquad \boldsymbol{\lambda}^{(k)} \geq \mathbf{0}, \ \mathbf{v}^{(k)} \geq \mathbf{g}, \\ \left(\boldsymbol{\lambda}^{(k)}\right)^T \left(\mathbf{v}^{(k)} - \mathbf{g}\right) = 0. \end{cases} \tag{18}$$

The first step (17) uses the Lagrange multiplier vector $\boldsymbol{\lambda}^{(k+1)}$ from the previous step and not $\boldsymbol{\lambda}^{(k)}$ which leads to the decoupling of the linear system and the constraints. The second step does not have any spatial couplings and the update can be made quickly by going through components of the vectors $\mathbf{v}^{(k)}$ and $\boldsymbol{\lambda}^{(k)}$ one by one. Due to this, the main computational cost is the solution of the linear system in the first step (17). Under reasonable assumptions it can be shown that the difference between the solutions of the original time stepping and the operator splitting time stepping is second-order with respect to the time step size [IT04b]. Hence, it does not reduce the order of accuracy compared to second-order accurate time stepping method like the Rannacher scheme.

6 Solution of Linear Systems

In each time step with the operator splitting method it is necessary to solve a system of linear equations with the matrix $\mathbf{B}$ defined in (14). Here and in the following we have omitted the subscript (k) in order to simplify the notations. The Black–Scholes PDE leads to a tridiagonal $\mathbf{B}$ with the above finite difference discretization. In this case the linear systems can be solved efficiently using the $\mathbf{LU}$ decomposition.

With the jump-diffusion models $\mathbf{B}$ is a full matrix and the use of $\mathbf{LU}$ decomposition would be computationally too expensive. We adopt the approach proposed in [AO05, dFV05] which is an iterative method based on a regular splitting of $\mathbf{B}$. We use the splitting

$$\mathbf{B} = \mathbf{T} - \mathbf{R}, \tag{19}$$

where $\mathbf{R}$ is the full matrix resulting from the integral term and, thus, $\mathbf{T}$ is a tridiagonal matrix defined by other terms. Now the iterative method for a system $\mathbf{Bv} = \mathbf{b}$ reads

$$\mathbf{v}^{l+1} = \mathbf{T}^{-1}\left(\mathbf{b} + \mathbf{Rv}^l\right), \qquad l = 0, 1, \dots, \tag{20}$$

where $\mathbf{v}^0$ is the initial guess taken to be the solution from the previous time step. The solutions with $\mathbf{T}$, that is, multiplications with $\mathbf{T}^{-1}$ can be computed efficiently using $\mathbf{LU}$ decomposition. The multiplications with $\mathbf{R}$ can be performed using the fast recursion formulas in [Toi06] when Kou's model is used. Furthermore, it has been shown in [dFV05] that the iteration (20) converges fast. As the numerical experiments will demonstrate, usually two or three iterations are enough to obtain the solution with sufficient accuracy.

With Heston's model $\mathbf{B}$ is a block tridiagonal matrix corresponding to a two-dimensional PDE. Furthermore, $\mathbf{B}$ is usually not well conditioned partly due to varying coefficient in the PDE. In order to obtain a method with optimal computational complexity, we will employ a multigrid method. The analysis in [Oos03] shows that a multigrid with an alternating direction smoother is robust with respect to all parameters in the problem and discretization. This smoother is computationally more expensive and complicated to implement than point smoothers, but we used it as it guarantees a fast multigrid convergence. The grid transfers are performed using full weighting restriction and bilinear prolongation.

7 Numerical Results

In our numerical examples we price American put options with the parameters

$$\sigma = 0.25, \quad r = 0.1, \quad T = 0.25, \quad \text{and} \quad K = 10. \tag{21}$$

The additional parameters for Kou's and Heston's models are defined in the subsequent sections. In Table 1, we have collected reference option prices for three asset values. They are computed with very fine discretizations for the one-dimensional models on the interval $[0, 40]$ and the prices under Heston's model are from [IT06b] with $y = 0.0625$. Fig. 1 shows the price of the option as a function of x computed with the different models in the interval $8.5 \leq x \leq 12.5$.

In the following tables all CPU times are given in milliseconds on a PC with 3.8 GHz Intel Xeon processor and implementations have been made using Fortran.

