Springer Proceedings in Mathematics & Statistics

Robert Klöfkorn
Eirik Keilegavlen
Florin A. Radu
Jürgen Fuhrmann *Editors*

Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples

FVCA 9, Bergen, Norway, June 2020

Springer

Springer Proceedings in Mathematics & Statistics

Volume 323

Springer Proceedings in Mathematics & Statistics

This book series features volumes composed of selected contributions from workshops and conferences in all areas of current research in mathematics and statistics, including operation research and optimization. In addition to an overall evaluation of the interest, scientific quality, and timeliness of each proposal at the hands of the publisher, individual contributions are all refereed to the high quality standards of leading journals in the field. Thus, this series provides the research community with well-edited, authoritative reports on developments in the most exciting areas of mathematical and statistical research today.

More information about this series at http://www.springer.com/series/10533

Robert Klöfkorn · Eirik Keilegavlen ·
Florin A. Radu · Jürgen Fuhrmann
Editors

Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples

FVCA 9, Bergen, Norway, June 2020

In two volumes

Springer

Editors
Robert Klöfkorn
NORCE Norwegian Research Centre AS
Bergen, Norway

Eirik Keilegavlen
Department of Mathematics
University of Bergen
Bergen, Norway

Florin A. Radu
Department of Mathematics
University of Bergen
Bergen, Norway

Jürgen Fuhrmann
Weierstrass Institute for Applied Analysis
and Stochastics
Berlin, Germany

ISSN 2194-1009 ISSN 2194-1017 (electronic)
Springer Proceedings in Mathematics & Statistics
ISBN 978-3-030-43650-6 ISBN 978-3-030-43651-3 (eBook)
https://doi.org/10.1007/978-3-030-43651-3

Mathematics Subject Classification: 65-06, 65Mxx, 65Nxx, 76xx, 78xx, 85-08, 86-08, 92-08

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Organization

Program Chairs

Robert Klöfkorn, NORCE Norwegian Research Centre AS, Norway
Eirik Keilegavlen, University of Bergen, Norway
Florin A. Radu, University of Bergen, Norway
Jürgen Fuhrmann, Weierstrass-Institut fur Angewandte Analysis und Stochastik, Germany

Program Committee

Sarah Gasda, NORCE Norwegian Research Centre, Norway
Robert Eymard, Université Paris-Est Marne-la-Vallée, France
Philipp Birken, Lund University, Sweden
Claire Chainais-Hillairet, Université Lille, France
Peter Frolkovic, Slovak University of Technology in Bratislava, Slovakia
Andreas Dedner, University of Warwick, UK
Clément Cancès, Inria Lille - Nord Europe, France
Paola Antonietti, Politecnico di Milano, Italy
Christoph Erath, Technical University Darmstadt, Germany
Knut-Andreas Lie, SINTEF Digital, Norway
Iuliu Sorin Pop, Hasselt University, Belgium
Donna Calhoun, Boise State University, Idaho, USA
Roland Masson, Université Cote d'Azur, France
Inga Berre, University of Bergen, Norway
Ivan Yotov, University of Pittsburgh, Pennsylvania, USA
Thierry Gallouet, Aix-Marseille Université, France
Christian Rohde, University of Stuttgart, Germany

Preface

The Finite Volume method in its various forms is a discretization technique for partial differential equations based on the fundamental physical principle of conservation. It has been used successfully in many applications including fluid dynamics, magnetohydrodynamics, structural analysis, nuclear physics, semiconductor theory, and electrochemistry. Recent decades have brought significant success in the theoretical understanding of the method. Many finite volume methods preserve further qualitative or asymptotic properties including maximum principles, dissipativity, monotone decay of the free energy, or asymptotic stability.

Due to these properties, finite volume methods belong to the wider class of compatible discretization methods, which preserve qualitative properties of continuous problems at the discrete level. This structural approach to the discretization of partial differential equations becomes particularly important for complex multiphysics and multiscale applications.

The triennal series of conferences “Finite Volumes for Complex Applications (FVCA)” brings together mathematicians, physicists, and engineers interested in this kind of physically motivated discretizations. The focus of the conference series is two-fold. Further development and advancement of the theoretical understanding of suitable finite volume, finite element, discontinuous Galerkin, and other discretization schemes provides a sound foundation for these methods. On the other hand, practical examples showcase the usefulness of the approach for modeling, simulation, and optimization in academia and industry.

Previous conferences of this series have been held in Rouen (1996), Duisburg (1999), Porquerolles (2002), Marrakech (2005), Aussois (2008), Prague (2011), Berlin (2014), and Lille (2017).

The present volumes contains the invited and contributed papers presented as posters or talks at the 9th International Symposium on Finite Volumes for Complex Applications held as an online event organized by NORCE Norwegian Research Centre AS and University of Bergen June 15–19, 2020.

The contributions in the first volume deal with theoretical aspects of the method. They focus on topics like preservation of physical properties on the discrete level, convergence, stability and error analysis, physically consistent coupling between discretizations for different processes, connections to other discretization methods, relationship between grids and discretization schemes, complex geometries and adaptivity shock waves and other flow discontinuities, new and existing schemes and their limitations, and bottlenecks in the solution of large scale problems.

The practical value of finite volume and related methods is demonstrated by the contributions to the second volume of the proceedings. Application fields include atmosphere and ocean modeling, chemical engineering and combustion, energy generation and storage, and electro-reaction-diffusion systems and porous media.

The volume editors would like to thank the authors for their high quality contributions, the members of the program committee for supporting the organization of the review process, and all reviewers for their thorough work on the evaluation of each of the contributions.

The production of the proceedings was continuously supported by the Editor's team at Springer Verlag.

Bergen, Norway Robert Klöfkorn
Bergen, Norway Eirik Keilegavlen
Bergen, Norway Florin A. Radu
Berlin, Germany Jürgen Fuhrmann
February 2020

Contents

Invited Contributions

Theoretical Aspects

Practical Examples

Invited Contributions

Interplay Between Diffusion Anisotropy and Mesh Skewness in Hybrid High-Order Schemes

J. Droniou

Abstract We explore the effects of mesh skewness on the accuracy of standard Hybrid High-Order (HHO) schemes for anisotropic diffusion equations. After defining a notion of regular skewed mesh sequences, which allows, e.g., for elements that become more and more elongated during mesh refinement, we establish an error estimate in which we precisely track the dependency of the local multiplicative constants in terms of the diffusion tensor and mesh skewness. This dependency makes explicit an interplay between the local diffusion properties and the distortion of the elements. We then provide several numerical results to assess the practical convergence properties of HHO for highly anisotropic diffusion or highly distorted meshes. These tests indicate a more robust behaviour than the theoretical estimate predicts.

Keywords Hybrid high-order schemes · Anisotropy · Diffusion equation · Skewed meshes

1 Introduction

The last few years have seen a increased interest in novel discretisation methods, for diffusion equations, that support polytopal meshes (made of general polygons/polyhedra) and allow for arbitrary approximation orders: Hybridisable Discontinuous Galerkin methods [3], Virtual Element Methods (VEM) [2], Weak Galerkin Methods [11], etc. The Hybrid High-Order (HHO) method [6, 7] is one of these arbitrary-order polytopal methods, and shares with the aforementioned ones the hybrid structure of unknowns (contrary to Discontinuous Galerkin methods [5]), that is, unknowns located in the elements and on their faces. We refer to the introduction of [7] for a thorough review of the literature on polytopal methods. The HHO method can be seen as a high-order extension of the Hybrid Mimetic Mixed method [8] and, contrary to some other polytopal methods, it has a flux formulation

J. Droniou (✉)
School of Mathematics, Monash University, Melbourne, Victoria 3800, Australia
e-mail: jerome.droniou@monash.edu

R. Klöfkorn et al. (eds.), *Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples*, Springer Proceedings in Mathematics & Statistics 323,
https://doi.org/10.1007/978-3-030-43651-3_1

that makes it a Finite Volume method. Additionally, the design of HHO schemes is dimension-independent and has an enhanced compliance with the physics due to the construction of local problem-dependent reconstruction operators.

HHO schemes have been applied to and analysed for a variety of models (see [7] and references therein), with error estimates that have an explicit dependency on the physical data. These estimates are however obtained for “regular” polytopal meshes, that is, meshes whose elements are “isotropic” (not elongated in any particular direction, and whose faces have a diameter comparable to their elements’ diameters). In this work we analyse and numerically test the HHO scheme for highly anisotropic diffusion equations and families of distorted meshes, that no longer satisfy the usual regularity conditions. We consider the archetypal linear diffusion model

$$\begin{cases} -\nabla\cdot(\boldsymbol{K}\nabla u) = f & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega, \end{cases} \tag{1}$$

where Ω is a polytopal domain of $\mathbb{R}^d$, $\boldsymbol{K} : \Omega \to \mathbb{R}^{d\times d}_{\rm sym}$ is a symmetric bounded uniformly coercive diffusion tensor, and $f \in L^2(\Omega)$. The solution to (1) is taken in the classical weak sense.

A review of historical or polytopal numerical methods on distorted meshes is out of this paper’s scope. We however mention the recent works [1, 12] about the VEM on anisotropic meshes, which present numerical results for a Poisson problem with internal layer, and derive approximation properties of the relevant interpolators. The novelty of our work, besides considering the HHO method instead of the VEM, is to establish complete error estimates (not just interpolator approximation properties) that take into account not only the distortion of the mesh, but also the high anisotropy of the diffusion tensor and the subtle interplay between these two features. The approach used here can be adapted to other methods, such as VEM, to yield error estimates that account for this interplay.

This paper is organised as follows. The concept of regular skewed mesh sequences, for which the error analysis will be carried out, is introduced in Sect. 2; these meshes can have very elongated elements, provided that some local linear map transforms them into isotropic elements. The oblique elliptic projector is at the core of HHO schemes; its approximation properties on skewed elements are presented in Sect. 3, and are used in Sect. 4 to perform the error analysis of HHO schemes on skewed meshes. This analysis is based on local transports of each skewed element T into an isotropic element $\widehat{T}$; this transport identifies a new diffusion tensor on $\widehat{T}$, whose anisotropy properties dictate the contribution of T to the global error estimate. The error estimate stated in Theorem 2 therefore highlights how the diffusion anisotropy and the mesh skewness are combined in the multiplicative constants. This approach has the added advantage of leading to an error estimate that is as optimal as the standard error estimate for anisotropic diffusion models on regular (non-skewed) mesh sequences. In Sect. 5, we perform a series of tests to evaluate the practical impact of high diffusion anisotropy and mesh skewness on the accuracy of HHO schemes. Some of the conclusions drawn from these tests are predicted by the error estimate but, overall, the HHO scheme is found to be more robust with respect to the diffusion anisotropy and mesh skewness than what the theoretical analysis seems to indicate. A conclusion is provided in Sect. 6.

Notations. The Euclidean norm of a vector $\boldsymbol{\xi} \in \mathbb{R}^d$ is denoted by $|\boldsymbol{\xi}|$. If $L : (\mathbb{R}^d)^s \to \mathbb{R}$ is an s-linear map, we define the norm of L by

$$\mathrm{N}_s(L) := \sup\{|L(\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_s)| \,:\, \boldsymbol{\xi}_i \in \mathbb{R}^d\,,\ |\boldsymbol{\xi}_i| \le 1\,, \forall i = 1, \dots, s\}.$$

For X an open subset of $\mathbb{R}^n$, $n \in \{d, d-1\}$, $(\cdot, \cdot)_X$ and $\|\cdot\|_X$ denote respectively the $L^2(X)$- or $L^2(X)^n$-inner product and norm. Letting $D^s v$ be the s-th order differential of v, the $H^s(X)$-seminorm of a function $v \in H^s(X)$ is $|v|_{H^s(X)} := \|\mathrm{N}_s(D^s v)\|_X$.

2 Regular Skewed Mesh Sequences

Let us first briefly recall the definition of polytopal mesh, referring to [7, Sect. 1.1] for details. A polytopal mesh of Ω is a couple $\mathcal{M}_h = (\mathcal{T}_h, \mathcal{F}_h)$ where $\mathcal{T}_h$ is a collection of disjoint polytopes T—the elements—such that $\overline{\Omega} = \cup_{T \in \mathcal{T}_h} \overline{T}$, and $\mathcal{F}_h$ is the set of mesh faces whose closures form a partition of $\cup_{T \in \mathcal{T}_h} \partial T$, and such that each face is contained in one or two elements boundaries. Mesh faces can be different from the geometrical faces of the polytopes, the latter being possibly cut in two mesh faces in case of non-conforming mesh [7, Fig. 1.2]. The diameter of a subset X of $\mathbb{R}^d$ is denoted by h_X. The index h in $\mathcal{M}_h$ is the meshsize $h = \max_{T \in \mathcal{T}_h} h_T$. For $T \in \mathcal{T}_h$, we let $\mathcal{F}_T$ be the set of faces $F \in \mathcal{F}_h$ such that $\partial T = \cup_{F \in \mathcal{F}_T} \overline{F}$. The outer normal to T on $F \in \mathcal{F}_T$ is $\boldsymbol{n}_{TF}$. A matching simplicial mesh of $T \in \mathcal{T}_h$ is a polytopal mesh of T made of simplices and whose faces correspond to the geometrical simplicial faces.

We now define the concept of regular skewed mesh sequence, which allows for elements that become more and more stretched as h decreases, provided that each element can be linearly mapped onto an "isotropic" element, that satisfies the regularity conditions of a standard regular mesh sequence [7, Definition 1.9].

Definition 1 (*Regular Skewed Mesh Sequence*) Let $\mathcal{H} \subset (0, +\infty)$ be a countable set with 0 as only accumulation point. For each $h \in \mathcal{H}$, let $\mathcal{M}_h$ be a polytopal mesh and $\phi_h = (\phi_T)_{T \in \mathcal{T}_h}$ be a family of isomorphisms of $\mathbb{R}^d$. The sequence $(\mathcal{M}_h, \phi_h)_{h \in \mathcal{H}}$ is a *regular skewed mesh sequence* if there exists $\varrho \in (0, 1)$ such that, for all $h \in \mathcal{H}$ and all $T \in \mathcal{T}_h$, the following properties hold:

1. Setting $\widehat{T} = \phi_T(T)$, it holds $\varrho h_T \le h_{\widehat{T}}$ and $\varrho h_{\widehat{T}} \le h_T$.
2. There is a matching simplicial mesh $(\mathfrak{T}_{\widehat{T}}, \mathfrak{F}_{\widehat{T}})$ of $\widehat{T}$ such that, letting $\mathcal{F}_{\widehat{T}} := \{\widehat{F} := \phi_T(F) \,:\, F \in \mathcal{F}_T\}$ be the set of faces of $\widehat{T}$, for any face $\sigma \in \mathfrak{F}_{\widehat{T}}$, either $\sigma \cap \partial\widehat{T} = \emptyset$ or there is $\widehat{F} \in \mathcal{F}_{\widehat{T}}$ such that $\sigma \subset \widehat{F}$.
3. For all $\tau \in \mathfrak{T}_{\widehat{T}}$, it holds $\varrho h_{\widehat{T}} \le h_\tau$ and $\varrho h_\tau \le r_\tau$, where r_τ is the inradius of τ.

Remark 1 (Comparison with [12]) The notion of regular skewed mesh sequence is close to the notion of regular anisotropic mesh of [12], in particular in the usage of maps from skewed elements to isotropic elements. A noticeable difference, however, is the requirement in [12] that two neighbouring elements T, T' must have similar isotropy (that is, the corresponding mappings $\phi_T, \phi_{T'}$ must be close in a proper measure); this is due to the type of interpolators considered in [12], which are adapted to VEM and therefore require to compute averaged values around each vertex. Such a requirement of similar isotropy for neighbouring elements is absent from Definition 1, which is geared towards methods—such as HHO—whose interpolators are L^2-projections on cell and face polynomials; as a consequence, this definition allows for example for meshes with layers of very thin rectangles neighbouring layers of squares.

In the rest of the paper, we consider a regular skewed mesh sequence $(\mathcal{M}_h, \phi_h)_{h\in\mathcal{H}}$ with parameter ϱ, and we write $a \lesssim b$ if $a \le Cb$ with $C > 0$ depending only on Ω and ϱ and, when the inequality involves H^s seminorms, also on the exponent s. We write $a \approx b$ if $a \lesssim b$ and $b \lesssim a$. We also make the following assumption.

Assumption 1 (*Piecewise constant diffusion tensor*) For all $h \in \mathcal{H}$, the diffusion tensor $\boldsymbol{K}$ is piecewise constant on $\mathcal{T}_h$. For any $T \in \mathcal{T}_h$ we set $\boldsymbol{K}_T = \boldsymbol{K}_{|T}$.

Let T be an element of one of the meshes $\mathcal{M}_h$. If $\boldsymbol{x} \in T$ we set $\widehat{\boldsymbol{x}} = \phi_T(\boldsymbol{x}) \in \widehat{T}$. The gradient (resp. differential) with respect to $\widehat{\boldsymbol{x}}$ is denoted by $\widehat{\nabla}$ (resp. $\widehat{D}$). For $w \in L^2(T)$, the transport $\widehat{w} \in L^2(\widehat{T})$ of w on $\widehat{T}$ is $\widehat{w}(\widehat{\boldsymbol{x}}) = w(\boldsymbol{x}) = w(\phi_T^{-1}(\widehat{\boldsymbol{x}}))$. We also set $J\phi_T = |\det \phi_T|$, and define $J\phi_{T|F}$ as the absolute value of the determinant of the restriction $\phi_{T|F} : H_F \to H_{\widehat{F}}$, where H_X denotes the hyperplane generated by $X = F$ or $\widehat{F}$; $J\phi_{T|F}$ can be computed using any pair of orthonormal bases in F and $\widehat{F}$. Letting ϕ_T^t be the transpose of ϕ_T, the relevant diffusion tensor on $\widehat{T}$ is:

$$\boldsymbol{K}_{\phi,\widehat{T}} = \phi_T \boldsymbol{K}_T \phi_T^t. \tag{2}$$

The maximal and minimal eigenvalues of $\boldsymbol{K}_{\phi,\widehat{T}}$ are denoted by $\overline{K}_{\phi,\widehat{T}}$ and $\underline{K}_{\phi,\widehat{T}}$.

Lemma 1 (Transport relations)

1. Geometrical properties. *It holds* $\mathrm{N}_1(\phi_T^{-1}) \le \varrho^{-3}$ *and, for all* $F \in \mathcal{F}_T$,

$$\phi_T^t \boldsymbol{n}_{\widehat{T}\widehat{F}} = \frac{J\phi_T}{J\phi_{T|F}} \boldsymbol{n}_{TF}. \tag{3}$$

2. Transport of L^2-inner products and norms. *For all* w, z *in* $L^2(T)$ *or* $L^2(T)^d$,

$$(w, z)_T = J\phi_T^{-1} (\widehat{w}, \widehat{z})_{\widehat{T}} \quad \textit{and} \quad \|w\|_T = J\phi_T^{-1/2} \|\widehat{w}\|_{\widehat{T}}. \tag{4}$$

For all $F \in \mathcal{F}_T$ *and* $w, z \in L^2(F)$,

$$(w, z)_F = J\phi_{T|F}^{-1} (\widehat{w}, \widehat{z})_{\widehat{F}} \quad \textit{and} \quad \|w\|_F = J\phi_{T|F}^{-1/2} \|\widehat{w}\|_{\widehat{F}}. \tag{5}$$

3. Transport of derivatives. *For all $s \in \mathbb{N}$, $w \in H^s(T)$, $\boldsymbol{x} \in T$, it holds*

$$\mathrm{N}_s(\widehat{D}^s\widehat{w}(\widehat{\boldsymbol{x}})) \lesssim \mathrm{N}_s(\widehat{D^s w}(\widehat{\boldsymbol{x}})). \tag{6}$$

For all $w, z \in H^1(T)$,

$$\widehat{\nabla w}(\widehat{\boldsymbol{x}}) = \nabla w(\boldsymbol{x}) = \phi_T^t \widehat{\nabla}\widehat{w}(\widehat{\boldsymbol{x}}) \qquad \forall \boldsymbol{x} \in T, \tag{7}$$

$$\begin{aligned}(\boldsymbol{K}_T\nabla w, \nabla z)_T &= J\phi_T^{-1}(\boldsymbol{K}_{\phi,\widehat{T}}\widehat{\nabla}\widehat{w}, \widehat{\nabla}\widehat{z})_{\widehat{T}},\\ \|\boldsymbol{K}_T^{1/2}\nabla w\|_T &= J\phi_T^{-1/2}\|\boldsymbol{K}_{\phi,\widehat{T}}^{1/2}\widehat{\nabla}\widehat{w}\|_{\widehat{T}}.\end{aligned} \tag{8}$$

Proof 1. We have $\phi_T^{-1}(\widehat{T}) = T$. Since $\widehat{T}$ contains a ball of radius $\varrho^2 h_{\widehat{T}}$ (Point 3 in Definition 1) and T has diameter $h_T \le \varrho^{-1}h_{\widehat{T}}$, we see that ϕ_T^{-1} maps a ball of radius $\varrho^2 h_{\widehat{T}}$ into a ball of radius $\varrho^{-1}h_{\widehat{T}}$. Hence, $\mathrm{N}_1(\phi_T^{-1}) \le (\varrho^{-1}h_{\widehat{T}})/(\varrho^2 h_{\widehat{T}}) = \varrho^{-3}$.

Select two orthonormal bases $\mathcal{B} = (\mathcal{B}_F, \boldsymbol{n}_{TF})$ and $\widehat{\mathcal{B}} = (\widehat{\mathcal{B}_F}, \boldsymbol{n}_{\widehat{T}\widehat{F}})$ of $\mathbb{R}^d$, where $\mathcal{B}_F$ is a basis of H_F and $\widehat{\mathcal{B}_F}$ is a basis of $H_{\widehat{F}}$. Since $\phi_T(F) = \widehat{F}$, the matrix of ϕ_T in $(\mathcal{B}, \widehat{\mathcal{B}})$ is written

$$\begin{bmatrix} A & * \\ 0 & \lambda \end{bmatrix},$$

where A is the matrix of $\phi_{T|F}$ in $(\mathcal{B}_F, \widehat{\mathcal{B}_F})$. In particular, $J\phi_T = |\det A|\lambda = J\phi_{T|F}\lambda$ and thus $\lambda = J\phi_T/J\phi_{T|F}$. Transposing the matrix above gives the matrix of ϕ_T^t in the orthonormal bases $(\widehat{\mathcal{B}}, \mathcal{B})$. Since the last vector of $\widehat{\mathcal{B}}$ (resp. $\mathcal{B}$) is $\boldsymbol{n}_{\widehat{T}\widehat{F}}$ (resp. $\boldsymbol{n}_{TF}$), reading the last column of this transposed matrix gives $\phi_T^t\boldsymbol{n}_{\widehat{T}\widehat{F}} = \lambda\boldsymbol{n}_{TF}$ and proves (3).

2. Simple changes of variables (in T or F) establish (4) and (5).

3. Since $\widehat{w}(\widehat{\boldsymbol{x}}) = w(\phi_T^{-1}(\widehat{\boldsymbol{x}}))$, an induction on s shows that, for all $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_s \in \mathbb{R}^d$,

$$\begin{aligned}\widehat{D}^s\widehat{w}(\widehat{\boldsymbol{x}})(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_s) &= D^s w(\phi_T^{-1}(\widehat{\boldsymbol{x}}))(\phi_T^{-1}(\boldsymbol{\xi}_1), \cdots, \phi_T^{-1}(\boldsymbol{\xi}_s))\\ &= \widehat{D^s w}(\widehat{\boldsymbol{x}})(\phi_T^{-1}(\boldsymbol{\xi}_1), \cdots, \phi_T^{-1}(\boldsymbol{\xi}_s)).\end{aligned} \tag{9}$$

We infer that $|\widehat{D}^s\widehat{w}(\widehat{\boldsymbol{x}})(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_s)| \le \mathrm{N}_s(\widehat{D^s w}(\widehat{\boldsymbol{x}}))\,|\phi_T^{-1}(\boldsymbol{\xi}_1)| \cdots |\phi_T^{-1}(\boldsymbol{\xi}_s)|$. By Point 1, $|\phi_T^{-1}(\boldsymbol{\xi}_i)| \lesssim |\boldsymbol{\xi}_i|$ for all $i = 1, \ldots, s$, and the proof of (6) is complete.

The relation (7) is obtained transposing (9) for $s = 1$. The second relation in (8) follows from the first with $z = w$. To prove this first relation, apply (4) to $\boldsymbol{K}_T\nabla w$ and ∇z instead of w and z, use the fact that $\boldsymbol{K}_T$ is constant and invoke (7) to write

$$(\boldsymbol{K}_T\nabla w, \nabla z)_T = J\phi_T^{-1}(\boldsymbol{K}_T\widehat{\nabla w}, \widehat{\nabla z})_{\widehat{T}} = J\phi_T^{-1}(\phi_T\boldsymbol{K}_T\phi_T^t\widehat{\nabla}\widehat{w}, \widehat{\nabla}\widehat{z})_{\widehat{T}}.$$

□

3 Oblique Elliptic Projector on Skewed Elements

Here, T is a generic element of $\mathcal{M}_h$. Fix a polynomial degree $\ell \geq 0$ and recall the definition in [7, Sect. 3.2.1] of the oblique elliptic projector $\pi_{\boldsymbol{K},T}^{1,\ell} : H^1(T) \to \mathbb{P}^\ell(T)$: for all $v \in H^1(T)$,

$$(\boldsymbol{K}_T \nabla \pi_{\boldsymbol{K},T}^{1,\ell} v, \nabla w)_T = (\boldsymbol{K}_T \nabla v, \nabla w)_T \quad \forall w \in \mathbb{P}^\ell(T), \tag{10}$$

$$(\pi_{\boldsymbol{K},T}^{1,\ell} v, 1)_T = (v, 1)_T. \tag{11}$$

The approximation properties of the oblique elliptic projector form an essential component of the analysis of HHO schemes for (1). To establish these approximation properties, let us first describe how the elliptic projector is transported through ϕ_T.

Lemma 2 (Transport of the elliptic projector) *Letting $\pi_{\boldsymbol{K},\phi,\widehat{T}}^{1,\ell}$ be the oblique elliptic projector on $\widehat{T}$ for the tensor $\boldsymbol{K}_{\phi,\widehat{T}}$ defined by (2), it holds*

$$\widehat{\pi_{\boldsymbol{K},T}^{1,\ell} v} = \pi_{\boldsymbol{K},\phi,\widehat{T}}^{1,\ell} \widehat{v} \qquad \forall v \in H^1(T). \tag{12}$$

Proof Take $w \in \mathbb{P}^\ell(T)$ and write, using the definition (10) of $\pi_{\boldsymbol{K},T}^{1,\ell}$, the transport relation (8) applied to (v, w) instead of (w, z), and the definition of $\pi_{\boldsymbol{K},\phi,\widehat{T}}^{1,\ell}$ together with $\widehat{w} \in \mathbb{P}^\ell(\widehat{T})$,

$$\begin{aligned}(\boldsymbol{K}_T \nabla \pi_{\boldsymbol{K},T}^{1,\ell} v, \nabla w)_T = (\boldsymbol{K}_T \nabla v, \nabla w)_T &= J\phi_T^{-1} (\boldsymbol{K}_{\phi,\widehat{T}} \widehat{\nabla} \widehat{v}, \widehat{\nabla} \widehat{w})_{\widehat{T}} \\ &= J\phi_T^{-1} (\boldsymbol{K}_{\phi,\widehat{T}} \widehat{\nabla} \pi_{\boldsymbol{K},\phi,\widehat{T}}^{1,\ell} \widehat{v}, \widehat{\nabla} \widehat{w})_{\widehat{T}}.\end{aligned} \tag{13}$$

On the other hand, (8) applied to $(\pi_{\boldsymbol{K},T}^{1,\ell} v, w)$ instead of (w, z) gives

$$(\boldsymbol{K}_T \nabla \pi_{\boldsymbol{K},T}^{1,\ell} v, \nabla w)_T = J\phi_T^{-1} (\boldsymbol{K}_{\phi,\widehat{T}} \widehat{\nabla} \widehat{\pi_{\boldsymbol{K},T}^{1,\ell} v}, \widehat{\nabla} \widehat{w})_{\widehat{T}}.$$

Combining this relation with (13) and using the fact that $\widehat{w}$ is arbitrary in $\mathbb{P}^\ell(\widehat{T})$ yields $\widehat{\nabla} \pi_{\boldsymbol{K},\phi,\widehat{T}}^{1,\ell} \widehat{v} = \widehat{\nabla} \widehat{\pi_{\boldsymbol{K},T}^{1,\ell} v}$. To prove (12) it remains to show that $\pi_{\boldsymbol{K},\phi,\widehat{T}}^{1,\ell} \widehat{v}$ and $\widehat{\pi_{\boldsymbol{K},T}^{1,\ell} v}$ have the same average on $\widehat{T}$. This is done by using (4) and (11) (for both $\pi_{\boldsymbol{K},T}^{1,\ell}$ and $\pi_{\boldsymbol{K},\phi,\widehat{T}}^{1,\ell}$) to write $(\pi_{\boldsymbol{K},\phi,\widehat{T}}^{1,\ell} \widehat{v}, 1)_{\widehat{T}} = (\widehat{v}, 1)_{\widehat{T}} = J\phi_T (v, 1)_T = J\phi_T (\pi_{\boldsymbol{K},T}^{1,\ell} v, 1)_T = (\widehat{\pi_{\boldsymbol{K},T}^{1,\ell} v}, 1)_{\widehat{T}}$. □

Let $|\cdot|_n$ be the n-dimensional Lebesgue measure. The following characteristic lengths will be used to state boundary approximation properties of $\pi_{\boldsymbol{K},T}^{1,\ell}$:

$$d_{TF} = \frac{|T|_d}{|F|_{d-1}} \qquad \forall F \in \mathcal{F}_T. \tag{14}$$

Using $J\phi_T|T|_d = |\widehat{T}|_d$, $J\phi_{T|F}|F|_{d-1} = |\widehat{F}|_{d-1}$, and $|\widehat{T}|_d \approx h_{\widehat{F}}|\widehat{F}|_{d-1}$ and $h_{\widehat{F}} \approx h_{\widehat{T}}$ (owing to the isotropy of $\widehat{T}$ and to [7, Lemma 1.12]), we see that

$$d_{TF} \approx \frac{J\phi_{T|F}}{J\phi_T} h_{\widehat{F}} \approx \frac{J\phi_{T|F}}{J\phi_T} h_{\widehat{T}}. \tag{15}$$

Proposition 1 (Approximation properties of the elliptic projector on skewed elements) *For all $s \in \{1, \ldots, \ell + 1\}$ and all $v \in H^s(T)$,*

$$\|\boldsymbol{K}_T^{1/2} \nabla(v - \pi_{\boldsymbol{K},T}^{1,\ell} v)\|_T \lesssim \overline{K}_{\phi,\widehat{T}}^{1/2} h_T^{s-1} |v|_{H^s(T)} \tag{16}$$

and, if $s \geq 2$, for all $F \in \mathcal{F}_T$,

$$d_{TF}^{1/2} \|\boldsymbol{K}_T^{1/2} \nabla(v - \pi_{\boldsymbol{K},T}^{1,\ell} v)\|_F \lesssim \overline{K}_{\phi,\widehat{T}}^{1/2} h_T^{s-1} |v|_{H^s(T)}. \tag{17}$$

Proof Since $\widehat{T}$ satisfies Points 2 and 3 in Definition 1, [7, Theorem 3.3] yields

$$\|\boldsymbol{K}_{\phi,\widehat{T}}^{1/2} \widehat{\nabla}(\widehat{v} - \pi_{\boldsymbol{K},\phi,\widehat{T}}^{1,\ell} \widehat{v})\|_{\widehat{T}} \lesssim \overline{K}_{\phi,\widehat{T}}^{1/2} h_{\widehat{T}}^{s-1} |\widehat{v}|_{H^s(\widehat{T})}, \tag{18}$$

$$h_{\widehat{T}}^{1/2} \|\boldsymbol{K}_{\phi,\widehat{T}}^{1/2} \widehat{\nabla}(\widehat{v} - \pi_{\boldsymbol{K},\phi,\widehat{T}}^{1,\ell} \widehat{v})\|_{\widehat{F}} \lesssim \overline{K}_{\phi,\widehat{T}}^{1/2} h_{\widehat{T}}^{s-1} |\widehat{v}|_{H^s(\widehat{T})} \quad \forall \widehat{F} \in \mathcal{F}_{\widehat{T}} \quad (\text{if } s \geq 2). \tag{19}$$

The volumetric (16) and trace (17) estimates are obtained transporting these estimates with (12). We start with the volumetric estimate. Using (12) and (8) we have

$$\|\boldsymbol{K}_{\phi,\widehat{T}}^{1/2} \widehat{\nabla}(\widehat{v} - \pi_{\boldsymbol{K},\phi,\widehat{T}}^{1,\ell} \widehat{v})\|_{\widehat{T}} = \|\boldsymbol{K}_{\phi,\widehat{T}}^{1/2} \widehat{\nabla}(\widehat{v - \pi_{\boldsymbol{K},T}^{1,\ell} v})\|_{\widehat{T}} = J\phi_T^{1/2} \|\boldsymbol{K}_T^{1/2} \nabla(v - \pi_{\boldsymbol{K},T}^{1,\ell} v)\|_T.$$

Hence, applying (18) and using the estimate $h_{\widehat{T}} \lesssim h_T$ (see Point 1 in Definition 1),

$$\|\boldsymbol{K}_T^{1/2} \nabla(v - \pi_{\boldsymbol{K},T}^{1,\ell} v)\|_T \lesssim J\phi_T^{-1/2} \overline{K}_{\phi,\widehat{T}}^{1/2} h_T^{s-1} |\widehat{v}|_{H^s(\widehat{T})}. \tag{20}$$

By the definition of the H^s-seminorm, the relation (6) and the transport (4) give

$$|\widehat{v}|_{H^s(\widehat{T})} \lesssim \|\mathrm{N}_s(\widehat{D^s v})\|_{\widehat{T}} \lesssim J\phi_T^{1/2} \|\mathrm{N}_s(D^s v)\|_T = J\phi_T^{1/2} |v|_{H^s(T)}. \tag{21}$$

Plugged into (20), this concludes the proof of (16). We now turn to (17). The transport relations (12), (7) and (5) together with the definition (2) of $\boldsymbol{K}_{\phi,\widehat{T}}$ yield

$$\|\boldsymbol{K}_{\phi,\widehat{T}}^{1/2} \widehat{\nabla}(\widehat{v} - \pi_{\boldsymbol{K},\phi,\widehat{T}}^{1,\ell} \widehat{v})\|_{\widehat{F}} = \|\boldsymbol{K}_{\phi,\widehat{T}}^{1/2} \widehat{\nabla}(\widehat{v - \pi_{\boldsymbol{K},T}^{1,\ell} v})\|_{\widehat{F}} = J\phi_{T|F}^{1/2} \|\boldsymbol{K}_T^{1/2} \nabla(v - \pi_{\boldsymbol{K},T}^{1,\ell} v)\|_F.$$

Estimate (17) follows plugging this relation into (19), using (21) and recalling (15) and that $h_{\widehat{T}} \lesssim h_T$. □

Remark 2 (Optimality of the approximation properties) This proof shows that (16) and (17) come from the corresponding inequalities (18) and (19) for isotropic elements, and from (21), itself derived from (6). The latter inequality is optimal in the sense that, for any ϕ_T, there are functions w for which it is an equality. Hence, the approximation properties (16) and (17) for skewed elements are as optimal as the corresponding approximation properties for isotropic elements.

4 Analysis of HHO Schemes on Skewed Meshes

We briefly recall the construction of HHO schemes for (1) (referring to [7, Chap. 3.1] for a comprehensive presentation), and establish key properties for proving error estimates on skewed meshes. In the following, $k \geq 0$ is a fixed polynomial degree.

4.1 Local Space and Potential Reconstruction

For $T \in \mathcal{T}_h$, the local space of unknowns is

$$\underline{U}_T^k := \{\underline{v}_T = (v_T, (v_F)_{F\in\mathcal{F}_T}) \,:\, v_T \in \mathbb{P}^k(T)\,,\ v_F \in \mathbb{P}^k(F)\quad \forall F \in \mathcal{F}_h\}.$$

Setting $K_{TF} = \boldsymbol{K}_T \boldsymbol{n}_{TF} \cdot \boldsymbol{n}_{TF}$, this space is endowed with the seminorm

$$\|\underline{v}_T\|_{1,\boldsymbol{K},T} := \left(\|\boldsymbol{K}_T^{1/2}\nabla v_T\|_T^2 + \sum_{F\in\mathcal{F}_T} \frac{K_{TF}}{d_{TF}} \|v_F - v_T\|_F^2 \right)^{1/2} \qquad \forall \underline{v}_T \in \underline{U}_T^k. \tag{22}$$

For isotropic elements, this norm is usually defined using h_F instead of d_{TF}, see [7, Sect. 3.1.3.2]. The choice made in (22) ensures, for skewed elements, optimal estimates in terms of ϕ_T. The potential reconstruction $\mathrm{p}_{\boldsymbol{K},T}^{k+1} : \underline{U}_T^k \to \mathbb{P}^{k+1}(T)$ is such that, for all $\underline{v}_T \in \underline{U}_T^k$ and $w \in \mathbb{P}^{k+1}(T)$,

$$(\boldsymbol{K}_T \nabla \mathrm{p}_{\boldsymbol{K},T}^{k+1}, \nabla w)_T = (\boldsymbol{K}_T \nabla v_T, \nabla w)_T + \sum_{F\in\mathcal{F}_T} (v_F - v_T, \boldsymbol{K}_T \nabla w \cdot \boldsymbol{n}_{TF})_F, \tag{23}$$

$$(\mathrm{p}_{\boldsymbol{K},T}^{k+1}\underline{v}_T, 1)_T = (v_T, 1)_T. \tag{24}$$

Lemma 3 (Transport of potential reconstruction) It holds

$$\widehat{\mathrm{p}_{\boldsymbol{K},T}^{k+1}\underline{v}_T} = \mathrm{p}_{\boldsymbol{K},\phi,\widehat{T}}^{k+1}\widehat{\underline{v}_T} \qquad \forall \underline{v}_T \in \underline{U}_T^k, \tag{25}$$

where $\widehat{\underline{v}_T} = (\widehat{v_T}, (\widehat{v_F})_{F\in\mathcal{F}_T}) \in \underline{U}_{\widehat{T}}^k$ is the transported $\underline{v}_T$, and $\mathrm{p}_{\boldsymbol{K},\phi,\widehat{T}}^{k+1}$ is the potential reconstruction on $\widehat{T}$ for the diffusion tensor $\boldsymbol{K}_{\phi,\widehat{T}}$.

Proof For all $w \in \mathbb{P}^{k+1}(T)$,

$$\begin{aligned}
(\boldsymbol{K}_{\phi,\widehat{T}}\widehat{\nabla}\widehat{\mathrm{p}^{k+1}_{\boldsymbol{K},T}\underline{v}_T}, \widehat{\nabla}\widehat{w})_{\widehat{T}} &= J\phi_T(\boldsymbol{K}_T\nabla \mathrm{p}^{k+1}_{\boldsymbol{K},T}\underline{v}_T, \nabla w)_T\\
&= J\phi_T(\boldsymbol{K}_T\nabla v_T, \nabla w)_T + J\phi_T\sum_{F\in\mathcal{F}_T}(v_F - v_T, \boldsymbol{K}_T\nabla w\cdot\boldsymbol{n}_{TF})_F\\
&= (\boldsymbol{K}_{\phi,\widehat{T}}\widehat{\nabla}\widehat{v_T}, \widehat{\nabla}\widehat{w})_{\widehat{T}} + J\phi_T\sum_{F\in\mathcal{F}_T}J\phi_{T|F}^{-1}(\widehat{v_F}-\widehat{v_T}, \boldsymbol{K}_T\widehat{\nabla w}\cdot\boldsymbol{n}_{TF})_{\widehat{F}},
\end{aligned}\tag{26}$$

where we have used in this order the transport relation (8), the definition (23) of $\mathrm{p}^{k+1}_{\boldsymbol{K},T}$, the transport relations (8) and (5), and the fact that $\boldsymbol{K}_T$ and $\boldsymbol{n}_{TF}$ are constant. Invoking (7), (3) and (2), we have

$$\boldsymbol{K}_T\widehat{\nabla w}\cdot\boldsymbol{n}_{TF} = \boldsymbol{K}_{\phi,\widehat{T}}\widehat{\nabla}\widehat{w}\cdot(\phi_T^{-1})^t\boldsymbol{n}_{TF} = \frac{J\phi_{T|F}}{J\phi_T}\boldsymbol{K}_{\phi,\widehat{T}}\widehat{\nabla}\widehat{w}\cdot\boldsymbol{n}_{\widehat{T}\widehat{F}}$$

and (26) gives

$$\begin{aligned}
(\boldsymbol{K}_{\phi,\widehat{T}}\widehat{\nabla}\widehat{\mathrm{p}^{k+1}_{\boldsymbol{K},T}\underline{v}_T}, \widehat{\nabla}\widehat{w})_{\widehat{T}} &= (\boldsymbol{K}_{\phi,\widehat{T}}\widehat{\nabla}\widehat{v_T}, \widehat{\nabla}\widehat{w})_{\widehat{T}} + \sum_{F\in\mathcal{F}_T}(\widehat{v_F}-\widehat{v_T}, \boldsymbol{K}_{\phi,\widehat{T}}\widehat{\nabla}\widehat{w}\cdot\boldsymbol{n}_{\widehat{T}\widehat{F}})_{\widehat{F}}\\
&= (\boldsymbol{K}_{\phi,\widehat{T}}\widehat{\nabla}\mathrm{p}^{k+1}_{\boldsymbol{K},\phi,\widehat{T}}\widehat{\underline{v}_T}, \widehat{\nabla}\widehat{w})_{\widehat{T}},
\end{aligned}$$

the conclusion following from the definition of $\mathrm{p}^{k+1}_{\boldsymbol{K},\phi,\widehat{T}}$. Since $\widehat{w}$ is arbitrary in $\mathbb{P}^{k+1}(\widehat{T})$, this proves that $\widehat{\mathrm{p}^{k+1}_{\boldsymbol{K},T}\underline{v}_T}$ and $\mathrm{p}^{k+1}_{\boldsymbol{K},\phi,\widehat{T}}\widehat{\underline{v}_T}$ have the same gradient. Using (4) and (24) we also see that they have same average on $\widehat{T}$, which concludes the proof of (25). □

4.2 Local Bilinear Form

The difference operators $\delta^k_{\boldsymbol{K},T} : \underline{U}^k_T \to \mathbb{P}^k(T)$ and, for $F\in\mathcal{F}_T$, $\delta^k_{\boldsymbol{K},TF} : \underline{U}^k_T \to \mathbb{P}^k(F)$ are defined by: for all $\underline{v}_T \in \underline{U}^k_T$,

$$\delta^k_{\boldsymbol{K},T}\underline{v}_T := \pi^{0,k}_T(\mathrm{p}^{k+1}_{\boldsymbol{K},T}\underline{v}_T - v_T)\,,\quad \delta^k_{\boldsymbol{K},TF}\underline{v}_T = \pi^{0,k}_F(\mathrm{p}^{k+1}_{\boldsymbol{K},T}\underline{v}_T - v_F)\qquad \forall F\in\mathcal{F}_T,\tag{27}$$

where, for $X = T$ or F, $\pi^{0,k}_X : L^2(X)\to\mathbb{P}^k(X)$ is the $L^2(X)$-orthogonal projection. We note that, for any $w\in L^2(X)$,

$$\widehat{\pi^{0,k}_X w} = \pi^{0,k}_{\widehat{X}}\widehat{w}.\tag{28}$$

The local stabilisation bilinear form is given by: for all $\underline{u}_T, \underline{v}_T \in \underline{U}^k_T$,

$$\mathrm{s}_{\boldsymbol{K},T}(\underline{u}_T,\underline{v}_T) := \sum_{F\in\mathcal{F}_T}\frac{K_{TF}}{d_{TF}}(\delta^k_{\boldsymbol{K},TF}\underline{u}_T-\delta^k_{\boldsymbol{K},T}\underline{u}_T,\delta^k_{\boldsymbol{K},TF}\underline{v}_T-\delta^k_{\boldsymbol{K},T}\underline{v}_T)_F. \tag{29}$$

The local HHO bilinear form $\mathrm{a}_{\boldsymbol{K},T}:\underline{U}_T^k\times\underline{U}_T^k\to\mathbb{R}$ is then defined by:

$$\mathrm{a}_{\boldsymbol{K},T}(\underline{u}_T,\underline{v}_T) := (\boldsymbol{K}_T\nabla\mathrm{p}^{k+1}_{\boldsymbol{K},T}\underline{u}_T,\nabla\mathrm{p}^{k+1}_{\boldsymbol{K},T}\underline{v}_T)_T+\mathrm{s}_{\boldsymbol{K},T}(\underline{u}_T,\underline{v}_T)\quad\forall\underline{u}_T,\underline{v}_T\in\underline{U}_T^k. \tag{30}$$

In the right-hand side above, the first term is responsible for the consistency of the bilinear form, while the addition of the second term ensures the stability and boundedness property stated in the following proposition. Other choices of $\mathrm{s}_{\boldsymbol{K},T}$ are possible [7, Assumption 3.9], and, on isotropic meshes, the factor d_{TF} in this stabilisation bilinear form can be replaced by h_F.