7.1 Black–Scholes Model

Based on a few numerical experiments using the model parameters in (21) we observed that the truncation boundary can be chosen to be $X = 2K = 20$ with the truncation error being so small that it does not influence the first five decimals of the prices at $x = 9$, 10, and 11. We define the spatial grid as

$$x_i = \left(1 + \frac{\sinh(\beta(i/n - \gamma))}{\sinh(\beta\gamma)}\right) K, \qquad i = 0, 1, \ldots, m, \tag{22}$$

Table 1. Reference prices for options with the different models

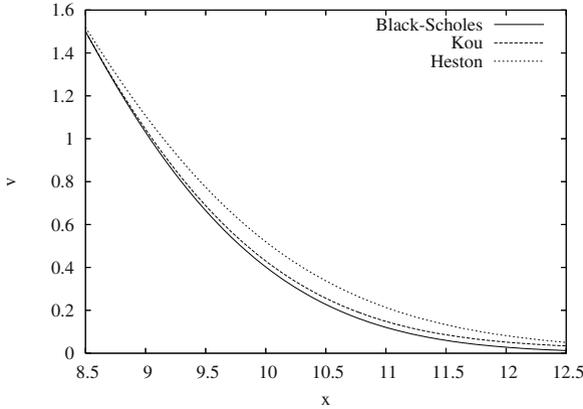| model \ asset value	$x = 9$	$x = 10$	$x = 11$
Black–Scholes	1.030463	0.402425	0.120675
Kou	1.043796	0.429886	0.148625
Heston	1.107621	0.520030	0.213677

Fig. 1. The price of the option with respect to the value of the underlying asset for the three different models.

Table 2. Results for different grids with Black–Scholes model

l	m	error	ratio	time
10	20	0.01056		0.02
18	40	0.00208	5.1	0.06
34	80	0.00058	3.6	0.21
66	160	0.00022	2.7	0.79
130	320	0.00007	3.3	3.08

where we have chosen $\beta = 6$ and $\gamma = 1/2$ which leads to some refinement near the strike price K. For the temporal discretization, we choose the approximation times to be

$$t_k = \left(\frac{a^{-k/(l-2)} - 1}{a^{-1} - 1} \right) T, \qquad k = 0, 1, \ldots, l - 4, \tag{23}$$

and

$$t_k = \left(\frac{a^{-(k+l-4)/(2l-4)} - 1}{a^{-1} - 1} \right) T, \qquad k = l - 3, \ldots, l. \tag{24}$$

The parameter a in (23) and (24) has been chosen to be $a = 2$ which leads to a mild refinement near the expiry.

Table 2 reports the l_2 errors computed using the reference prices in Table 1 at $x = 9$, 10, and 11 for five different space-time grids. The ratio column in the table gives the ratios between two successive l_2 errors. The time is the CPU time in milliseconds needed to price the options.

7.2 Kou's Jump-Diffusion Model

The parameters defining the jump probability and its distribution in Kou's model are chosen to be

Table 3. Results for different grids with Kou's model

l	m	error	ratio	iter	time
10	20	0.01050		3.1	0.10
18	40	0.00231	4.5	3.0	0.29
34	80	0.00056	4.1	3.0	0.97
66	160	0.00022	2.6	2.3	2.95
130	320	0.00006	3.7	2.0	10.17

Table 4. Results for different grids with Heston's model

l	m	n	error	ratio	iter	time
10	20	8	0.02576		1.0	0.7
18	40	16	0.00574	4.5	1.3	5.7
34	80	32	0.00420	1.4	2.0	59.4
66	160	64	0.00049	8.5	2.0	487.5
130	320	128	0.00012	4.1	2.0	4373.7

$$\alpha_1 = 3, \quad \alpha_2 = 3, \quad p = \frac{1}{3}, \quad \text{and} \quad \mu = 0.1. \tag{25}$$

We have used the same space-time grids as with the Black–Scholes model. Table 3 reports the errors, their ratios and CPU times in milliseconds. The column iter in the table gives the average number of the iterations (20). The stopping criterion for the iterations was that the norm of the residual vector is less than 10^{-11} times the norm of the right-hand side vector.

7.3 Heston's Stochastic Volatility Model

In Heston's model the behavior of the stochastic volatility and its correlation with the value of the asset are described by the parameters

$$\alpha = 5, \quad \beta = 0.16, \quad \gamma = 0.9, \quad \text{and} \quad \rho = 0.1. \tag{26}$$

The values of these parameters are the same as in many previous studies including [CP99, IT07, Oos03, ZFV98]. The computational domain is truncated at $X = 20$ and $Y = 1$ like also in [Oos03, IT07], for example. We use the same non-uniform grids as in [IT05] and the parameter w in the discretization of the cross derivative (not discussed in this paper) is chosen using the formula in [IT07]. For the time stepping we use uniform time steps.