Proposition 2 (Stability and boundedness of $\mathrm{a}_{\boldsymbol{K},T}$) *It holds*

$$\mathrm{a}_{\boldsymbol{K},T}(\underline{v}_T,\underline{v}_T)\approx\|\underline{v}_T\|^2_{1,\boldsymbol{K},T}\qquad\forall\underline{v}_T\in\underline{U}_T^k. \tag{31}$$

Proof *Step 1: transport of seminorms*. Let $\|\cdot\|_{1,\boldsymbol{K},\phi,\widehat{T}}$ be defined on $\underline{U}^k_{\widehat{T}}$ by:

$$\|\widehat{\underline{v}_T}\|^2_{1,\boldsymbol{K},\phi,\widehat{T}} := \|\boldsymbol{K}^{1/2}_{\phi,\widehat{T}}\widehat{\nabla}\widehat{v_T}\|^2_{\widehat{T}}+\sum_{\widehat{F}\in\mathcal{F}_{\widehat{T}}}\frac{K_{\phi,\widehat{T}\widehat{F}}}{h_{\widehat{F}}}\|\widehat{v_F}-\widehat{v_T}\|^2_{\widehat{F}}\qquad\forall\widehat{\underline{v}_T}\in\underline{U}^k_{\widehat{T}},$$

where $K_{\phi,\widehat{T}\widehat{F}} := \boldsymbol{K}_{\phi,\widehat{T}}\boldsymbol{n}_{\widehat{T}\widehat{F}}\cdot\boldsymbol{n}_{\widehat{T}\widehat{F}}$. If $\underline{v}_T\in\underline{U}_T^k$, the transport relations (8) and (5) yield

$$\|\underline{v}_T\|^2_{1,\boldsymbol{K},T} = J\phi_T^{-1}\|\boldsymbol{K}^{1/2}_{\phi,\widehat{T}}\widehat{\nabla}\widehat{v_T}\|^2_{\widehat{T}}+\sum_{F\in\mathcal{F}_T}\frac{K_{TF}}{d_{TF}}J\phi^{-1}_{T|F}\|\widehat{v_F}-\widehat{v_T}\|^2_{\widehat{F}}. \tag{32}$$

Starting from $K_{TF}=\boldsymbol{K}_T\boldsymbol{n}_{TF}\cdot\boldsymbol{n}_{TF}$, the relations (3), (15) and (2) yield

$$\frac{K_{TF}}{d_{TF}}J\phi^{-1}_{T|F}=\frac{\boldsymbol{K}_T\frac{J\phi_{T|F}}{J\phi_T}\phi_T^t\boldsymbol{n}_{\widehat{T}\widehat{F}}\cdot\frac{J\phi_{T|F}}{J\phi_T}\phi_T^t\boldsymbol{n}_{\widehat{T}\widehat{F}}}{d_{TF}J\phi_{T|F}}\approx J\phi_T^{-1}\frac{K_{\phi,\widehat{T}\widehat{F}}}{h_{\widehat{F}}}. \tag{33}$$

Plugged into (32), this gives

$$\|\underline{v}_T\|^2_{1,\boldsymbol{K},T}\approx J\phi_T^{-1}\|\widehat{\underline{v}_T}\|^2_{1,\boldsymbol{K},\phi,\widehat{T}}. \tag{34}$$

Step 2: transport of bilinear forms. Let $\mathrm{a}_{\boldsymbol{K},\phi,\widehat{T}}:\underline{U}^k_{\widehat{T}}\times\underline{U}^k_{\widehat{T}}\to\mathbb{R}$ be the standard local HHO bilinear form on $\widehat{T}$ for $\boldsymbol{K}_{\phi,\widehat{T}}$:

$$\mathrm{a}_{\boldsymbol{K},\phi,\widehat{T}}(\widehat{\underline{v}_T},\widehat{\underline{w}_T}) := (\boldsymbol{K}_{\phi,\widehat{T}}\widehat{\nabla}\mathrm{p}^{k+1}_{\boldsymbol{K},\phi,\widehat{T}}\widehat{\underline{v}_T},\widehat{\nabla}\mathrm{p}^{k+1}_{\boldsymbol{K},\phi,\widehat{T}}\widehat{\underline{w}_T})_{\widehat{T}}+\mathrm{s}_{\boldsymbol{K},\phi,\widehat{T}}(\widehat{\underline{v}_T},\widehat{\underline{w}_T}),\ \text{where}$$

$$\mathrm{s}_{\boldsymbol{K},\phi,\widehat{T}}(\widehat{\underline{v}_T},\widehat{\underline{w}_T}) := \sum_{\widehat{F}\in\mathcal{F}_{\widehat{T}}}\frac{K_{\phi,\widehat{T}\widehat{F}}}{h_{\widehat{F}}}(\delta^k_{\boldsymbol{K},\phi,\widehat{T}\widehat{F}}\widehat{\underline{v}_T}-\delta^k_{\boldsymbol{K},\phi,\widehat{T}}\widehat{\underline{v}_T},\delta^k_{\boldsymbol{K},\phi,\widehat{T}\widehat{F}}\widehat{\underline{w}_T}-\delta^k_{\boldsymbol{K},\phi,\widehat{T}}\widehat{\underline{w}_T})_{\widehat{F}}$$

with difference operators $\delta_{\boldsymbol{K},\phi,\widehat{T}}$ and $(\delta_{\boldsymbol{K},\phi,\widehat{T}\widehat{F}})_{\widehat{F}\in\mathcal{F}_{\widehat{T}}}$ defined on $\underline{U}^k_{\widehat{T}}$ in a similar way as in (27), using $\mathrm{p}^{k+1}_{\boldsymbol{K},\phi,\widehat{T}}$ instead of $\mathrm{p}^{k+1}_{\boldsymbol{K},T}$. Let $\underline{v}_T \in \underline{U}^k_T$. Relations (27), (25) and (28) show that $\widehat{\delta^k_{\boldsymbol{K},T}\underline{v}_T} = \delta_{\boldsymbol{K},\phi,\widehat{T}\widehat{F}}\widehat{\underline{v}_T}$ and $\widehat{\delta^k_{\boldsymbol{K},TF}\underline{v}_T} = \delta_{\boldsymbol{K},\phi,\widehat{T}\widehat{F}}\widehat{\underline{v}_T}$. Hence, by (5) and (33),

$$\mathrm{s}_{\boldsymbol{K},T}(\underline{v}_T, \underline{v}_T) \approx J\phi_T^{-1}\mathrm{s}_{\boldsymbol{K},\phi,\widehat{T}}(\widehat{\underline{v}_T}, \widehat{\underline{w}_T}) \tag{35}$$

and, recalling (8),

$$(\boldsymbol{K}_T\nabla\mathrm{p}^{k+1}_{\boldsymbol{K},T}\underline{u}_T, \nabla\mathrm{p}^{k+1}_{\boldsymbol{K},T}\underline{v}_T)_T = J\phi_T^{-1}(\boldsymbol{K}_{\phi,\widehat{T}}\widehat{\nabla}\mathrm{p}^{k+1}_{\boldsymbol{K},\phi,\widehat{T}}\widehat{\underline{v}_T}, \widehat{\nabla}\mathrm{p}^{k+1}_{\boldsymbol{K},\phi,\widehat{T}}\widehat{\underline{v}_T})_T.$$

This leads to

$$\mathrm{a}_{\boldsymbol{K},T}(\underline{v}_T, \underline{v}_T) \approx J\phi_T^{-1}\mathrm{a}_{\boldsymbol{K},\phi,\widehat{T}}(\widehat{\underline{v}_T}, \widehat{\underline{v}_T}). \tag{36}$$

Step 3: conclusion. Since $\widehat{T}$ is isotropic, [7, Proposition 3.13] yields $\mathrm{a}_{\boldsymbol{K},\phi,\widehat{T}}(\widehat{\underline{v}_T}, \widehat{\underline{v}_T}) \approx \|\widehat{\underline{v}_T}\|^2_{1,\boldsymbol{K},\phi,\widehat{T}}$. Using (34) and (36), the proof of (31) is complete. □

4.3 HHO Scheme and Error Estimate

The global discrete space of unknowns is obtained patching local spaces and enforcing homogeneous Dirichlet boundary conditions:

$$\begin{aligned}\underline{U}^k_{h,0} := \{\underline{v}_h = ((v_T)_{T\in\mathcal{T}_h}, (v_F)_{F\in\mathcal{F}_h}) \ : \ & v_T \in \mathbb{P}^k(T) \quad \forall T \in \mathcal{T}_h\,, \\ & v_F \in \mathbb{P}^k(F) \quad \forall F \in \mathcal{F}_h\,, \ v_F = 0 \quad \forall F \subset \partial\Omega\}.\end{aligned}$$

The restriction of $\underline{v}_h \in \underline{U}^k_{h,0}$ to an element T is $\underline{v}_T = (v_T, (v_F)_{F\in\mathcal{F}_T}) \in \underline{U}^k_T$. The interpolator $\underline{I}^k_h : H^1_0(\Omega) \to \underline{U}^k_{h,0}$ is such that, for $v \in H^1_0(\Omega)$,

$$\underline{I}^k_h v := ((\pi^{0,k}_T v)_{T\in\mathcal{T}_h}, (\pi^{0,k}_F v_{|F})_{F\in\mathcal{F}_h}).$$

The local interpolator on $T \in \mathcal{T}_h$ is $\underline{I}^k_T : H^1(T) \to \underline{U}^k_T$ such that, for $v \in H^1(T)$, $\underline{I}^k_T v = (\pi^{0,k}_T v, (\pi^{0,k}_F v_{|F})_{F\in\mathcal{F}_T})$. The global HHO bilinear form $\mathrm{a}_{\boldsymbol{K},h} : \underline{U}^k_{h,0} \times \underline{U}^k_{h,0} \to \mathbb{R}$ is assembled from local contributions: for $\underline{v}_h, \underline{w}_h \in \underline{U}^k_{h,0}$,

$$\mathrm{a}_{\boldsymbol{K},h}(\underline{u}_h, \underline{v}_h) := \sum_{T\in\mathcal{T}_h}\mathrm{a}_{\boldsymbol{K},T}(\underline{u}_T, \underline{v}_T).$$

This global bilinear form defines the energy norm such that, for $\underline{v}_h \in \underline{U}^k_{h,0}$,

$$\|\underline{v}_h\|_{\mathrm{a},\boldsymbol{K},h} := \mathrm{a}_{\boldsymbol{K},h}(\underline{v}_h, \underline{v}_h)^{1/2}. \tag{37}$$

The HHO scheme for (1) is written: find $\underline{u}_h \in \underline{U}_{h,0}^k$ such that

$$\mathrm{a}_{\boldsymbol{K},h}(\underline{u}_h, \underline{v}_h) = \sum_{T\in\mathcal{T}_h} (f, v_T)_T \qquad \forall \underline{v}_h \in \underline{U}_{h,0}^k. \tag{38}$$

This scheme is well-posed, and is a Finite Volume scheme in the sense that it has a flux formulation [7, Lemma 3.17]. Our main result is the following theorem.

Theorem 2 (Discrete energy error estimate for HHO schemes on skewed meshes) *Assume that the weak solution $u \in H_0^1(\Omega)$ to (1) is such that, for some $r \in \{0, \dots, k\}$, $u_{|T} \in H^{r+2}(T)$ for all $T \in \mathcal{T}_h$. Let $\underline{u}_h \in \underline{U}_{h,0}^k$ be the solution to the HHO scheme (38). Then, it holds*

$$\|\underline{I}_h^k u - \underline{u}_h\|_{\mathrm{a},\boldsymbol{K},h} \lesssim \left(\sum_{T\in\mathcal{T}_h} \overline{K}_{\phi,\widehat{T}} \alpha_{\boldsymbol{K},\phi,\widehat{T}} h_T^{2(r+1)} |u|^2_{H^{r+2}(T)} \right)^{1/2}, \tag{39}$$

where $\alpha_{\boldsymbol{K},\phi,\widehat{T}}$ is the anisotropy ratio of $\boldsymbol{K}_{\phi,\widehat{T}}$, defined by $\alpha_{\boldsymbol{K},\phi,\widehat{T}} := \frac{\overline{K}_{\phi,\widehat{T}}}{\underline{K}_{\phi,\widehat{T}}}$.

Remark 3 (Optimality of the error estimate) Following Remark 2, Estimate (39) is as optimal with respect to the mesh skewness as the corresponding estimate [7, Theorem 3.18], for isotropic meshes, is optimal with respect to the diffusion tensor.

Remark 4 (Interplay between mesh skewness and diffusion anisotropy) Assume for simplicity that $d = 2$ and that, for any $T \in \mathcal{T}_h$, there is an orthonormal basis in which

$$\boldsymbol{K}_T = \begin{bmatrix} \lambda_T & 0 \\ 0 & 1 \end{bmatrix} \quad \text{and} \quad \phi_T = \begin{bmatrix} a_T & 0 \\ 0 & b_T \end{bmatrix}. \tag{40}$$

Then $\boldsymbol{K}_{\phi,\widehat{T}}$ is diagonal with coefficients $a_T^2\lambda_T$ and b_T^2, and (39) leads to the estimate

$$\|\underline{I}_h^k u - \underline{u}_h\|_{\mathrm{a},\boldsymbol{K},h} \lesssim \max_{T\in\mathcal{T}_h} \left[\max(a_T\lambda_T^{1/2}, b_T) \max\left(\frac{a_T\lambda_T^{1/2}}{b_T}, \frac{b_T}{a_T\lambda_T^{1/2}} \right) \right] h^{r+1} |u|_{H^{r+2}(\mathcal{T}_h)}, \tag{41}$$

where $|u|_{H^{r+2}(\mathcal{T}_h)}$ is the usual broken H^{r+2}-seminorm of u. The first term in the right-hand side of (41) encodes the interaction between the skewness of the mesh elements and the local anisotropy of the diffusion tensor.

Proof (Theorem 2) Applying the 3rd Strang lemma [4], we have

$$\|\underline{I}_h^k u - \underline{u}_h\|_{\mathrm{a},\boldsymbol{K},h} \leq \sup_{\underline{v}_h \in \underline{U}_{h,0}^k,\ \|\underline{v}_h\|_{\mathrm{a},\boldsymbol{K},h} \leq 1} \mathcal{E}_{\boldsymbol{K},h}(u; \underline{v}_h), \tag{42}$$

where $\mathcal{E}_{\boldsymbol{K},h}(u; \underline{v}_h) := \sum_{T\in\mathcal{T}_h} (f, v_T)_T - \mathrm{a}_{\boldsymbol{K},h}(\underline{I}_h^k u, \underline{v}_h)$. The following relation is established in the proof of [7, Lemma 3.15]:

$$
\begin{aligned}
\mathcal{E}_{\boldsymbol{K},h}(u;\underline{v}_h) = {} & \sum_{T\in\mathcal{T}_h}\sum_{F\in\mathcal{F}_T} (\boldsymbol{K}_T\nabla(u-\pi^{1,k+1}_{\boldsymbol{K},T}u)\cdot\boldsymbol{n}_{TF}, v_F - v_T)_F \\
& - \sum_{T\in\mathcal{T}_h} \mathrm{s}_{\boldsymbol{K},T}(\underline{I}^k_T u, \underline{v}_T) = \mathfrak{T}_1+\mathfrak{T}_2.
\end{aligned}
\tag{43}
$$

Let $\underline{v}_h \in \underline{U}^k_{h,0}$ be such that $\|\underline{v}_h\|_{\mathrm{a},\boldsymbol{K},h} \le 1$. Writing $\boldsymbol{K}_T\nabla(u-\pi^{1,k+1}_{\boldsymbol{K},T}u)\cdot\boldsymbol{n}_{TF} = \boldsymbol{K}_T^{1/2}\nabla(u-\pi^{1,k+1}_{\boldsymbol{K},T}u)\cdot\boldsymbol{K}_T^{1/2}\boldsymbol{n}_{TF}$, using Cauchy–Schwarz inequalities and $|\boldsymbol{K}_T^{1/2}\boldsymbol{n}_{TF}| = K_{TF}^{1/2}$, we have

$$
\begin{aligned}
|\mathfrak{T}_1| &\le \sum_{T\in\mathcal{T}_h}\sum_{F\in\mathcal{F}_T} K_{TF}^{1/2}\|\boldsymbol{K}_T^{1/2}\nabla(u-\pi^{1,k+1}_{\boldsymbol{K},T}u)\|_F\|v_F-v_T\|_F \\
&\le \left(\sum_{T\in\mathcal{T}_h}\sum_{F\in\mathcal{F}_T} d_{TF}\|\boldsymbol{K}_T^{1/2}\nabla(u-\pi^{1,k+1}_{\boldsymbol{K},T}u)\|_F^2\right)^{1/2}\left(\sum_{T\in\mathcal{T}_h}\sum_{F\in\mathcal{F}_T}\frac{K_{TF}}{d_{TF}}\|v_F-v_T\|_F^2\right)^{1/2} \\
&\lesssim \left(\sum_{T\in\mathcal{T}_h}\overline{K}_{\phi,\widehat{T}}h_T^{2(r+1)}|u|^2_{H^{r+2}(T)}\right)^{1/2},
\end{aligned}
\tag{44}
$$

where we have used (17) (with $\ell = k+1$ and $s = r+2$) and the norm equivalence (31) to write $\sum_{T\in\mathcal{T}_h}\sum_{F\in\mathcal{F}_T}\frac{K_{TF}}{d_{TF}}\|v_F-v_T\|_F^2 \lesssim \sum_{T\in\mathcal{T}_h}\mathrm{a}_{\boldsymbol{K},T}(\underline{v}_T,\underline{v}_T) = \mathrm{a}_{\boldsymbol{K},h}(\underline{v}_h,\underline{v}_h) \le 1$. To estimate $\mathfrak{T}_2$, we also use Cauchy–Schwarz inequalities, the bound $\sum_{T\in\mathcal{T}_h}\mathrm{s}_{\boldsymbol{K},T}(\underline{v}_T,\underline{v}_T) \le \sum_{T\in\mathcal{T}_h}\mathrm{a}_{\boldsymbol{K},T}(\underline{v}_T,\underline{v}_T) \le 1$ and the transport relation (35) to write

$$
|\mathfrak{T}_2| \le \left(\sum_{T\in\mathcal{T}_h}\mathrm{s}_{\boldsymbol{K},T}(\underline{I}^k_T u,\underline{I}^k_T u)\right)^{1/2} \lesssim \left(\sum_{T\in\mathcal{T}_h}J\phi_T^{-1}\mathrm{s}_{\boldsymbol{K},\phi,\widehat{T}}(\widehat{\underline{I}^k_T u},\widehat{\underline{I}^k_T u})\right)^{1/2}.
$$

Since $\widehat{T}$ is isotropic and $\widehat{\underline{I}^k_T u} = \underline{I}^k_{\widehat{T}}\widehat{u}$ (owing to (28)), the consistency properties [7, Lemma 3.10] of $\mathrm{s}_{\boldsymbol{K},\phi,\widehat{T}}$ and the relations $h_{\widehat{T}} \lesssim h_T$ and (21) yield

$$
\begin{aligned}
|\mathfrak{T}_2| &\lesssim \left(\sum_{T\in\mathcal{T}_h}J\phi_T^{-1}\overline{K}_{\phi,\widehat{T}}\alpha_{\boldsymbol{K},\phi,\widehat{T}}h_{\widehat{T}}^{2(r+1)}|\widehat{u}|^2_{H^{r+2}(\widehat{T})}\right)^{1/2} \\
&\lesssim \left(\sum_{T\in\mathcal{T}_h}\overline{K}_{\phi,\widehat{T}}\alpha_{\boldsymbol{K},\phi,\widehat{T}}h_T^{2(r+1)}|u|^2_{H^{r+2}(T)}\right)^{1/2}.
\end{aligned}
$$

Plug this estimate and (44) into (43), use $\alpha_{\boldsymbol{K},\phi,\widehat{T}} \ge 1$ and recall (42) to conclude.

5 Numerical Evaluation of the Effects of Diffusion Anisotropy and Mesh Skewness

We provide here a series of numerical results, on the domain $\Omega = (0,1)^2$ (and with non-homogeneous Dirichlet boundary conditions —see [7, Sect. 2.4] for the adaptation of the scheme (38) to this case), to assess the practical optimality of the error estimate (39) and its consequence (41) in cases of highly anisotropic diffusion tensor and/or skewed mesh families. The accuracy of the HHO scheme is measured through the following two relative errors:

$$E_{\mathrm{a},\boldsymbol{K},h} := \frac{\|\underline{I}_h^k u - \underline{u}_h\|_{\mathrm{a},\boldsymbol{K},h}}{\|\underline{I}_h^k u\|_{\mathrm{a},\boldsymbol{K},h}} \quad \text{and} \quad E_{1,h} := \frac{\|\underline{I}_h^k u - \underline{u}_h\|_{1,h}}{\|\underline{I}_h^k u\|_{1,h}},$$

where $\|\cdot\|_{\mathrm{a},\boldsymbol{K},h}$ is defined by (37), and $\|\cdot\|_{1,h}$ is the diffusion-independent discrete H^1-norm obtained adding together the local seminorms (22) with $\boldsymbol{K} = I_d$, that is:

$$\|\underline{v}_h\|_{1,h} := \left(\sum_{T\in\mathcal{T}_h}\left[\|\nabla v_T\|_T^2 + \sum_{F\in\mathcal{F}_T} d_{TF}^{-1}\|v_F - v_T\|_F^2\right]\right)^{1/2}. \tag{45}$$

The numerical tests have been performed using the code “HHO-Diffusion” in the C++ open source library HArDCore [9]. This library provides generic tools for implementing 2D and 3D numerical methods with unknowns made of polynomials on the edges/faces and cells of the mesh; it also naturally handles generic polygonal and polyhedral meshes. The variety of possible tests to assess the practical efficiency of the scheme (38) with anisotropic diffusion/skewed meshes is infinite, given the numerous possible parameters (polynomial degrees k, diffusion tensors, exact solutions, type of meshes, etc). We only report a few relevant results here, but all the meshes and data used in the tests below are available in HArDCore for the interested reader to run additional tests.

5.1 Test A: Anisotropic Diffusion Tensor

This test focuses on the effect of an anisotropic and heterogeneous diffusion tensor. For $\lambda \in \{10^{-6}, 1, 10^6\}$, we consider the tensor

$$\boldsymbol{K}(x,y) = \begin{bmatrix}\lambda & 0\\ 0 & 1\end{bmatrix} \quad \text{if } y < 0.5, \qquad \boldsymbol{K}(x,y) = I_d \quad \text{if } y \ge 0.5,$$

and fix the exact solution $u(x,y) = \cos(\pi x)\cos(\pi y)$; the source term and boundary conditions are computed from this solution. Since $(\partial_x u)_{|y=0.5} = 0$, we still have $\nabla\cdot(\boldsymbol{K}\nabla u) \in L^2(\Omega)$ despite the discontinuity of $\boldsymbol{K}$ along $y = 0.5$. We consider a family of locally refined meshes from the FVCA5 benchmark [10] (see Fig. 1), for which the setting of Remark 4 holds with $\lambda_T = \lambda$ and $a_T = b_T = 1$; the estimate (41) therefore predicts a dependency of the energy error on $\max(\lambda, \lambda^{-\frac{1}{2}})$. The results for $k = 1, 3$ are presented in Fig. 2; tests with other polynomial degrees present the same trend. The energy error $E_{\mathrm{a},\boldsymbol{K},h}$ appears to depend much less on the anisotropy ratio than predicted; the error $E_{1,h}$ shows a more pronounced dependency on the tensor anisotropy, especially for low degrees where a factor of about 30 is seen on the finest mesh between $\lambda = 1$ and $\lambda = 10^{-6}, 10^{6}$.

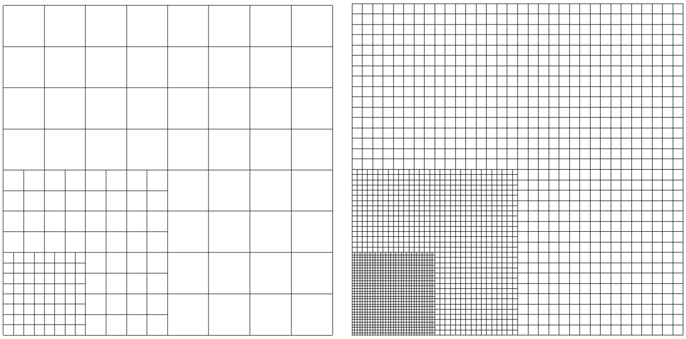
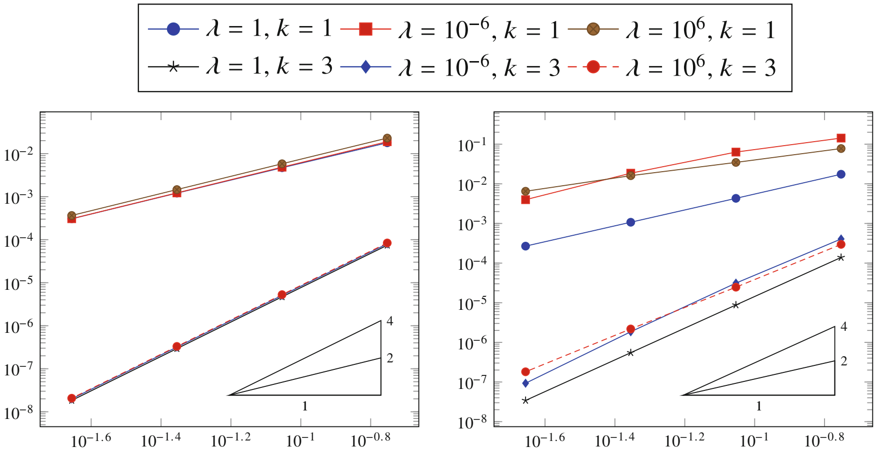


Fig. 1 Two members of the family of meshes used in Test A



(a) $E_{\mathrm{a},\boldsymbol{K},h}$ vs. h.

(b) $E_{1,h}$ vs. h.

Fig. 2 Errors versus meshsize for Test A. Slopes = rates expected from (39)

5.2 *Test B: Skewed Mesh*

In this test, we study the impact of the mesh skewness. We take $\boldsymbol{K} = I_d$ and the exact solution $u(x, y) = \cos(\pi x)\cos(\pi y)$. The meshes are (mostly) hexagonal, and more and more skewed as their size decreases (see Fig. 3). The results in Fig. 4 show a clear loss of rate of convergence, compared to the expected rate for isotropic meshes.

To estimate more precisely the effect of mesh skewness, we introduce the *flatness* factor defined as $\mathrm{fl}_h := \max_{T\in\mathcal{T}_h} \mathrm{fl}_T$ with $\mathrm{fl}_T := \frac{h_T}{\rho_T}$, where ρ_T is the radius of the largest ball centred at the centre of mass of T and contained in T. The skewness of the considered meshes comes from the large flatness factors fl_T of some elements T. It is easy to convince oneself that this setting is compatible with Remark 4 with $\lambda = 1$, $a_T = \mathrm{fl}_T$ and $b_T = 1$. As a consequence, (41) predicts an upper bound

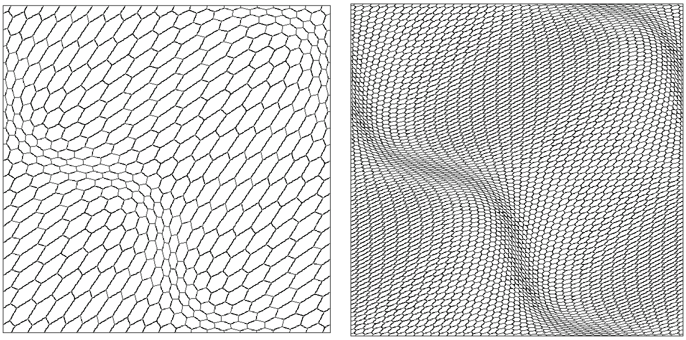
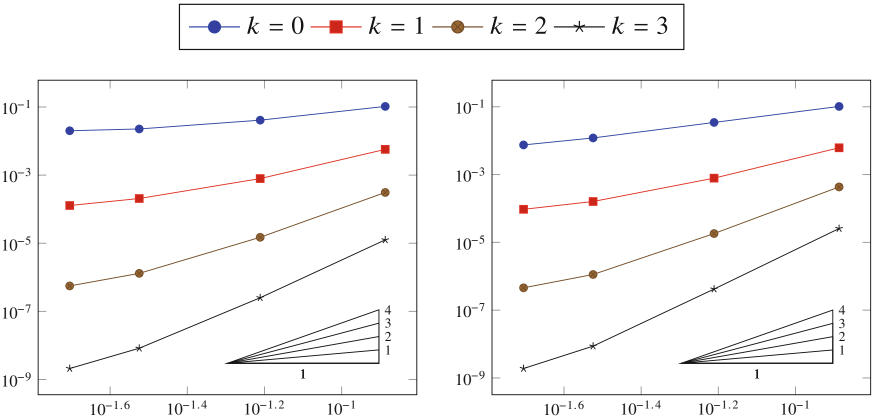


Fig. 3 First two meshes in the skewed family used in Test B



(a) $E_{\mathrm{a},\boldsymbol{K},h}$ vs. h.

(b) $E_{1,h}$ vs. h.

Fig. 4 Errors versus meshsize for Test B. The slopes indicate the expected rates of convergence h^{k+1}, disregarding the effects of the mesh skewness

$$E_{\mathrm{a},\boldsymbol{K},h} \lesssim \mathrm{fl}_h^2 h^{k+1} |u|_{H^{k+2}(\mathcal{T}_h)}. \tag{46}$$

To evaluate the accuracy of this estimate with respect to the mesh flatness, for each error $E_h \in \{E_{\mathrm{a},\boldsymbol{K},h}, E_{1,h}\}$ we provide in Table 1 an evaluation of the rates of growth of E_h/h^{k+1} with respect to fl_h. Estimate (46) tells us that, at least for the energy error, this rate should be at a maximum of 2. As can be seen in Table 1, the actual rates are much smaller than 2, and both errors are less sensitive to the mesh flatness than (46) predicts; the diffusion-independent norm $E_{1,h}$ is the least sensitive of both.

Table 1 also reports the condition numbers (CN) in 1-norm for the statically condensed system. For regular mesh sequences, CNs of HHO systems grow as h^{-2}. Here, the growth is in h^{-4} (but the CNs do not depend much on k). The additional power of 2 could come from a factor fl^2 (since, here, $\mathrm{fl} \sim 1/h$). Further analysis and tests are however necessary to reach a definitive conclusion, and it should also be noted that the meshes considered here contain a large portion of skewed elements; the condition numbers could be reduced for meshes with a smaller portion of distorted cells.

5.3 Test C

We assess here the interplay between mesh skewness and diffusion anisotropy, taking $\boldsymbol{K}(x, y) = \mathrm{diag}(10^6, 1)$, and $u(x, y) = \cos(\pi x)\cos(\pi y)$ as before. We consider two families of meshes: regular hexagonal, and skewed hexagonal with flatness factor multiplied by two from one mesh member to the next; see Fig. 5.

The setting of Remark 4 is valid with $(a_T, b_T) = (1, \mathrm{fl}_h)$ with $\mathrm{fl}_h \leq 10^3$ for the considered meshes; (41) thus predicts a bound $E_{\mathrm{a},\boldsymbol{K},h} \lesssim 10^6 \mathrm{fl}_h^{-1} h^{k+1} |u|_{H^{k+2}(\Omega)}$. For the skewed meshes, we have $\mathrm{fl}_h \sim 1/h$ and we therefore expect a better rate of convergence than the usual h^{k+1} rate for isotropic meshes. Figure 6 confirms this improvement, albeit in a non-uniform way.

The improvement is clearer if we superimpose the errors for the families of regular and skewed meshes, see Figs. 7a, b. For a given meshsize, selecting a mesh that is stretched in the direction of strong diffusion improves the convergence in both norms; this gain is valid for all degrees, but especially prominent for the lowest-order case $k = 0$ (for which, at the considered meshsizes, there is no apparent convergence on non-stretched meshes). In Fig. 7c, d the same errors are plotted against the number of globally coupled degrees of freedom, which for HHO schemes correspond to the edge unknowns (the element unknowns can be eliminated by static condensation [7, Appendix B]). In terms of errors vs. number of degrees of freedom, the gain in using skewed meshes is less clear, except for $k = 0$; the reason is that meshes entirely made of stretched elements usually have, for a given meshsize, more edges than regular meshes.

Table 1 Rates of convergence of the errors with respect to the mesh flatness, Test B

h	fl_h	CN		$\frac{E_{\mathrm{a},\boldsymbol{K},h}}{h^{k+1}}$	rate		$\frac{E_{1,h}}{h^{k+1}}$	rate
0.13	10	875		8e−01	–		7.9e−01	–
0.06	22	1.8e+04		6.7e−01	−0.2		5.6e−01	−0.5
0.03	46	2.6e+05		7.6e−01	0.2		4.0e−01	−0.4
0.02	70	1.3e+06		1e+00	0.7		3.8e−01	−0.1

$k = 0$

h	fl_h	CN		$\frac{E_{\mathrm{a},\boldsymbol{K},h}}{h^{k+1}}$	rate		$\frac{E_{1,h}}{h^{k+1}}$	rate
0.13	10	1.7e+03		3.4e−01	–		3.7e−01	–
0.06	22	3.1e+04		2.1e−01	−0.6		2.1e−01	−0.7
0.03	46	5.0e+05		2.3e−01	0.1		1.8e−01	−0.2
0.02	70	2.6e+06		3.3e−01	0.8		2.4e−01	0.7

$k = 1$

h	fl_h	CN		$\frac{E_{\mathrm{a},\boldsymbol{K},h}}{h^{k+1}}$	rate		$\frac{E_{1,h}}{h^{k+1}}$	rate
0.13	10	2.7e+03		1.4e−01	–		2.0e−01	–
0.06	22	4.3e+04		6.4e−02	−1		7.8e−02	−1.2
0.03	46	7.7e+05		4.9e−02	−0.4		4.2e−02	−0.8
0.02	70	4.0e+06		7.3e−02	0.9		5.9e−02	0.8

$k = 2$

h	fl_h	CN		$\frac{E_{\mathrm{a},\boldsymbol{K},h}}{h^{k+1}}$	rate		$\frac{E_{1,h}}{h^{k+1}}$	rate
0.13	10	3.9e+03		4.4e−02	–		9.1e−02	–
0.06	22	5.6e+04		1.8e−02	−1.2		2.9e−02	−1.5
0.03	46	1.1e+06		1.0e−02	−0.7		1.1e−02	−1.3
0.02	70	5.6e+06		1.4e−02	0.7		1.3e−02	0.3

$k = 3$

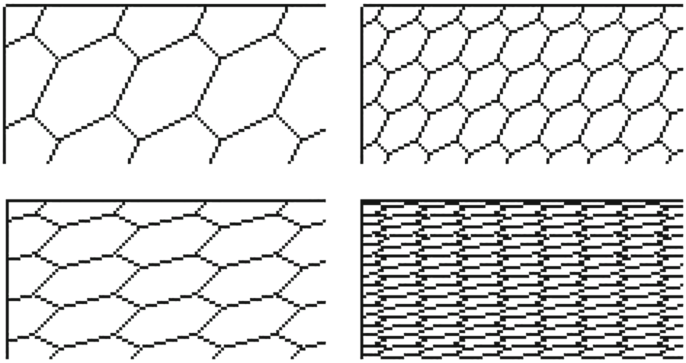
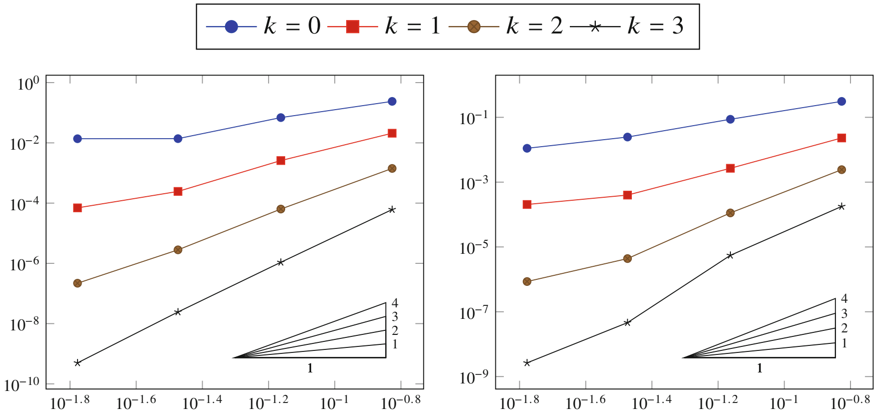


Fig. 5 Upper left corner of the meshes in Test C: regular hexagonal (top); skewed hexagonal (bottom)

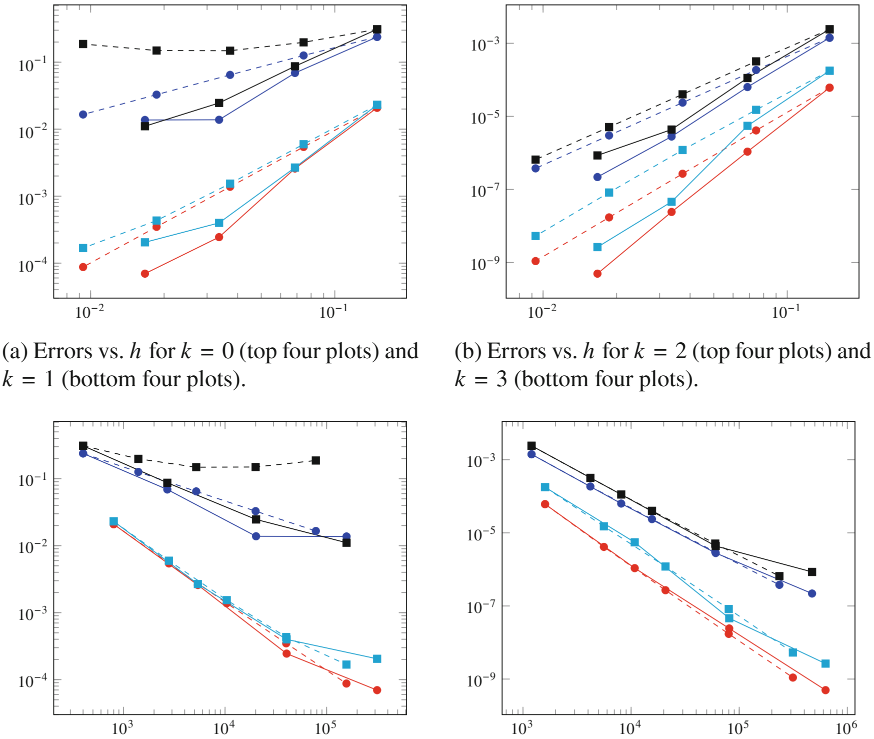


(a) $E_{\mathrm{a},\boldsymbol{K},h}$ vs. h. (b) $E_{1,h}$ vs. h.

Fig. 6 Errors in Test C for the family of skewed hexagonal meshes. The slopes correspond to the h^{k+1} rates expected for non-skewed meshes

6 Conclusion

We presented a theoretical and numerical study of the accuracy and robustness of the classical HHO method, when applied to anisotropic diffusion equations on distorted meshes. We defined a notion of mesh sequences that accepts in particular elements that become more and more elongated as the mesh is refined, and we established an error estimate that tracks the dependency of the constants with respect to the local diffusion anisotropy and elements skewness. We then presented the results of several numerical tests designed to explore the optimality of the error estimate. These results indicate that some behaviours highlighted by the theoretical estimate (such as the interplay between diffusion anisotropy and mesh skewness) are perceptible



(a) Errors vs. h for $k = 0$ (top four plots) and $k = 1$ (bottom four plots).

(b) Errors vs. h for $k = 2$ (top four plots) and $k = 3$ (bottom four plots).

(c) Errors vs. nb DOFs for $k = 0$ (top four plots) and $k = 1$ (bottom four plots).

(d) Errors vs. nb DOFs for $k = 2$ (top four plots) and $k = 3$ (bottom four plots).