Table 4 reports the errors, their ratios, the average number of multigrid iterations, and CPU times in milliseconds. The stopping criterion for the multigrid iterations was that the norm of the residual vector is less than 10^{-6} times the norm of the right-hand side vector.

8 Conclusions

We described an operator splitting method for solving linear complementarity problems (LCPs) resulting from American option pricing problems. We considered it in the case of the Black–Scholes model, Kou's jump-diffusion model, and Heston's stochastic volatility model for the value of the underlying asset. The numerical results demonstrated that with all these models the prices can be computed in a few milliseconds on a PC.

As future research one could consider the construction of adaptive discretization; see [AP05, LPvST07], for example. Also the robustness and accuracy of discretizations for Heston's model with higher correlations could be studied. A natural generalization would be to extent the methods for stochastic volatility models including jumps like the ones in [Bat96, DPS00].

References

[AA00]   L. Andersen and J. Andreasen. Jump-diffusion processes: Volatility smile fitting and numerical methods for option pricing. *Rev. Deriv. Res.*, 4:231–262, 2000.

[AO05]   A. Almendral and C. W. Oosterlee. Numerical valuation of options with jumps in the underlying. *Appl. Numer. Math.*, 53:1–18, 2005.

[AP05]   Y. Achdou and O. Pironneau. *Computational methods for option pricing*, volume 30 of *Frontiers in Applied Mathematics*. SIAM, Philadelphia, PA, 2005.

[Bat96]   D. S. Bates. Jumps and stochastic volatility: Exchange rate processes implicit Deutsche mark options. *Review Financial Stud.*, 9:69–107, 1996.

[BC83]   A. Brandt and C. W. Cryer. Multigrid algorithms for the solution of linear complementarity problems arising from free boundary problems. *SIAM J. Sci. Statist. Comput.*, 4:655–684, 1983.

[BS73]   F. Black and M. Scholes. The pricing of options and corporate liabilities. *J. Polit. Econ.*, 81:637–654, 1973.

[BS77]   M. J. Brennan and E. S. Schwartz. The valuation of American put options. *J. Finance*, 32:449–462, 1977.

[CP99]   N. Clarke and K. Parrott. Multigrid for American option pricing with stochastic volatility. *Appl. Math. Finance*, 6:177–195, 1999.

[Cry71]   C. W. Cryer. The solution of a quadratic programming problem using systematic overrelaxation. *SIAM J. Control*, 9:385–392, 1971.

[CT04]   R. Cont and P. Tankov. *Financial modelling with jump processes*. Chapman & Hall/CRC, Boca Raton, FL, 2004.

[CV05]   R. Cont and E. Voltchkova. A finite difference scheme for option pricing in jump diffusion and exponential Lévy models. *SIAM J. Numer. Anal.*, 43:1596–1626, 2005.

[dFL04]   Y. d'Halluin, P. A. Forsyth, and G. Labahn. A penalty method for American options with jump diffusion processes. *Numer. Math.*, 97:321–352, 2004.

[dFV05] Y. d'Halluin, P. A. Forsyth, and K. R. Vetzal. Robust numerical methods for contingent claims under jump diffusion processes. *IMA J. Numer. Anal.*, 25:87–112, 2005.

[DPS00] D. Duffie, J. Pan, and K. Singleton. Transform analysis and asset pricing for affine jump-diffusions. *Econometrica*, 68(6):1343–1376, 2000.

[Dup94] B. Dupire. Pricing with a smile. *Risk*, 7:18–20, 1994.

[FPS00] J.-P. Fouque, G. Papanicolaou, and K. R. Sircar. *Derivatives in financial markets with stochastic volatility*. Cambridge University Press, Cambridge, 2000.

[FV02] P. A. Forsyth and K. R. Vetzal. Quadratic convergence for valuing American options using a penalty method. *SIAM J. Sci. Comput.*, 23:2095–2122, 2002.

[GC06] M. B. Giles and R. Carter. Convergence analysis of Crank-Nicolson and Rannacher time-marching. *J. Comput. Finance*, 9:89–112, 2006.