Fig. 7 Test C: comparison between regular (dashed lines) and skewed (continuous lines) hexagonal meshes. Round markers: $E_{\mathrm{a},\boldsymbol{K},h}$; square markers: $E_{1,h}$

in practical numerical results, but they also show that this estimate appears to be pessimistic in its prediction of the behaviour of the error in case of strong anisotropy or skewness.

Further work remains to be done to obtain more optimal estimates in terms of dependency with respect to the tensor anisotropy (this only has to be done for non-skewed meshes, as our approach would then provide an optimal estimate for skewed meshes). An aspect that is not covered by our definition of regular skewed mesh sequences is the case of small edges/faces in otherwise isotropic elements; another approach has to be adopted to derive error estimates in such situations. Finally, even though the standard HHO scheme displays some level of robustness on distorted meshes, it would be interesting to develop a variant that is specifically adapted to such meshes, and leads to better condition numbers than the standard method.

Acknowledgements This work was partially supported by the Australian Government through the Australian Research Council's Discovery Projects funding scheme (project number DP170100605).

References

1. Antonietti, P.F., Berrone, S., Verani, M., Weißer, S.: The virtual element method on anisotropic polygonal discretizations. In: Numerical Mathematics and Advanced Applications-ENUMATH 2017, vol. 126. Lecture Notes for Computer Science Engineering. Springer, Cham, pp. 725–733 (2019)
2. Beirão da Veiga, L., Brezzi, F., Cangiani, A., Manzini, G., Marini, L.D., Russo, A.: Basic principles of virtual element methods. Math. Models Methods Appl. Sci. (M3AS) **199**(23), 199–214 (2013)
3. Cockburn, B., Gopalakrishnan, J., Lazarov, R.: Unified hybridization of discontinuous Galerkin, mixed, and continuous Galerkin methods for second order elliptic problems. SIAM J. Numer. Anal. **47**(2), 1319–1365 (2009)
4. Di Pietro, D.A., Droniou, J.: A third Strang lemma for schemes in fully discrete formulation. Calcolo **55**(40) (2018)
5. Di Pietro, D.A., and A. Ern. Mathematical aspects of discontinuous Galerkin methods. Mathématiques & Applications, vol. 69. Springer, Berlin, Heidelberg, pp. xviii+384 (2012). ISBN: 978-3-642-22979-4
6. Di Pietro, D.A., Ern, A., Lemaire, S.: An arbitrary-order and compactstencil discretization of diffusion on general meshes based on local reconstruction operators. Comput. Methods Appl. Math. **14**(4), 461–472 (2014)
7. Di Pietro, D.A., Droniou, J.: The Hybrid High-Order Method for Polytopal Meshes: Design, Analysis, and Applications. Modeling, Simulation and Applications. To appear. Springer International Publishing, pp. xxxii + 498p (2020). https://hal.archives-ouvertes.fr/hal-02151813
8. Droniou, J., Eymard, R., Gallouët, T., Herbin, R.: A unified approach to mimetic finite difference, hybrid finite volume and mixed finite volume methods. Math. Models Methods Appl. Sci. (M3AS) **20**(2), pp. 1–31 (2010)
9. HArDCore—Hybrid Arbitrary Degree: Core. Version 2.0. https://github.com/jdroniou/HArDCore
10. Herbin, R., Hubert, F.: Benchmark on discretization schemes for anisotropic diffusion problems on general grids. In: Eymard, R., Hérard, J.-M. (eds.) Finite Volumes for Complex Applications V. Wiley, pp. 659–692 (2008)
11. Wang, J., Ye, X.: A weak Galerkin finite element method for second-order elliptic problems. J. Comput. Appl. Math. **241**, 103–115 (2013)
12. Weißer, S.: Anisotropic polygonal and polyhedral discretizations in finite element analysis. ESAIM Math. Model. Numer. Anal. **53**(2), 475–501 (2019). ISSN: 0764-583X

$\mathscr{K}$-Convergence of Finite Volume Solutions of the Euler Equations

Mária Lukáčová-Medvid'ová

Abstract We review our recent results on the convergence of invariant domain-preserving finite volume solutions to the Euler equations of gas dynamics. If the classical solution exists we obtain strong convergence of numerical solutions to the classical one applying the weak-strong uniqueness principle. On the other hand, if the classical solution does not exist we adapt the well-known Prokhorov compactness theorem to space-time probability measures that are generated by the sequences of finite volume solutions and show how to obtain the strong convergence in space and time of observable quantities. This can be achieved even in the case of ill-posed Euler equations having possibly many oscillatory solutions.

Keywords Convergence analysis · Finite volume methods · Euler equations · Ill-posedness · Dissipative measure–valued solutions · $\mathscr{K}$-convergence

MSC (2010) 65M08 · 76N10 · 35L65 · 35R06

1 Introduction

Hyperbolic conservation laws are fundamental for most of physical, biological and mechanical processes. The iconic example of this class of partial differential equations are the Euler equations of gas dynamics. Being published in 1757 by Leonhard Euler in Mémoires de l'Académie des Sciences de Berlin in his article "Principes généraux du mouvement des fluides" the Euler equations are one of the first written partial differential equations at all. Recently, multidimensional Euler equations have achieved renewed interest in mathematical community. Indeed, it is a well-known fact that the classical (i.e., continuously differentiable) solution exists in general only for a short time since discontinuities (shocks) may develop. A suitable gener-

M. Lukáčová-Medvid'ová (✉)
Institut für Mathematik, Johannes Gutenberg-Universität Mainz,
Staudingerweg 9, 55 128 Mainz, Germany
e-mail: lukacova@uni-mainz.de

R. Klöfkorn et al. (eds.), *Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples*, Springer Proceedings in Mathematics & Statistics 323,
https://doi.org/10.1007/978-3-030-43651-3_2

alization is to consider weak solutions, which moreover satisfy the second law of thermodynamics.

As shown by De Lellis and Székelyhidi [8] and by Chiodaroli et al. [6] infinitely many weak entropy solutions can be constructed for the multidimensional compressible Euler equations. Their ill-posedness is related to the lack of compactness of a set of weak entropy solutions. Such a failure of well-posedness (i.e., uniqueness) of the multidimensional Euler equations in the class of weak entropy solutions is connected to the turbulence effects which are apparently not appropriately described by the concept of distributional solutions.

On the other hand, we can find in literature a large variety of powerful numerical schemes, typically finite volume or discontinuous Galerkin methods, that are successfully used in order to approximate multidimensional hyperbolic conservation laws and the Euler equations, in particular. We refer to monographs [11, 12, 18, 22, 28, 30, 33] and the references therein. Despite the popularity of the finite volume and discontinuous Galerkin methods for practical applications their theoretical convergence analysis for multidimensional hyperbolic conservation laws is still incomplete. We should mention for example the convergence and error analysis obtained in [26] for the Cauchy problem of a general multidimensional hyperbolic conservation law. Under the assumption of the existence of the classical solution the authors applied the stability result due to Dafermos [7] and DiPerna’s method [9] in order to derive the error estimates for the explicit finite volume schemes satisfying the discrete entropy inequality. Consequently, they proved the strong convergence of the entropy stable finite volume schemes.

In view of the facts that the classical solution may not exist and the weak entropy solutions are non-unique new probabilistic concepts have been developed. In [9, 10, 29, 31] the so-called measure–valued solutions to hyperbolic conservation laws are studied. The latter are represented by the Young measures, which are space-time parametrized probability measures acting on a (solution) phase space. Measure–valued solutions have been also successfully used in [19, 21] in order to show convergence of the entropy stable finite volume schemes for general hyperbolic conservation laws under additional assumptions on the boundedness of numerical solutions or a growth condition on the flux function. Another interesting contribution to the convergence analysis of the Euler equations was presented in [29], where the limit of higher order viscous regularization to the Euler equations was identified with a measure–valued solution that exists globally in time.

Clearly, the set of (entropy) measure–valued solutions is larger than that of (entropy) weak solutions and thus the question of uniqueness remains still open. However, a recently introduced concept of dissipative measure–valued (DMV) solutions [5, 25] allows to show the DMV-strong uniqueness principle. It means that DMV solutions coincide with the strong solution as long as the latter exists.

The aim of the present paper is to review our recent results on the convergence analysis of some finite volume methods. It turned out that some invariant domain-preserving properties, such as the entropy stability, preservation of positivity of density and internal energy and minimum entropy principle are important in order to obtain convergence of a numerical scheme without any additional nonphysical assumptions [15]. We also report on the recently established concept of $\mathscr{K}$-convergence which allows to compute observable quantities of possibly strongly oscillating dissipative measure–valued solutions [14, 17]. We wish to give a clear overview of main convergence results without going into deep theoretical justifications. In this way we hope to attract the attention of more experimentally oriented computational scientists and to encourage them to apply $\mathscr{K}$-convergence to other well-known finite volume and discontinuous Galerkin methods. A reader interested in further theoretical details and proofs is referred to [14–17] and the references therein.

In what follows we firstly introduce the dissipative measure–valued and dissipative weak solutions of the Euler equations and describe a suitable invariant domain-preserving finite volume method. Consequently, we present the strong convergence results for single realizations under the assumption that the classical solution to the Euler equation exists and the strong convergence result of observable quantities in a general case.

2 Euler Equations and Dissipative Solutions

The gas dynamics of inviscid compressible flows is governed by the Euler equations

$$\begin{aligned} \partial_t \rho + \operatorname{div} \boldsymbol{m} &= 0, \\ \partial_t \boldsymbol{m} + \operatorname{div}(\boldsymbol{m} \otimes \boldsymbol{u}) + \nabla p &= 0, \\ \partial_t E + \operatorname{div}((E + p)\boldsymbol{u}) &= 0, \end{aligned} \tag{1}$$

where ρ, $\boldsymbol{m}$ and E represent the conservative variables, the density, momentum and the total energy, respectively. Further, p and $\boldsymbol{u} = \boldsymbol{m}/\rho$ stand for the pressure and velocity. The total energy $E = \frac{1}{2}\frac{\boldsymbol{m}^2}{\rho} + \rho e$ consists of the kinetic energy and the internal energy e.

System (1) is closed by the standard pressure law for a perfect gas $p(\rho, \vartheta) = R\rho\vartheta$, ϑ is the temperature and R the gas constant. We assume without loss of generality that $R = 1$. We denote by γ the adiabatic coefficient and by c_V the specific heat at constant volume, $c_V = \frac{1}{\gamma-1}$. In what follows we will assume that $1 < \gamma < 2$ and note that this covers the physically reasonable range for gases $1 < \gamma \le 5/3$. Denoting s the specific physical entropy and S the total entropy we have

$$s(\rho, \vartheta) = \log\left(\frac{\vartheta^{c_V}}{\rho}\right) = \frac{1}{\gamma - 1}\log\left(\frac{p}{\rho^\gamma}\right), \quad S = \rho s \quad \text{and} \quad e(\rho, \vartheta) = c_V \vartheta.$$

On a space-time cylinder $\Omega \times (0, T)$, $\Omega \subset \mathbb{R}^d$, $d = 2, 3$, $T > 0$, the system of the Euler equations is accompanied by appropriate boundary and initial conditions. Here we assume the periodic or the no flux boundary conditions

$$\boldsymbol{u}|_{\partial\Omega} \cdot \boldsymbol{n} = 0, \ \frac{\partial\vartheta}{\partial\boldsymbol{n}} = 0$$

and set

$$\rho(t=0) = \rho_0, \quad \boldsymbol{m}(t=0) = \boldsymbol{m}_0, \quad E(t=0) = E_0.$$

In [16] it has been proved that numerical solutions obtained by suitable numerical schemes (such as invariant domain-preserving finite volume methods) either converge strongly in suitable Bochner spaces or their (weak) limit is not a weak entropy solution. Clearly, such a result calls for a new concept of generalized solutions to the Euler equations. Following [2, 5] we introduce the dissipative measure–valued solutions and dissipative weak solutions. The latter can be seen as the statistical mean values with respect to the corresponding Young measures.

Definition 1 (*Dissipative measure–valued solution*) [5, 16] A parametrized probability measure $\{\mathscr{V}_{t,x}\}_{(t,x)\in(0,T)\times\Omega}$,

$$\mathscr{V} \in L^\infty((0,T)\times\Omega; \mathscr{P}(\mathbb{R}^{d+2})), \ \mathbb{R}^{d+2} = \left\{[\tilde{\rho}, \tilde{\boldsymbol{m}}, \widetilde{S}] \,\middle|\, \tilde{\rho} \in \mathbb{R}, \ \tilde{\boldsymbol{m}} \in \mathbb{R}^d, \ \widetilde{S} \in \mathbb{R}\right\},$$

is called *dissipative measure valued (DMV) solution* of the Euler system (1) if the following holds:

- **lower bound on density and entropy**
  there exists $\underline{s} \in \mathbb{R}$ such that

$$\mathscr{V}_{t,x}\Big[\ \big\{\rho \geq 0, \ S \equiv s\rho \geq \underline{s}\rho\big\}\ \Big] = 1 \text{ for a.a. } (t,x); \tag{2}$$

- **integral energy inequality**[1]

$$\int_\Omega \left\langle \mathscr{V}_{\tau,x}; \frac{1}{2}\frac{|\tilde{\boldsymbol{m}}|^2}{\tilde{\rho}} + \tilde{\rho} e(\tilde{\rho}, \widetilde{S}) \right\rangle \,\mathrm{d}x + \int_{\overline{\Omega}} \mathrm{d}\mathfrak{E}_{cd}(\tau) \leq \int_\Omega \left[\frac{1}{2}\frac{|\boldsymbol{m}_0|^2}{\rho_0} + \rho_0 e(\rho_0, S_0)\right] \,\mathrm{d}x \tag{3}$$

  holds for a.a. $0 \leq \tau \leq T$, with the energy concentration defect

$$\mathfrak{E}_{cd} \in L^\infty(0,T; M^+(\overline{\Omega})),$$

  where $M^+(\overline{\Omega})$ denotes the space of positive Radon measures on $\overline{\Omega}$;
- **equation of continuity**

$$\left[\int_\Omega \langle \mathscr{V}; \tilde{\rho}\rangle \ \mathrm{d}x\right]_{t=0}^{t=\tau} = \int_0^\tau \int_\Omega \Big[\langle \mathscr{V}; \tilde{\rho}\rangle\, \partial_t\varphi + \langle \mathscr{V}; \tilde{\boldsymbol{m}}\rangle \cdot \nabla\varphi\Big] \,\mathrm{d}x\,\mathrm{d}t \tag{4}$$

[1] Here the mean value $\left\langle \mathscr{V}_{t,x}; b\left(\tilde{\boldsymbol{U}}\right)\right\rangle \equiv \int_{\mathbb{R}^{d+2}} b\left(\tilde{\boldsymbol{U}}\right) d\mathscr{V}_{t,x}(\tilde{\boldsymbol{U}})$ for $\boldsymbol{U} \in \mathbb{R}^{d+2}$ and b bounded continuous function.

for any $0 \leq \tau \leq T$, and any $\varphi \in W^{1,\infty}((0,T) \times \Omega)$;

- **momentum equation**

$$\left[\int_{\Omega} \langle \mathscr{V}; \tilde{\boldsymbol{m}} \rangle \cdot \boldsymbol{\varphi} \ \mathrm{d}x \right]_{t=0}^{t=\tau} = \int_0^{\tau} \int_{\Omega} \left[\langle \mathscr{V}; \tilde{\boldsymbol{m}} \rangle \cdot \partial_t \boldsymbol{\varphi} + \left\langle \mathscr{V}; 1_{\tilde{\rho}>0} \frac{\tilde{\boldsymbol{m}} \otimes \tilde{\boldsymbol{m}}}{\tilde{\rho}} \right\rangle : \nabla \boldsymbol{\varphi} + \left\langle \mathscr{V}; 1_{\tilde{\rho}>0} p(\tilde{\rho}, \widetilde{S}) \right\rangle \mathrm{div} \boldsymbol{\varphi} \right] \mathrm{d}x \, \mathrm{d}t + \int_0^{\tau} \int_{\overline{\Omega}} \nabla \boldsymbol{\varphi} : \mathrm{d}\mathfrak{R}_{cd}(t) \, \mathrm{d}t \tag{5}$$

for any $0 \leq \tau \leq T$, and any $\boldsymbol{\varphi} \in C^m([0,T] \times \overline{\Omega}; R^d)$, $\boldsymbol{\varphi} \cdot \boldsymbol{n}|_{\partial \Omega} = 0, m \geq 1$, with the Reynolds concentration defect

$$\mathfrak{R}_{cd} \in L^{\infty}(0,T; M^+(\overline{\Omega}; R^{d \times d}_{\mathrm{sym}}))$$

satisfying

$$\underline{d} \ \mathfrak{E}_{cd} \leq \mathrm{tr}[\mathfrak{R}_{cd}] \leq \overline{d} \ \mathfrak{E}_{cd} \text{ for some constants } 0 < \underline{d} \leq \overline{d}; \tag{6}$$

- **entropy inequality**

$$\left[\int_{\Omega} \left\langle \mathscr{V}; \widetilde{S} \right\rangle \varphi \ \mathrm{d}x \right]_{t=\tau_1-}^{t=\tau_2+} \geq \int_{\tau_1}^{\tau_2} \int_{\Omega} \left[\left\langle \mathscr{V}; \widetilde{S} \right\rangle \partial_t \varphi + \left\langle \mathscr{V}; 1_{\tilde{\rho}>0} \left(\widetilde{S} \tilde{\boldsymbol{u}} \right) \right\rangle \cdot \nabla \varphi \right] \ \mathrm{d}x \, \mathrm{d}t \tag{7}$$

for any $0 \leq \tau_1 \leq \tau_2 < T$, and any $\varphi \in W^{1,\infty}((0,T) \times \Omega)$, $\varphi \geq 0$.

The DMV solution is a very general concept that allows to show the convergence of invariant domain-preserving schemes in an elegant way. Despite its generality it still satisfies the DMV-strong uniqueness principle [5] and thus the DMV solutions coincide with the classical solution as long as the latter exists. To prove the latter the crucial properties are the energy dissipation (3) and (6) controlling the Reynolds defect in the momentum equation by the energy concentration defect. It is to be pointed out that the Reynolds concentration defect brings an additional freedom to model turbulent flow behaviour.

To simplify the viewpoint on this generalized solutions it often suffices to consider only the mean values of DMV solutions, which are the below-mentioned dissipative solutions.

Definition 2 (*dissipative weak solution*) [16] A triple $[\rho, \boldsymbol{m}, S]$ is *dissipative weak (DW) solution* of the full Euler system (1) if the following holds:

- **weak continuity** in time

$$\begin{aligned} &\rho \in C_{\mathrm{weak}}([0,T]; L^{\gamma}(\Omega)), \ (\gamma \text{ being the adiabatic constant}) \\ &\boldsymbol{m} \in C_{\mathrm{weak}}([0,T]; L^{\frac{2\gamma}{\gamma+1}}(\Omega; \mathbb{R}^d)), \\ &S \in L^{\infty}(0,T; L^{\gamma}(\Omega)) \cap BV_{\mathrm{weak}}([0,T]; L^{\gamma}(\Omega)); \end{aligned} \tag{8}$$

- **energy inequality**: there exists a measure

$$\mathfrak{E} \in L^\infty(0,T;\mathcal{M}^+(\overline{\Omega})),$$

such that the inequality

$$\int_\Omega \left[\frac{1}{2}\frac{|\boldsymbol{m}|^2}{\rho} + \rho e(\rho,S)\right](\tau,\cdot)\ \mathrm{d}x + \int_{\overline{\Omega}} \mathrm{d}\mathfrak{E}(\tau) \leq \int_\Omega \left[\frac{1}{2}\frac{|\boldsymbol{m}_0|^2}{\rho_0} + \rho_0 e(\rho_0,S_0)\right]\ \mathrm{d}x \tag{9}$$

holds for a.a. $0 \leq \tau \leq T$;

- **equation of continuity**

$$\left[\int_\Omega \rho\varphi\ \mathrm{d}x\right]_{t=0}^{t=\tau} = \int_0^\tau \int_\Omega \left[\rho\partial_t\varphi + \boldsymbol{m}\cdot\nabla\varphi\right]\ \mathrm{d}x\,\mathrm{d}t \tag{10}$$

holds for any $0 \leq \tau \leq T$;

- **momentum equation**

$$\left[\int_\Omega \boldsymbol{m}\cdot\boldsymbol{\varphi}\ \mathrm{d}x\right]_{t=0}^{t=\tau} = \int_0^\tau \int_\Omega \left[\boldsymbol{m}\cdot\partial_t\boldsymbol{\varphi} + 1_{\rho>0}\frac{\boldsymbol{m}\otimes\boldsymbol{m}}{\rho} : \nabla\boldsymbol{\varphi} + 1_{\rho>0}p(\rho,S)\mathrm{div}\boldsymbol{\varphi}\right]\ \mathrm{d}x\,\mathrm{d}t \\ + \int_0^\tau \nabla\boldsymbol{\varphi} : \mathrm{d}\mathfrak{R} \tag{11}$$

for any $0 \leq \tau \leq T$, any test function $\boldsymbol{\varphi} \in C^m([0,T]\times\overline{\Omega};\mathbb{R}^d)$, $\boldsymbol{\varphi}\cdot\boldsymbol{n}|_{\partial\Omega} = 0$, and a defect measure

$$\mathfrak{R} \in L^\infty(0,T;\mathcal{M}^+(\overline{\Omega};\mathbb{R}^d));$$

- **entropy inequality**

$$\left[\int_\Omega S\varphi\ \mathrm{d}x\right]_{t=\tau_1-}^{t=\tau_2+} \geq \int_{\tau_1}^{\tau_2} \int_\Omega \left[S\partial_t\varphi + \left\langle\mathcal{V};1_{\tilde{\rho}>0}\left(\widetilde{S}\tilde{\boldsymbol{u}}\right)\right\rangle\cdot\nabla\varphi\right]\ \mathrm{d}x\,\mathrm{d}t,\ S(0-,\cdot) = S_0, \tag{12}$$

for any $0 \leq \tau_1 \leq \tau_2 < T$, any $\varphi \in W^{1,\infty}((0,T)\times\Omega)$, $\varphi \geq 0$, where $\{\mathcal{V}_{t,x}\}_{(t,x)\in(0,T)\times\Omega}$ is the aforementioned DMV solution

- **defect compatibility conditions**

$$\underline{d}\ \mathfrak{E} \leq \mathrm{tr}\,[\mathfrak{R}] \leq \overline{d}\ \mathfrak{E} \text{ for some constants } 0 \leq \underline{d} \leq \overline{d}, \tag{13}$$

and

$$\mathfrak{E} \geq \left\langle\mathcal{V};\frac{1}{2}\frac{|\tilde{\boldsymbol{m}}|^2}{\tilde{\rho}} + \tilde{\rho}e(\tilde{\rho},\widetilde{S})\right\rangle - \left(\frac{1}{2}\frac{|\boldsymbol{m}|^2}{\rho} + \rho e(\rho,S)\right). \tag{14}$$

The existence of DMV or DW solutions can be shown by the convergence of suitable invariant domain-preserving finite volume schemes. In what follows we will present such a finite volume method and review its convergence results for multidimensional

Euler equations. We mention in passing that in [13] the convergence of the standard Lax-Friedrichs finite volume method has been shown in an analogous way as presented below.

3 A Finite Volume Method Based on the Brenner Model

In [15] the two-velocity model for compressible flows by Brenner [3, 4] was revisited and a new invariant domain-preserving finite volume method, denoted here by the FLM method, has been proposed and analysed. To fix the notation we start by introducing a suitable discrete space and a finite volume mesh.

The finite volume grid $\mathbb{T}_h$ consists of finite volumes, denoted by K, that can be triangles, rectangles or polygons and cover the physical domain Ω

$$\overline{\Omega} = \bigcup_{K \in \mathbb{T}_h} K.$$

The parameter $h \in (0, 1)$ is the maximum element size, i.e., the size of the mesh $\mathbb{T}_h$. We assume that $\mathbb{T}_h$ is regular and quasi-uniform. The set of all faces is denoted by Σ, while the set of faces on the boundary is denoted by Σ_{ext}, and the set of interior faces by $\Sigma_{int} = \Sigma \backslash \Sigma_{ext}$. For periodic boundary conditions we set $\Sigma_{ext} = \emptyset$ and $\Sigma_{int} = \Sigma$. Further, we associate each face with its outer normal vector $\boldsymbol{n}$.

We denote by Q_h the set of piecewise constant functions on $\mathbb{T}_h$ and define for any $v \in Q_h$, $x \in \sigma \in \Sigma_{int}$

$$v^{\text{out}}(x) = \lim_{\delta \to 0+} v(x + \delta \boldsymbol{n}), \qquad v^{\text{in}}(x) = \lim_{\delta \to 0+} v(x - \delta \boldsymbol{n}),$$
$$\overline{v}(x) = \frac{v^{\text{in}}(x) + v^{\text{out}}(x)}{2}, \qquad [\![v]\!] = v^{\text{out}}(x) - v^{\text{in}}(x).$$

A numerical flux function in our finite volume method is based on the so-called *dissipative upwinding*. Let a velocity $\boldsymbol{u}_h \in Q_h$ and a function $r_h \in Q_h$, then the (classical) upwinding reads

$$\begin{aligned} Up[r_h, \boldsymbol{u}_h] &= r_h^{\text{up}} \boldsymbol{u}_h \cdot \boldsymbol{n} = r_h^{\text{in}} [\overline{\boldsymbol{u}_h} \cdot \boldsymbol{n}]^+ + r_h^{\text{out}} [\overline{\boldsymbol{u}_h} \cdot \boldsymbol{n}]^- \\ &= \overline{r_h}\, \overline{\boldsymbol{u}_h} \cdot \boldsymbol{n} - \frac{1}{2} |\overline{\boldsymbol{u}_h} \cdot \boldsymbol{n}|\, [\![r_h]\!]\,, \end{aligned}$$

where

$$[f]^{\pm} = \frac{f \pm |f|}{2} \quad \text{and} \quad r^{\text{up}} = \begin{cases} r^{\text{in}} & \text{if } \overline{\boldsymbol{u}_h} \cdot \boldsymbol{n} \geq 0, \\ r^{\text{out}} & \text{if } \overline{\boldsymbol{u}_h} \cdot \boldsymbol{n} < 0. \end{cases}$$

The numerical flux function is defined in the following way

$$F_h(r_h, \boldsymbol{u}_h) = Up[r_h, \boldsymbol{u}_h] - h^\beta \, [\![r_h]\!] , \;\; 0 < \beta < 1.$$

Note that the term $h^\beta \, [\![r_h]\!]$ leads to an additional vanishing viscosity term in the approximation of the Euler equations. Now we proceed to the formulation of a semi-discrete finite volume method for the Euler system (1).

Definition 3 (*FLM method*) Given the initial values $(\rho_{0,h}, \boldsymbol{m}_{0,h}, E_{0,h}) \in Q_h \times Q_h \times Q_h$, we seek a piecewise constant approximation $(\rho_h, \boldsymbol{m}_h, E_h) \in Q_h \times Q_h \times Q_h$ which solves at any time $t \in (0, T]$ the following equations:

$$\begin{aligned}
&D_t \rho_h \Big|_K + \sum_{\sigma \in \partial K} \frac{|\sigma|}{|K|} F_h(\rho_h, \boldsymbol{u}_h) = 0, \\
&D_t \boldsymbol{m}_h \Big|_K + \sum_{\sigma \in \partial K} \frac{|\sigma|}{|K|} \left(\mathbf{F}_h(\boldsymbol{m}_h, \boldsymbol{u}_h) + \overline{p_h} \boldsymbol{n} \right) = h^{\alpha - 1} \sum_{\sigma \in \partial K} \frac{|\sigma|}{|K|} [\![\boldsymbol{u}_h]\!] , \\
&D_t E_h \Big|_K + \sum_{\sigma \in \partial K} \frac{|\sigma|}{|K|} \Big(F_h(E_h, \boldsymbol{u}_h) + (\overline{p_h} \, [\![\boldsymbol{u}_h]\!] + [\![p_h]\!] \, \overline{\boldsymbol{u}_h}) \cdot \boldsymbol{n} \Big) = \frac{h^{\alpha - 1}}{2} \sum_{\sigma \in \partial K} \frac{|\sigma|}{|K|} \left[\!\!\left[\boldsymbol{u}_h^2 \right]\!\!\right] ,
\end{aligned} \tag{15}$$

for any $K \in \mathbb{T}_h$.

By D_t we have denoted (continuous) time derivative; in practical implementation one can use any suitable ODE solver in order to approximate (15). In our recent work [15] we have shown that the FLM method (15) satisfies the following *invariant domain-preserving properties*, see [23, 24] where this notion was firstly introduced.

- **Positivity of the discrete density, pressure and internal energy**.
  For any fixed $h > 0$ the approximate density, pressure and internal energy remain strictly positive on any finite time interval. We refer the reader to [15, Sects. 4.3 and 4.4] for more details.
- **Discrete entropy inequality**.
  The discrete (renormalized) entropy inequality in the sense of Tadmor is satisfied, cf. [32]. More precisely, it holds that

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} \int_{\mathbb{T}_h} \rho_h \chi(s_h) \Phi_h \, \mathrm{d}x \geq \sum_{\sigma \in \Sigma_{int}} \int_\sigma Up[\rho_h \chi(s_h), \boldsymbol{u}_h][[\Phi_h]] \mathrm{d}S_x + \\
+ \sum_{\sigma \in \Sigma_{int}} \int_\sigma h^\beta \left(\overline{\nabla_\rho (\rho_h \chi(s_h))} [[\rho_h]] + \overline{\nabla_p (\rho_h \chi(s_h))} [[p_h]] \right) [[\Phi_h]] \mathrm{d}S_x ,
\end{aligned}$$

  where χ is a non-decreasing, concave, twice continuously differentiable function on $\mathbb{R}$ that is bounded from above. For the derivation and proof see [15, Sect. 3.2].
- **Minimum entropy principle**
  The discrete physical entropy $s_h = \log \left(\vartheta_h^{c_v} / \rho_h \right)$ attains its minimum at the initial time, i.e.,

$$s_h(t) \geq \underline{s}, \;\; t \geq 0, \quad \text{where} \quad -\infty < \underline{s} < \min s_h(0).$$

The entropy is either constant or produced over time, cf. [15, Sect. 4.2].

The above invariant domain-preserving properties are crucial in order to show that the approximate solutions obtained by the FLM method yield a consistent approximation to the Euler equations (1). Moreover, the discrete mass and energy conservation and some standard estimates, cf. [15], imply the stability of the FLM method, i.e., we have uniformly w.r.t. $h \to 0$

$$\|\rho_h\|_{L^\infty(0,T;L^\gamma(\Omega))} \lesssim 1, \quad \|\boldsymbol{m}_h\|_{L^\infty(0,T;L^{2\gamma/(\gamma+1)}(\Omega))} \lesssim 1, \quad \|E_h\|_{L^\infty(0,T;L^1(\Omega))} \lesssim 1.$$

In [15] the following type of nonlinear generalization of the *Lax-equivalence theorem* has been proven: Having consistent FLM method (15) for the Euler system (1), the stability of the FLM method is equivalent to its convergence. More precisely, we have shown the following results.

Theorem 1 (Existence of a DMV solution) *Let the initial data* $(\rho_{0,h}, \boldsymbol{m}_{0,h}, E_{0,h})$ *satisfy*

$$\rho_{0,h} \geq \underline{\rho} > 0, \quad E_{0,h} - \frac{1}{2}\frac{|\boldsymbol{m}_{0,h}|^2}{\rho_{0,h}} > 0.$$

Let $(\rho_h, \boldsymbol{m}_h, E_h) \in Q_h \times Q_h \times Q_h$ *be the solution of the FLM scheme* (15) *with*

$$0 < \beta < 1, \; 0 < \alpha < \frac{4}{3},$$

and there exist $\underline{\rho}, \overline{\vartheta} \in \mathbb{R}$, *such that the numerical solutions stay in a non-degenerate gas region*

$$0 < \underline{\rho} \leq \rho_h(t), \;\; \vartheta_h(t) \leq \overline{\vartheta} \textit{ for all } t \in [0, T] \textit{ uniformly for } h \to 0.$$

Then the family of approximate solutions $\{\rho_h, \boldsymbol{m}_h, E_h\}_{h \searrow 0}$ *generates a dissipative measure–valued (DMV) solution of the complete Euler system* (1) *in the sense of Definition 1.*

Further, taking into account the DMV–strong uniqueness principle proved in [5, Theorem 3.3] we obtain the desired strong convergence result.

Theorem 2 (Strong convergence of the FLM method) *In addition to the hypotheses of Theorem 1, suppose that the Euler system* (1) *admits the strong (Lipschitz–continuous) solution* $(\rho, \boldsymbol{m}, E)$ *defined on* $[0, T]$.

Then for $h \longrightarrow 0$ *it holds*

$$\rho_h \to \rho, \; \boldsymbol{m}_h \to \boldsymbol{m}, \; E_h \to E \textit{ (strongly) in } L^1((0,T) \times \Omega_h).$$

4 $\mathscr{K}$-Convergence

As demonstrated by numerical experiments, cf. [17, 19, 21], the finite volume approximations may not converge strongly. A typical example is the Kelvin-Helmholtz problem, where new and new small vortex substructures arise by refining the mesh. On the other hand, one can consider coarse-grained quantities, such as the mean or variance, averaged over different meshes. In our recent work [17] we have studied the question of strong convergence for these observable quantities. The aim of this section is to give an overview of our main results on the strong convergence without going deep into the theory of Young measures. Moreover, we would like to point out some connections to well-known and recent probabilistic concepts.

To start we recall a beautiful result of Komlós [27] on the pointwise convergence of the so-called Cèsaro averages.

Any sequence $\{F_n\}_{n=1}^{\infty}$ of uniformly L^1-bounded real valued functions on a set $Q \subset R^K$ admits a subsequence $\{F_{n_k}\}_{k=1}^{\infty}$, such that the arithmetic averages (Cèsaro averages)

$$\frac{1}{N}\sum_{k=1}^{N} F_{n_k} \text{ converge a.e. to a function } F \in L^1(Q).$$

Moreover, any subsequence of $\{F_{n_k}\}_{k=1}^{\infty}$ enjoys the same property.

We note that analogous result holds also for sequences in the reflexive L^p spaces, $1 < p < \infty$, due to the Banach-Sachs theorem. Komlós theorem has been adapted by Balder [1] who introduced the concept of $\mathscr{K}$ (Komlós)-convergence for sequences of Young measures. Applying the Young measure adapted variant of the celebrated Prokhorov theorem for random processes one obtains compactness of the empirical measures and the strong convergence in space and time of mean values and variances, see [14, 16, 17].

Theorem 3 ($\mathscr{K}$-convergence of the FLM method) *Let $\{\rho_{h_n}, \boldsymbol{m}_{h_n}, S_{h_n}\}_{n=1}^{\infty}$ be a sequence of finite volume solutions obtained by the FLM method* (15) *with $0 < \beta < 1,\ 0 < \alpha < \frac{4}{3}$. Further, assume that the FLM solutions remain in a non-degenerate gas region, i.e., there exist $\underline{\rho}, \overline{\vartheta} \in \mathbb{R}$, such that*

$$0 < \underline{\rho} \le \rho_{h_n}(t),\quad \vartheta_{h_n}(t) \le \overline{\vartheta} \text{ for all } t \in [0,T] \text{ uniformly for } h_n \to 0.$$

Then there exists a subsequence of $\{\rho_{h_n}, \boldsymbol{m}_{h_n}, S_{h_n}\}_{n=1}^{\infty}$ denoted by $\{\rho_{n_k}, \boldsymbol{m}_{n_k}, S_{n_k}\}$, for which we have

- **strong convergence of Cesàro averages to a DW solution**

$$\frac{1}{N}\sum_{k=1}^{N} \rho_{n_k} \to \rho \text{ as } N \to \infty \text{ in } L^q(0,T; L^{\gamma}(\Omega)) \text{ for any } 1 \le q < \infty,$$

$$\frac{1}{N}\sum_{k=1}^{N}\boldsymbol{m}_{n_k} \to \boldsymbol{m} \; \textit{as } N \to \infty \textit{ in } L^q(0,T; L^{\frac{2\gamma}{\gamma+1}}(\Omega; R^d)) \textit{ for any } 1 \le q < \infty,$$

$$\frac{1}{N}\sum_{k=1}^{N}S_{n_k} \to S \textit{ as } N \to \infty \textit{ in } L^q(0,T; L^{\gamma}(\Omega)) \textit{ for any } 1 \le q < \infty, \tag{16}$$

where $\rho, \boldsymbol{m}, S$ are the density, momentum and total entropy components of the DW solution in the sense of Definition 2.

- **strong convergence to a DMV solution in the Wasserstein metric**[2]

$$W_q\left[\frac{1}{N}\sum_{k=1}^{N}\delta_{[\rho_{n_k}(t,x),\boldsymbol{m}_{n_k}(t,x),S_{n_k}(t,x)]}; \mathscr{V}_{t,x}\right] \to 0 \textit{ as } N \to \infty \;\; \textit{in } L^q((0,T)\times\Omega) \tag{17}$$

for any $1 \le q < \frac{2\gamma}{\gamma+1}$. Here δ denotes the Dirac measure acting on numerical solutions $[\rho_{n_k}, \boldsymbol{m}_{n_k}, S_{n_k}]$.

- **strong convergence of the variance**

Let $\widetilde{\boldsymbol{U}} = (\tilde{\rho}, \tilde{\boldsymbol{m}}, \widetilde{S})$ and $\boldsymbol{U}_{n_k} \equiv \left(\rho_{n_k}, \boldsymbol{m}_{n_k}, S_{n_k}\right)$, then

$$\left\|\frac{1}{N}\sum_{k=1}^{N}\left|\boldsymbol{U}_{n_k} - \frac{1}{N}\sum_{j=1}^{N}\boldsymbol{U}_{n_j}\right| - \left\langle\mathscr{V}_{t,x}; \left|\widetilde{\boldsymbol{U}} - \left\langle\mathscr{V}_{t,x}; \widetilde{\boldsymbol{U}}\right\rangle\right|\right\rangle\right\|_{L^1((0,T)\times\Omega)} \qquad \textit{as } \; N \to \infty. \tag{18}$$

Theorem 3 offers an elegant way how to compute DW solutions and the statistical moments of DMV solutions in the case that the strong solution does not exist. It indicates that we still have strong convergence to the observable quantities that can be approximated directly by averaging of numerical solutions over different meshes. We refer a reader to [17] where the numerical solutions obtained by the FLM method were presented for several tests. Depending on chosen numerical experiments it may happen that the mesh-convergence of single numerical solutions is not achieved. On the other hand, the strong convergence of empirical mean values and variances was clearly shown, see [17, Figs. 1–7]. In future it will be interesting to investigate the rate of $\mathscr{K}$-convergence.

In this context we should also mention an interesting work [20], where a different probabilistic concept of the so-called statistical solutions for general multidimensional hyperbolic conservation laws has been developed. Analogously to the DMV solutions the statistical solutions are probabilistic-type solutions. In fact, they are time-parametrized probability measures satisfying an infinite set of partial differential equations consistent with the underlying hyperbolic conservation laws. Thus,

[2] We recall that the Wasserstein metric of q-th order, $q \in [1,\infty)$, is defined in the following way $W_q(\mathscr{N},\mathscr{V}) := \left\{\inf_{\pi\in\Pi(\mathscr{N},\mathscr{V})}\int_{\mathbb{R}^{d+2}\times\mathbb{R}^{d+2}}|\zeta-\xi|^q d\pi(\zeta,\xi)\right\}^{1/q}$, where $\Pi(\mathscr{N},\mathscr{V})$ is the set of probability measures on $\mathbb{R}^{d+2}\times\mathbb{R}^{d+2}$ with marginals $\mathscr{N}$ and $\mathscr{V}$.

they are the measure–valued solutions augmented by information on multi-point spatial correlations. In order to obtain strong convergence of the entropy stable finite volume solutions (or more precisely, approximate statistical solutions) to a statistical solution one however needs to assume that a special condition on an approximate scaling of structure factors holds. The latter is related to the Kolmogorov compactness criterium. On the other hand, the concept of $\mathscr{K}$-convergence based on the averaging over different meshes naturally inherits compactness. Consequently, the empirical mean values (Cèsaro averages) converge strongly to a DW solution. In future it will be interesting to generalize the concept of DMV and DW solutions to general hyperbolic conservation laws.

Acknowledgements The author wishes to thank E. Feireisl (Prague), H. Mizerová (Bratislava) and B. She (Prague) for fruitful discussions on the topic. This research was supported by the German Science Foundation under the grants TRR 146 Multiscale simulation methods for soft matter systems and TRR 165 Waves to Weather.