[Glo03] R. Glowinski. Finite element methods for incompressible viscous flow. In P. G. Ciarlet and J.-L. Lions, editors, *Handbook of Numerical Analysis, Vol. IX*, pages 3–1176. North-Holland, Amsterdam, 2003.

[Hes93] S. Heston. A closed-form solution for options with stochastic volatility with applications to bond and currency options. *Rev. Financial Stud.*, 6:327–343, 1993.

[HIK03] M. Hintermüller, K. Ito, and K. Kunisch. The primal-dual active set strategy as a semismooth Newton method. *SIAM J. Optim.*, 13:865–888, 2003.

[IK06] K. Ito and K. Kunisch. Parabolic variational inequalities: The Lagrange multiplier approach. *J. Math. Pures Appl.*, 85:415–449, 2006.

[IT04a] S. Ikonen and J. Toivanen. Operator splitting methods for American option pricing. *Appl. Math. Lett.*, 17:809–814, 2004.

[IT04b] S. Ikonen and J. Toivanen. Operator splitting methods for pricing American options with stochastic volatility. Reports of the Department of Mathematical Information Technology, Series B, Scientific Computing B11/2004, University of Jyväskylä, Jyväskylä, 2004.

[IT05] S. Ikonen and J. Toivanen. Componentwise splitting methods for pricing American options under stochastic volatility. Reports of the Department of Mathematical Information Technology, Series B, Scientific Computing B7/2005, University of Jyväskylä, Jyväskylä, 2005.

[IT07] S. Ikonen and J. Toivanen. Componentwise splitting methods for pricing American options under stochastic volatility. *Int. J. Theor. Appl. Finance*, 10(2):331–361, 2007.

[IT06b] K. Ito and J. Toivanen. Lagrange multiplier approach with optimized finite difference stencils for pricing American options under stochastic volatility. Reports of the Department of Mathematical Information Technology, Series B, Scientific Computing B6/2006, University of Jyväskylä, Jyväskylä, 2006.

[KN00] R. Kangro and R. Nicolaides. Far field boundary conditions for Black-Scholes equations. *SIAM J. Numer. Anal.*, 38:1357–1368, 2000.

[Kou02] S. G. Kou. A jump-diffusion model for option pricing. *Management Sci.*, 48:1086–1101, 2002.

[LPvST07] P. Lötstedt, J. Persson, L. von Sydow, and J. Tysk. Space-time adaptive finite difference method for European multi-asset options. *Comput. Math. Appl.*, 53(8):1159–1180, 2007.

[Mer73]   R. C. Merton. Theory of rational option pricing. *Bell J. Econom. and Management Sci.*, 4:141–183, 1973.

[Mer76]   R. Merton. Option pricing when underlying stock returns are discontinuous. *J. Financial Econ.*, 3:125–144, 1976.

[MSW05]   A.-M. Matache, C. Schwab, and T. P. Wihler. Fast numerical solution of parabolic integrodifferential equations with applications in finance. *SIAM J. Sci. Comput.*, 27:369–393, 2005.

[MW86]   T. A. Manteuffel and A. B. White, Jr. The numerical solution of second-order boundary value problems on nonuniform meshes. *Math. Comp.*, 47:511–535, 1986.

[Oos03]   C. W. Oosterlee. On multigrid for linear complementarity problems with application to American-style options. *Electron. Trans. Numer. Anal.*, 15:165–185, 2003.

[Ran84]   R. Rannacher. Finite element solution of diffusion problems with irregular data. *Numer. Math.*, 43:309–327, 1984.

[RW04]   C. Reisinger and G. Wittum. On multigrid for anisotropic equations and variational inequalities: pricing multi-dimensional European and American options. *Comput. Vis. Sci.*, 7(3–4):189–197, 2004.

[Toi06]   J. Toivanen. Numerical valuation of European and American options under Kou's jump-diffusion model. Reports of the Department of Mathematical Information Technology, Series B, Scientific Computing B11/2006, University of Jyväskylä, Jyväskylä, 2006.

[TR00]   D. Tavella and C. Randall. *Pricing financial instruments: The finite difference method.* John Wiley & Sons, Chichester, 2000.

[Wil98]   P. Wilmott. *Derivatives.* John Wiley & Sons, Chichester, 1998.

[ZFV98]   R. Zvan, P. A. Forsyth, and K. R. Vetzal. Penalty methods for American options with stochastic volatility. *J. Comput. Appl. Math.*, 91:199–218, 1998.