References

1. Balder, E.: On Prohorov's theorem for transition probabilities. Sém. Anal. Convexe **19** (1989)
2. Breit, D., Feireisl, F., Hofmanová, M.: Solution semiflow to the isentropic Euler system. Arch. Ration. Mech. Anal. (2019). https://doi.org/10.1007/s00205-019-01420-6
3. Brenner, H.: Kinematics of volume transport. Phys. A **349**, 11–59 (2005)
4. Brenner, H.: Fluid mechanics revisited. Phys. A **349**, 190–224 (2006)
5. Březina, J., Feireisl, E.: Measure-valued solutions to the complete Euler system. J. Math. Soc. Japan **70**(4), 1227–1245 (2018)
6. Chiodaroli, E., De Lellis, C., Kreml, O.: Global ill-posedness of the isentropic system of gas dynamics. Comm. Pure Appl. Math. **68**(7), 1157–1190 (2015)
7. Dafermos, C.M.: The second law of thermodynamics and stability. Arch. Rational Mech. Anal. **70**(2), 167–179 (1979)
8. De Lellis, C., Székelyhidi, Jr., L.: On admissibility criteria for weak solutions of the Euler equations. Arch. Ration. Mech. Anal. **195**(1), 225–260 (2010)
9. DiPerna, R.J.: Convergence of approximate solutions to conservation laws. Arch. Rational Mech. Anal. **82**(1), 27–70 (1983)
10. DiPerna, R.J.: Measure-valued solutions to conservation laws. Arch. Rational Mech. Anal. **88**(3), 223–270 (1985)
11. Dolejší, V., Feistauer, M.: Discontinuous Galerkin Method. Springer Series in Computational Mathematics, vol. 48. Springer, Cham (2015)
12. Eymard, R., Gallouët, T., Herbin, R.: Finite volume methods. In: Ciarlet, P.G., et al. (eds.) Handbook of Numerical Analysis, pp. 713–1020. North-Holland, Amsterdam (2000)
13. Feireisl, E., Lukáčová-Medvid'ová, M., Mizerová, H.: Convergence of finite volume schemes for the Euler equations via dissipative measure-valued solutions. Found. Comput. Math. 1–44 (2019). https://doi.org/10.1007/s10208-019-09433-z
14. Feireisl, E., Lukáčová-Medvid'ová, M., Mizerová, H.: $\mathscr{K}-$convergence as a new tool in numerical analysis. IMA J. Numer. Anal. (2019). https://doi.org/10.1093/imanum/drz045
15. Feireisl, E., Lukáčová-Medvid'ová, M., Mizerová, H.: A finite volume scheme for the Euler system inspired by the two velocities approach. Numer. Math. **144**, 89–132 (2020)
16. Feireisl, E., Lukáčová-Medvid'ová, M., Mizerová, H., She, B.: Numerical Analysis of Compressible Fluid Flows (in preparation)

17. Feireisl, E., Lukáčová-Medvid'ová, M., She, B., Wang, Y.: Computing oscillatory solutions of the Euler system via K-convergence (2019). arXiv:1910.03161
18. Feistauer, M.: Mathematical methods in fluid dynamics. In: Pitman Monographs and Surveys in Pure and Applied Mathematics, vol. 67. Longman Scientific & Technical, Harlow (1993)
19. Fjordholm, U.S., Käppeli, R., Mishra, S., Tadmor, E.: Construction of approximate entropy measure-valued solutions for hyperbolic systems of conservation laws. Found. Comput. Math. **17**(3), 763–827 (2017)
20. Fjordholm, U.S., Lye, K., Mishra, S., Weber, F.: Statistical solutions of hyperbolic systems of conservation laws: numerical approximation (2020). arXiv:1906.02536v1
21. Fjordholm, U.S., Mishra, S., Tadmor, E.: On the computation of measure-valued solutions. Acta Numer. **25**, 567–679 (2016)
22. Godlewski, E., Raviart, P.A.: Numerical approximation of hyperbolic systems of conservation laws. In: Applied Mathematical Sciences, vol. 118. Springer, New York (1996)
23. Guermond, J.L., Popov, B.: Viscous regularization of the Euler equations and entropy principles. SIAM J. Appl. Math. **74**(2), 284–305 (2014)
24. Guermond, J.L., Popov, B.: Invariant domains and first-order continuous finite element approximation for hyperbolic systems. SIAM J. Numer. Anal. **54**(4), 2466–2489 (2016)
25. Gwiazda, P., Świerczewska-Gwiazda, A., Wiedemann, E.: Weak-strong uniqueness for measure-valued solutions of some compressible fluid models. Nonlinearity **28**(11), 3873–3890 (2015)
26. Jovanović, V., Rohde, C.: Error estimates for finite volume approximations of classical solutions for nonlinear systems of hyperbolic balance laws. SIAM J. Numer. Anal. **43**(6), 2423–2449 (2006)
27. Komlós, J.: A generalization of a problem of Steinhaus. Acta Math. Acad. Sci. Hungar. **18**, 217–229 (1967)
28. Kröner, D.: Numerical Schemes for Conservation Laws. Wiley-Teubner Series Advances in Numerical Mathematics. Wiley, Ltd. (1997)
29. Kröner, D., Zajaczkowski, W.M.: Measure-valued solutions of the Euler equations for ideal compressible polytropic fluids. Math. Methods Appl. Sci. **19**(3), 235–252 (1996)
30. LeVeque, R.J.: Finite Volume Methods for Hyperbolic Problems. Cambridge Texts in Applied Mathematics. Cambridge University Press (2002)
31. Schochet, S.: Examples of measure-valued solutions. Comm. Partial Diff. Eq. **14**(5), 545–575 (1989)
32. Tadmor, E.: Entropy stability theory for difference approximations of nonlinear conservation laws and related time-dependent problems. Acta Numer. **12**, 451–512 (2003)
33. Toro, E.F.: Riemann Solvers and Numerical Methods for Fluid Dynamics, 3rd edn. Springer, Berlin (2009)

Time-Dependent Conservation Laws on Cut Cell Meshes and the Small Cell Problem

Sandra May

Abstract When solving time-dependent conservation laws on cut cell meshes, one has to face the *small cell problem*: standard explicit schemes are not stable if the time step is chosen based on the size of the background cells. Therefore, special schemes must be developed. The first part of this contribution discusses the small cell problem in detail and summarizes several existing solution approaches in the context of both finite volume (FV) schemes and discontinuous Galerkin (DG) schemes. In the second part, we present our two fundamentally different solution approaches for overcoming the small cell problem: the FV based mixed explicit implicit scheme, developed in collaboration with Berger (J. Sci. Comput. 71, pp. 919–943, 2017), and the DG based Domain-of-Dependence (DoD) stabilization, joint work with Engwer, Nüßing, and Streitbürger (ArXiv:1906.05642).

Keywords Cut cell · Small cell problem · Finite volume method · Discontinuous Galerkin method · Hyperbolic conservation law

MSC (2010) 65M08 · 65M60 · 65M12 · 65M20 · 35L02 · 35L65

1 Introduction

As a result of the objective to tackle real-world problems, grid generation has become a big challenge in today's numerical schemes for solving partial differential equations (PDEs). Simulating flow around an airplane, flow in blood vessels, the process of metal forming, collisions, or phase transitions requires to mesh very complicated geometries, which are often implicitly given or as CAD models. The generation of corresponding body-fitted meshes, which exhibit the desired properties such as shape regularity, is a sophisticated and very time-consuming process, in particular in the case of evolving geometries.

S. May (✉)
Technical University of Munich, Boltzmannstraße 3, Garching Close to Munich 85748, Germany
e-mail: sandra.may@tum.de

R. Klöfkorn et al. (eds.), *Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples*, Springer Proceedings in Mathematics & Statistics 323,
https://doi.org/10.1007/978-3-030-43651-3_3

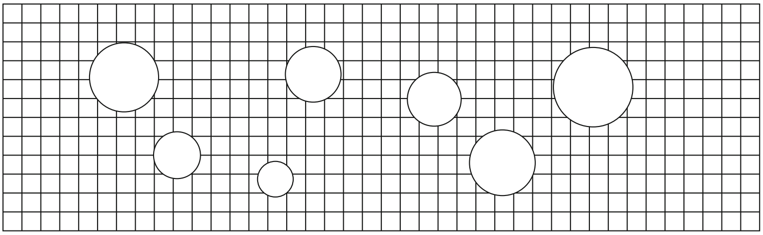


Fig. 1 Idea behind the cut cell approach: the embedded geometry is cut out of a structured background mesh (For second-order schemes, typically curved boundaries are replaced by piecewise linear approximations.)

An alternative approach is the usage of structured, e.g., Cartesian, background meshes that do not resolve the geometry combined with suitable numerical schemes that take over that role. There are different versions of so called *embedded boundary*, *immersed boundary*, or *cut cell* methods. In the following, we focus on the following approach: we consider a Cartesian background mesh and simply cut the given geometry out of the mesh, resulting in so called *cut cells* along the boundary of the embedded object as shown in Fig. 1. Cut cells can have various shapes and may in particular become arbitrarily small. Special schemes must be developed as standard schemes usually do not work well.

For solving time-dependent, first-order hyperbolic problems, probably the biggest challenge is the *small cell problem*, which is the main subject of this contribution. We describe the problem in detail in Sect. 1.1. In Sect. 2, we give a short overview of the state of the art for solving the small cell problem. Finally, in Sects. 3 and 4, we present our contributions for solving the small cell problem in the context of a finite volume (FV) scheme and of a discontinuous Galerkin (DG) scheme, respectively.

1.1 The Small Cell Problem

Typically, *explicit* time stepping schemes are used for solving first-order hyperbolic conservation laws: the associated CFL condition of having to choose $\Delta t = \mathscr{O}(h)$ for stability coincides with the standard choice for accuracy reasons (if time stepping scheme and spatial discretization are of the same order); further, limiters are significantly better understood in an explicit setting; and explicit schemes are typically cheaper than implicit schemes.

When using an explicit time stepping scheme on a cut cell mesh, one faces the so called *small cell problem*: one would like to choose the time step based on the size of the (Cartesian) background cells and use the same time step on the potentially arbitrarily small cut cells as well. This causes standard schemes to become unstable since the CFL condition is violated on small cut cells.

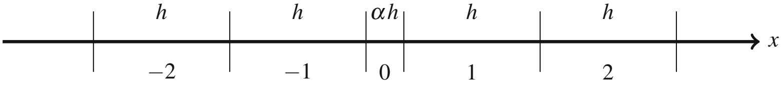


Fig. 2 1d model problem for examining the small cell problem: equidistant mesh of mesh width h with one small cell (cell 0) in the middle, which is of length αh, $\alpha \in (0, 1]$

Let us examine this problem in more detail. We consider the 1d model mesh shown in Fig. 2. Here, cell 0, which can become arbitrarily small for $\alpha \ll 1$, imitates the behavior of cut cells.

We consider the linear advection equation given by

$$u_t(x,t) + \beta u_x(x,t) = 0 \text{ in } I \times (0,T), \quad u(x,0) = u^0(x) \; \forall \, x \in I, \tag{1}$$

with $\beta > 0$ constant. We make the following definition.

Definition 1 The *model problem* refers to solving the advection equation (1) on the mesh shown in Fig. 2 on a finite domain I with periodic boundary conditions.

As numerical scheme we consider the standard FV upwind scheme given by

$$U_i^{n+1} = U_i^n - \frac{\beta \Delta t}{h_i}\left(U_i^n - U_{i-1}^n\right), \tag{2}$$

where U_i^n denotes the (piecewise constant) discrete solution on cell i at time t^n. Further, h_i denotes the cell length of cell i and Δt the time step. The scheme (2) is stable under the CFL condition $0 \leq \frac{\beta \Delta t}{h_i} \leq 1 \; \forall i$.

We set $\Delta t = \frac{\lambda h}{\beta}$ and choose $\lambda = 0.8$, independent of the size of the small cut cell 0. We determine the solution at time t^{n+1} by tracing back characteristics to the solution at time t^n as shown in Fig. 3. For cell 2, the solution at time t^{n+1} obviously depends on U_2^n and U_1^n, i.e., on the solution at time t^n in its own cell and in its left neighbor. This is the case for all cells, except for cells 0 and 1. Therefore, for a cell i with $i \neq 0, 1$, the numerical domain of dependence of the upwind scheme (2) contains the domain of dependence of the PDE.

The exact solution in cell 1 at t^{n+1}, however, depends on the solution at time t^n in cells -1, 0, and 1, which is *not* reflected in the upwind scheme (2). Therefore, we cannot expect the solution on cell 1 to be correct if the upwind scheme is used with the chosen time step.

Further, the upwind scheme for cell 0 reads

$$U_0^{n+1} = U_0^n - \frac{\beta \Delta t}{\alpha h}\left(U_0^n - U_{-1}^n\right).$$

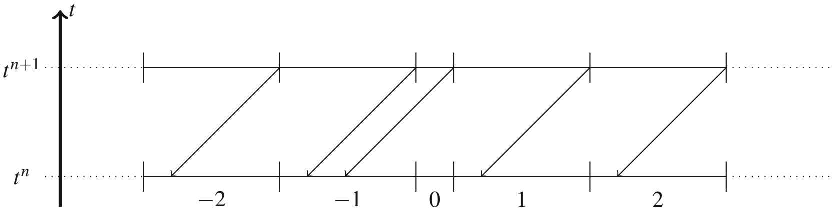


Fig. 3 Domain of dependence of the solution at time t^{n+1} in terms of the data at time t^n for $t^{n+1} - t^n = 0.8h$ and $\beta = 1$

Interpreting the solution unknowns as point values at the cell centroids, we expect for smooth (non-constant) solutions $U_0^n - U_{-1}^n = \mathscr{O}((1+\alpha)h)$ to hold. As a consequence, with $\Delta t = \mathscr{O}(h)$, the update explodes for $\alpha \ll 1$.

Therefore, to solve the small cell problem, one needs to address the following two problems:

1. The outflow neighbors of small cut cells need information from the cut cells' inflow neighbors.
2. One needs to stabilize the update on small cut cells.

2 Approaches for Solving the Small Cell Problem

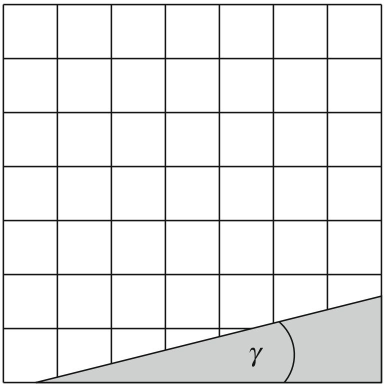
The most obvious approach for overcoming the small cell problem is *cell merging* or *cell agglomeration*: small cut cells are simply *merged* with their neighbors and thereby the problem is gone. This solution is used by a variety of authors, see, e.g., [12, 14, 16, 21, 23]. While this approach is intuitive, it is difficult to realize in a robust and fully automatic way (in 3d). Furthermore, it shifts the complexity caused by cut cells back into the mesh generation procedure.

In the following, we focus on approaches that are able to deal with potentially arbitrarily small cut cells and that solve the small cell problem in an algorithmic fashion. First though we briefly discuss the question of measuring the accuracy of a cut cell scheme.

2.1 Accuracy Considerations

We refer to the 2d ramp geometry shown in Fig. 4. Denote by N the number of cells of the background mesh in x- and y-direction, respectively. Then, the mesh width h behaves as $h = \mathscr{O}(\frac{1}{N})$. Consider a scheme that is second-order on Cartesian cells but only first-order accurate on cut cells.

Fig. 4 Simple cut cell geometry: a ramp of angle γ is cut out of a square domain



Denote by $|A|$ the size of a cut cell. The overall L^1 error is given by

$$\sum_{\text{Cart. cells}} h^2 \mathcal{O}(h^2) + \sum_{\text{cut cells}} |A| \mathcal{O}(h)$$
$$\leq \sum_{\text{Cart. cells}} \mathcal{O}(h^4) + \sum_{\text{cut cells}} \mathcal{O}(h^3).$$

The cut cell mesh contains $\mathcal{O}(N^2)$ Cartesian cells. The number of cut cells is $\mathcal{O}(N)$. Consequently, the L^1 error behaves like $\mathcal{O}(h^2)$ despite the scheme being only first-order accurate along the cut cell boundary. Therefore, it is not sufficient to just examine the L^1 error measured over the whole domain if one is also interested in the accuracy of the scheme along the cut boundary. Similar considerations apply for the L^2 norm [11].

2.2 FV Schemes for Solving the Small Cell Problem

There exist several approaches to stabilize explicit time stepping schemes on cut cell meshes in the context of FV schemes. We briefly review three approaches in the following.

2.2.1 The Flux Redistribution Method

The flux redistribution method had originally been introduced by Chern and Colella [4] and was then developed further by Colella and coworkers, see, e.g., [6, 22]. The method is designed to expand the range of influence of small cut cells to their neighbors in order to get stability: the idea is to use a stable, but non-conservative scheme on small cut cells and to restore conservation by redistributing the mass difference to the neighboring cells.

The scheme has successfully been used in complex simulations in 3d. However, due to the redistribution process, the scheme is second-order accurate in L^1 but only of first-order accurate along the cut boundary.

2.2.2 Dimensionally Split Approach

This is a more recent approach, introduced and further developed by Klein, Nikiforakis, and coworkers [8, 15]. The general idea in 1d is very similar to the idea behind the flux redistribution method but its extension to 2d and 3d relies on a dimensionally split approach. Like the flux redistribution scheme, it is second-order accurate in L^1 but only first order along the embedded boundary.

2.2.3 The *h*-box Method

The h-box method was developed by Berger, Helzel, and LeVeque [1, 2, 13]. It follows a different approach: it constructs boxes of length h, the so called h-boxes, for the flux computation on cut cell faces.

We refer to Fig. 5 and again consider solving the model problem with the upwind scheme. Due to using the upwind flux, we only construct boxes on the downwind sides of the cut faces. The box for the flux computation at face $x_{-\frac{1}{2}}$ is shown with a dotted line: it starts at $x_{-\frac{1}{2}}$ and has length h, and therefore it coincides with cell -1. The box for the flux computation at face $x_{\frac{1}{2}}$ (drawn with a dashed line) is more interesting: it includes all of cell 0 and some part of cell -1. One then reconstructs a new solution on this box, based on the solution of the underlying cells -1 and 0. This new solution is then used for the flux computation at edge $x_{\frac{1}{2}}$.

This approach addresses both issues described in Sect. 1.1:

- the outflow neighbor of cut cell 0 obtains information from the inflow neighbor of the cut cell;
- the fluxes for the update on cell 0 satisfy $F_{\frac{1}{2}} - F_{-\frac{1}{2}} = \mathscr{O}(\alpha)$; an intuitive explanation for this behavior is the observation that for $\alpha \ll 1$ the two h-boxes almost coincide, except for the two ends of length αh.

The h-box method can be proven to be fully second-order accurate in 1d and shows close to second-order accuracy in numerical experiments in 2d. As the construction of the appropriate h-boxes is fairly complex, it has not been implemented in 3d yet.

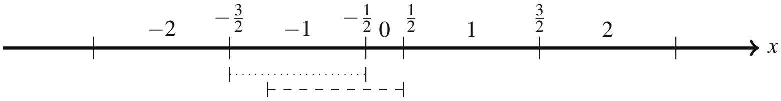


Fig. 5 Idea behind the h-box method for solving the model problem with the upwind scheme: at the cut faces boxes of length h are constructed to restore the proper domains of dependence

In Sect. 3, we present yet another approach for overcoming the small cell problem in a FV setting, which we developed together with Berger [20]. Here, the idea is to combine an explicit and an implicit time stepping scheme.

2.3 DG Schemes for Solving the Small Cell Problem

Solving the small cell problem in the context of DG schemes has not been researched to a comparable standard yet. There has been a lot of research for solving PDEs on cut cell meshes using finite element or DG schemes in recent years—but most people have focused on solving elliptic and parabolic problems with only few exceptions.

Sticko and Kreiss [24] have developed penalty terms for stabilizing the solution of the wave equation, written as a second-order equation. Gürkan and Massing [10] propose a stabilization for solving the *steady* advection-reaction equation. Both publications consider hyperbolic problems but do not solve the small cell problem for first-order hyperbolic problems.

In [7], we introduced together with Engwer, Nüßing, and Streitbürger the *Domain-of-Dependence stabilization* (DoD stabilization) to overcome the small cell problem for the time-dependent linear advection equation. We give a short summary of the scheme in Sect. 4. To the best of our knowledge, this is the first scheme for stabilizing explicit time stepping for solving time-dependent conservation laws on a cut cell mesh in the context of a DG scheme.

3 A Mixed Explicit Implicit Scheme

We introduced the FV based mixed explicit implicit scheme together with Berger [18–20].

The idea is fairly straight forward: an implicit scheme is used on cut cells for stability. In order to keep the cost low, an explicit scheme is used away from the cut cells. The switch between explicit and implicit scheme is done in a conservative and stable way. The scheme is extendable to 3d [20]. For combining a second-order accurate explicit scheme with a second-order accurate implicit scheme, the mixed scheme is provably second-order accurate in 1d in the L^1 and the L^∞ norm [17]. In 2d, we numerically observe second order in L^1 and between first and second order in L^∞ [20].

In the following, we describe how to switch between the explicit and implicit scheme and examine the accuracy of the mixed scheme in more detail.

3.1 Flux Bounding

We suggest to use what we call *flux bounding* to couple the explicit and implicit scheme. We refer to the model problem examined in Sect. 1.1. For the considered mesh, flux bounding proceeds as shown in Fig. 6:

- Step 1: Update all cells away from the cut cell using explicit fluxes F^E.
- Step 2: Update the neighborhood of the cut cell using implicit fluxes F^I on the cut faces.

In Step 2, cut cell 0 is treated fully implicitly using implicit fluxes F^I. Thereby, we solve both issues raised in Sect. 1.1: the update on cut cell 0 is stabilized and the proper domain of dependence for the update on cell 1 is used.

The Cartesian neighbors of the cut cell, which we refer to as *transition cells* use explicit fluxes on faces, which connect them to other Cartesian cells (in this case cells −2 and 2), and implicit fluxes on faces, which connect them to the cut cell (cell 0). By reusing the explicit fluxes $F^E_{-\frac{3}{2}}$ and $F^E_{\frac{3}{2}}$, which were used in Step 1 to compute U_{-2}^{n+1} and U_2^{n+1}, for the updates of cells U_{-1}^{n+1} and U_1^{n+1} in Step 2, one ensures *mass conservation*.

Example 1 We consider the upwind flux and use explicit and implicit Euler in time. Then, on all cells i with $i \neq -1, 0, 1$ the scheme corresponds to the standard upwind scheme (2). For cells $-1, 0, 1$ the scheme has the following form with $\lambda = \frac{\beta \Delta t}{h}$

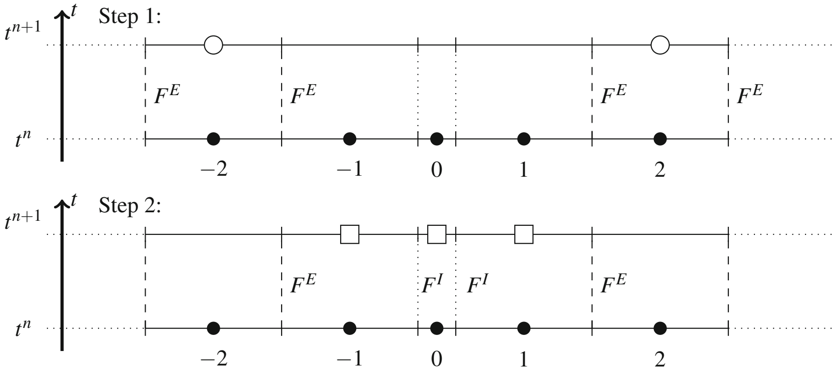


Fig. 6 Idea behind flux bounding for the time step $t^n \to t^{n+1}$: First, all cells away from the cut cell 0 are updated (indicated by the symbol '∘') using a standard explicit scheme based on the explicit flux F^E; then, the neighborhood of the cut cell is updated (indicated by the symbol '□') using implicit fluxes F^I for the cut faces

$$
\begin{aligned}
U_{-1}^{n+1} &= U_{-1}^{n} - \lambda\left(U_{-1}^{n+1} - U_{-2}^{n}\right) && \text{(transition cell)},\\
U_{0}^{n+1} &= U_{0}^{n} - \frac{\lambda}{\alpha}\left(U_{0}^{n+1} - U_{-1}^{n+1}\right) && \text{(impl. Euler)},\\
U_{1}^{n+1} &= U_{1}^{n} - \lambda\left(U_{1}^{n} - U_{0}^{n+1}\right) && \text{(transition cell)}.
\end{aligned}
$$

Concerning stability, one can show the following theorem, which corresponds to a rewording of [20, Theorem 1]. The result holds for the model problem, independent of the size of α, and employs the MUSCL scheme [5, 26] as explicit scheme. The MUSCL scheme is second-order accurate in space and time. On an equidistant mesh in one dimension, it is given by

$$
U_i^{n+1} = U_i^n - \frac{\Delta t}{h}\left(F_{i+1/2}^{n+1/2} - F_{i-1/2}^{n+1/2}\right), \quad \text{with } F_{i+1/2}^{n+1/2} = \beta\left(U_i^n + (1-\lambda)U_{x,i}^n \frac{h}{2}\right) \tag{3}
$$

for equation (1), with $U_{x,i}^n \approx \partial_x u(x_i, t^n)$ denoting suitably limited slopes.

Theorem 1 *The mixed scheme for the model problem consisting of MUSCL with minmod limiter as explicit scheme and implicit Euler with piecewise constant data as implicit scheme, coupled by means of flux bounding, is TVD for* $0 \leq \lambda \leq 1$*, if the exact solution has compact support.*

Therefore, flux bounding couples the explicit and implicit scheme in a stable and conservative way. Further, it is straight forward to extend the idea of flux bounding to 2d and 3d [20].

We note that so far flux bounding has only been applied to one stage time stepping schemes. The application to time stepping schemes with several stages is more complex. One possibility to extend the implicit zone correspondingly.

Concerning the costs of this approach: due to using implicit time stepping on cut cells, one needs to solve an implicit system in each time step that involves the cut cells and their direct Cartesian neighbors. However, as discussed in Sect. 2.1, the number of cut cells is usually one order of magnitude lower than the overall number of cells, which makes this approach inexpensive.

3.2 Accuracy

For a second-order accurate scheme, we need a second-order explicit scheme and a second-order accurate implicit scheme. We use MUSCL as explicit scheme and choose the implicit Trapezoidal rule combined with slope reconstruction in space as implicit scheme. We refer to the mixed scheme as *MUSCL-Trap*. The question is whether this combination of explicit and implicit scheme leads to a second-order accurate mixed scheme.

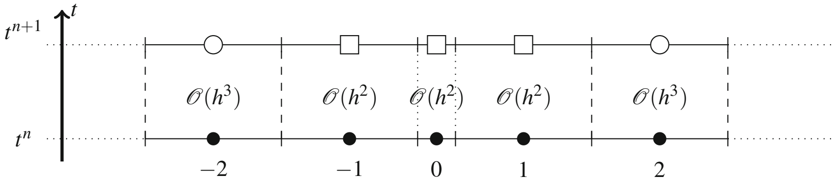


Fig. 7 The one step error for the MUSCL-Trap scheme: on cells $i = -1, 0, 1$ the error is of order $\mathscr{O}(h^2)$, on all other cells, it is $\mathscr{O}(h^3)$

3.2.1 Considerations in 1d

In Fig. 7, we show the one step error, i.e., the error for taking one time step with MUSCL-Trap on the model mesh for the linear advection equation. We observe a second-order one step error on cells -1, 0, and 1, and a third-order one step error on all other cells [17, 20]. We note that the order of the one step error is typically by one higher than the order of the overall error at time T. Therefore, this implies that the mixed scheme may not be second-order accurate.

For the reduced order of the one step error on cells -1, 0, and 1, there are two main reasons:

1. the switch in the time stepping between the explicit MUSCL scheme and the implicit Trapezoidal rule leads to an error of order $\mathscr{O}(h^2)$ on cells -1 and 1;
2. the fact that cell 0 has a different length also leads to errors of size $\mathscr{O}(h^2)$ on all three cells.

Fortunately, as is often observed on non-uniform meshes [27], the error does *not* accumulate in the standard way and together with Laakmann [17], we can show the following result for the model problem in 1d:

Theorem 2 *Let the MUSCL-Trap scheme be stable with respect to the L^1 and the L^∞ norm and use unlimited forward difference quotients for slope reconstruction. Then the scheme is second-order accurate with respect to the L^1 and the L^∞ norm for the model problem for smooth data u^0.*

This is backed up by numerical results that show full second-order accuracy for smooth test problems for the linear advection equation, measured in the L^1 and L^∞ norm.

3.2.2 Considerations in 2d

Numerical results in 2d unfortunately show worse convergence behavior with respect to the L^∞ norm. For advection along the ramp shown in Fig. 4 with varying ramp angle γ, we observe for MUSCL-Trap [20]:

- second-order convergence in the L^1 norm;
- rates between 1.3 and 1.7 in the L^∞ norm, with some dependence on the ramp angle γ.

Therefore, different to 1d, the error does accumulate in this setting. One major difference to 1d is the way that we switch between explicit and implicit time stepping. In 1d, a particle trajectory undergoes a switch from explicit to implicit and back to explicit. In 2d, this is not the case, as the transition cells are (roughly speaking) located parallel to the ramp. Therefore, one cannot expect the error caused by the switch in time stepping to cancel in the same way as it does in 1d.

One possibility to solve this problem is to use a different pair of explicit and implicit scheme. First tests show promising results.

4 DoD Stabilization

The DoD stabilization, jointly introduced with Engwer, Nüßing, and Streitbürger, is a very recent approach for solving the small cell problem in the context of DG schemes. In 2d, the stabilized scheme shows second-order convergence in the L^1 norm and rates of roughly 1.6 in the L^∞ norm if piecewise linear polynomials are used. Partially due to its short lifespan, it has not been extended to 3d yet.

4.1 Problem Setup in 1d

For examining the small cell problem in a DG setting, we use a slightly different model problem: we again consider the advection equation (1) but instead of using the mesh shown in Fig. 2, we now consider the mesh shown in Fig. 8. Again, cell 0 is the model for the behavior of a small cut cell. We use this model mesh to account for the fact that one might want to center the basis functions of the DG scheme with respect to the background mesh (instead of centering them with respect to cut cell centroids). In the following, we only stabilize the solution on the small cut cell 0, and do not stabilize the solution on the bigger cut cell 1. This corresponds to the approach taken in cell merging, where one typically also uses a slightly reduced CFL number.

We define the function space

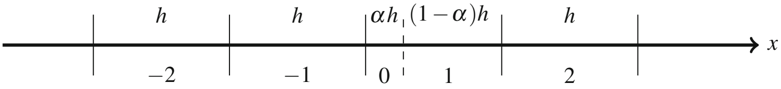


Fig. 8 1d model problem for the small cell problem in a DG setting: for a given equidistant mesh of mesh width h with N cells, one cell is split in two parts of lengths αh and $(1-\alpha)h$ with $\alpha \in (0, \frac{1}{2}]$

$$\mathcal{V}_h^k(I) := \left\{ v_h \in L^2(I) \,\middle|\, v_h \text{ is a polynomial of degree } k \text{ on cell } j = 1, \dots, N+1 \right\}. \tag{4}$$

The semi-discrete scheme, which uses the standard DG scheme with an upwind flux in space and is not yet discretized in time, is given by: Find $u_h \in \mathcal{V}_h^k(I)$ such that

$$\int_I d_t u_h(t)\, w_h\, dx + a_h^{\text{upw}}(u_h(t), w_h) = 0, \quad \forall w_h \in \mathcal{V}_h^k(I), \tag{5}$$

with the bilinear form defined as

$$a_h^{\text{upw}}(u_h, w_h) = -\sum_{j=1}^{N+1} \int_{\text{cell } j} \beta u_h \partial_x w_h dx + \sum_{j=1}^{N+1} \beta u_h(x_{j+\frac{1}{2}}^-) \left[\!\left[w_h \right]\!\right]_{j+\frac{1}{2}},$$

and the jump being given by

$$\left[\!\left[w_h \right]\!\right]_{j+\frac{1}{2}} = w_h(x_{j+\frac{1}{2}}^-) - w_h(x_{j+\frac{1}{2}}^+), \quad x_{j+\frac{1}{2}}^{\pm} = \lim_{\varepsilon \to 0^+} x_{j+\frac{1}{2}} \pm \varepsilon.$$

Using an explicit time stepping scheme and choosing Δt based on h and independent of α leads to instabilities. The goal is to stabilize the scheme by adding a suitable penalty term J_h, i.e., to define a new solution: Find $u_h \in \mathcal{V}_h^k(I)$ such that

$$\int_I d_t u_h(t)\, w_h\, dx + a_h^{\text{upw}}(u_h(t), w_h) + J_h(u_h(t), w_h) = 0, \quad \forall w_h \in \mathcal{V}_h^k(I). \tag{6}$$

4.2 The Case of Piecewise Constant Polynomials in 1d

The ghost penalty stabilization, introduced by Burman [3], has been used very successfully for stabilizing *elliptic* PDEs on cut cell meshes. For the case of $\mathcal{V}_h^0(I)$ and the considered model problem shown in Fig. 8, the penalty term for stabilizing the bilinear form a_h^{upw} on cell 0 is given by

$$J_h^{\text{GP}} = \beta\eta_1 \left[\!\left[u_h \right]\!\right]_{-\frac{1}{2}} \left[\!\left[w_h \right]\!\right]_{-\frac{1}{2}} + \beta\eta_2 \left[\!\left[u_h \right]\!\right]_{\frac{1}{2}} \left[\!\left[w_h \right]\!\right]_{\frac{1}{2}}. \tag{7}$$

Here, the parameters $\eta_1, \eta_2 \in \mathbb{R}$ are free to choose.

Let us use explicit Euler in time. Then, the scheme that we wish to stabilize simply corresponds to the upwind scheme. Straight forward computations show that it is *not* possible to choose η_1 and η_2 in such a way that the resulting scheme is monotone when the size of α is not reflected in the choice of the time step length [25]. Monotonicity however is a reasonable property to ask from a first-order scheme.

Therefore, we suggest to use the penalty term

$$J_h^{\text{DoD}}(u_h, w_h) := \beta\eta \left[\!\left[u_h \right]\!\right]_{-\frac{1}{2}} \left[\!\left[w_h \right]\!\right]_{\frac{1}{2}}. \tag{8}$$

This term introduces an additional flux between cells 0 and 1 with the size of the flux depending on the jump between the solution on cell -1 and on cell 0. Thereby, information from the inflow neighbor of cell 0 is passed to the outflow neighbor of cell 0, and the proper domain of dependence of cell 1 is restored. In addition, the stabilized flux difference for the update on cell 0 is of order $\mathscr{O}(\alpha)$ is η is chosen appropriately. Therefore, both problems described in Sect. 1.1 are taken care of. We also note that the new stabilization term J_h^{DoD} reflects the hyperbolic character of the equation of having a designated direction of information propagation, different to J_h^{GP}.

For a suitable choice of η, it is possible to design a stable and monotone scheme [7, 25]. For piecewise constant polynomials, we suggest to use $\eta = 1 - \min\left(\frac{\alpha}{\lambda}, 1\right)$ with $\lambda = \frac{\beta \Delta t}{h}$, for piecewise linear polynomials, the choice $\eta = 1 - \min\left(\frac{\alpha}{2\lambda}, 1\right)$ does better. There holds the following theorem [7, 25].

Theorem 3 *Consider the model problem with the mesh shown in Fig. 8. Use $\mathscr{V}_h^0$ in space and explicit Euler in time. Consider the stabilized scheme using the DoD-stabilization with $\eta = 1 - \min\left(\frac{\alpha}{\lambda}, 1\right)$ or $\eta = 1 - \min\left(\frac{\alpha}{2\lambda}, 1\right)$. Then, the scheme is monotone, L^1 stable, and TVD stable for $0 < \lambda < \frac{1}{2}$, independent of the size of α.*

The reduced CFL of $\lambda < \frac{1}{2}$ is mainly due to not stabilizing cell 1, which can become as small as $\frac{1}{2}h$.

Overall, the DoD stabilization has a certain similarity to the h-box method without the explicit reconstruction of the h-boxes. The stabilization is implemented by adding penalty terms, which is a natural approach for DG schemes. The h-box method and the DoD stabilization therefore mainly differ in the extension of the schemes to higher order.

4.3 The Case of Piecewise Linear Polynomials in 1d

The next step is the extension of the stabilization from piecewise constant to piecewise linear polynomials. Here, we suggest to use the following term

$$\begin{aligned} J_h^{\text{DoD}}(u_h, w_h) = &\beta\, \eta \left([\![u_h]\!]_{-\frac{1}{2}} + \alpha h\, [\![\partial_x u_h]\!]_{-\frac{1}{2}} \right) [\![w_h]\!]_{\frac{1}{2}} \\ &- \int_{\text{cell } 0} \beta\, \eta \left([\![u_h]\!]_{-\frac{1}{2}} + \frac{1}{2}\alpha h\ [\![\partial_x u_h]\!]_{-\frac{1}{2}} \right) \partial_x w_h \, dx \end{aligned} \tag{9}$$

with $\eta = 1 - \min\left(\frac{\alpha}{2\lambda}, 1\right)$. We note that the terms in the first line of J_h^{DoD} are mainly responsible for correcting the mass distribution between the cells, while the second line mainly stabilizes the gradient. An eigenvalue analysis shows stability of the resulting scheme if the standard second-order explicit SSP Runge Kutta scheme [9] is used with a CFL condition that is independent of α [7].

Numerical results in 1d for smooth initial data show that the scheme is second-order accurate in the L^1 and the L^∞ norm.

4.4 The Scheme in 2d

The extension of the scheme to 2d is more technical and we refer to the contribution [7] for details. Numerical results for the ramp test show second-order convergence in L^1 for piecewise linear polynomials. In the L^∞ norm, we observe convergence rates of roughly 1.6 for most angles γ. This behavior still needs to be examined in more detail.

We currently work on the extension of the scheme to Burgers equation and the compressible Euler equations. In 1d, the most challenging step is the extension of the fourth term in (9), which stabilizes the gradient. In the case of the linear advection equation, an eigenvalue analysis was used, which will not be feasible anymore.

References

1. Berger, M., Helzel, C.: A simplified h-box method for embedded boundary grids. SIAM J. Sci. Comput. **34**(2), A861–A888 (2012)
2. Berger, M.J., Helzel, C., LeVeque, R.J.: H-Box method for the approximation of hyperbolic conservation laws on irregular grids. SIAM J. Numer. Anal. **41**, 893–918 (2003)
3. Burman, E.: Ghost penalty. C. R. Math. **348**(21), 1217–1220 (2010)
4. Chern, I.L., Colella, P.: A conservative front tracking method for hyperbolic conservation laws. Tech. rep., Lawrence Livermore National Laboratory, Livermore, CA (1987). Preprint UCRL-97200
5. Colella, P.: A direct Eulerian MUSCL scheme for gas dynamics. SIAM J. Sci. Stat. Comput. **6**, 104–117 (1985)
6. Colella, P., Graves, D.T., Keen, B.J., Modiano, D.: A cartesian grid embedded boundary method for hyperbolic conservation laws. J. Comput. Phys. **211**, 347–366 (2006)
7. Engwer, C., May, S., Nüßing, C., Streitbürger, F.: A stabilized discontinuous Galerkin cut cell method for discretizing the linear transport equation (2019). ArXiv:1906.05642
8. Gokhale, N., Nikiforakis, N., Klein, R.: A dimensionally split cartesian cut cell method for hyperbolic conservation laws. J. Comput. Phys. **364**, 186–208 (2018)
9. Gottlieb, S., Shu, C.W.: Total-variation-diminishing Runge-Kutta schemes. Math. Comp. **67**, 73–85 (1998)
10. Gürkan, C., Massing, A.: A stabilized cut discontinuous Galerkin framework: II. Hyperbolic problems. ArXiv:1807.05634
11. Gustafsson, B.: The convergence rate for difference approximations to mixed initial boundary value problems. Math. Comp. **29**, 396–406 (1975)
12. Hartmann, D., Meinke, M., Schröder, W.: An adaptive multilevel multigrid formulation for cartesian hierarchical grid methods. Comput. Fluids **37**, 1103–1125 (2008)
13. Helzel, C., Berger, M.J., LeVeque, R.: A high-resolution rotated grid method for conservation laws with embedded geometries. SIAM J. Sci. Comput. **26**, 785–809 (2005)
14. Hunt, J.D.: An adaptive 3D cartesian approach for the parallel computation of inviscid flow about static and dynamic configurations. Ph.D. thesis, University of Michigan (2004)
15. Klein, R., Bates, K.R., Nikiforakis, N.: Well-balanced compressible cut-cell simulation of atmospheric flow. Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci. **367**, 4559–4575 (2009)
16. Krivodonova, L., Qin, R.: A discontinuous Galerkin method for solutions of the Euler equations on cartesian grids with embedded geometries. J. Comput. Sci-neth. **4**, 24–35 (2013)

17. May, S., Berger, M., Laakmann, F.: Accuracy considerations of mixed explicit implicit schemes for embedded boundary meshes. In: Eberhardsteiner, J., Schöberl, M. (eds.) PAMM. Proceedings in Applied Mathematics and Mechanics, Pamm.201900411, Wiley (2019)
18. May, S., Berger, M.: A mixed explicit implicit time stepping scheme for cartesian embedded boundary meshes. In: Fuhrmann, J., Ohlberger, M., Rohde, C. (eds.) Finite Volumes for Complex Applications VII-Methods and Theoretical Aspects, pp. 393–400. Springer International Publishing (2014)
19. May, S.: Embedded boundary methods for flow in complex geometries. Ph.D. thesis, Courant Institute of Mathematical Sciences, New York University (2013)
20. May, S., Berger, M.J.: An explicit implicit scheme for cut cells in embedded boundary meshes. J. Sci. Comput. **71**, 919–943 (2017)
21. Müller, B., Krämer-Eis, S., Kummer, F., Oberlack, M.: A high-order discontinuous Galerkin method for compressible flows with immersed boundaries. Int. J. Numer. Meth, Eng (2016)
22. Pember, R., Bell, J.B., Colella, P., Crutchfield, W., Welcome, M.L.: An adaptive cartesian grid method for unsteady compressible flow in irregular regions. J. Comput. Phys. **120**, 278–304 (1995)
23. Quirk, J.J.: An alternative to unstructured grids for computing gas dynamic flows around arbitrarily complex two-dimensional bodies. Comput. Fluids **23**(1), 125–142 (1994)
24. Sticko, S., Kreiss, G.: Higher order cut finite elements for the wave equation. J. Sci. Comput. **80**, 1867–1887 (2019)
25. Streitbürger, F., Engwer, C., May, S., Nüßing, C.: Monotonicity considerations for stabilized DG cut cell schemes for the unsteady advection equation (2019). ArXiv:1912.11933
26. van Leer, B.: Towards the ultimate conservative difference scheme. V. A second order sequel to Godunov's methods. J. Comput. Phys. **32**, 101–136 (1979)
27. Wendroff, B., White, A.B.: A supraconvergent scheme for nonlinear hyperbolic systems. Comput. Math. Appl **18**(8), 761–767 (1989)

Reactive Flow in Fractured Porous Media

Alessio Fumagalli and Anna Scotti

Abstract In this work we present a model reduction procedure to derive a hybrid-dimensional framework for the mathematical modeling of reactive transport in fractured porous media. Fractures are essential pathways in the underground which allow fast circulation of the fluids present in the rock matrix, often characterized by low permeability. However, due to infilling processes fractures may change their hydraulic properties and become barriers for the flow and creating impervious blocks in the underground. The geometrical as well as the physical properties of the fractures require a special treatment to allow the subsequent numerical discretization to be affordable and accurate. The aim of this work is to introduce a simple yet complete mathematical model to account for such diagenetic effects where chemical reactions will occlude or empty portions of the porous media and, in particular, fractures.

Keywords Reduced modeling · Hybrid-dimensional framework · Fracture porous media · Reactive flow

MSC (2010) 80A32 · 76S05 · 35Q86

1 Introduction

Fractures play a crucial role in determining fluid flow in a geological system. However, two critical parameters make the modeling of fractures challenging from both mathematical and numerical points of view. These are their apertures, which normally are several order of magnitude smaller than any other dimensions in the problem, and their microscopic structure: fractures can be open or filled by porous materials. Fractures thus can behave as highly conductive flow pathways that link distant parts

A. Fumagalli (✉) · A. Scotti
Politecnico di Milano, piazza Leonardo da Vinci 32, 20133 Milan, Italy
e-mail: Alessio.Fumagalli@polimi.it

A. Scotti
e-mail: Anna.Scotti@polimi.it

R. Klöfkorn et al. (eds.), *Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples*, Springer Proceedings in Mathematics & Statistics 323,
https://doi.org/10.1007/978-3-030-43651-3_4

of the geological system and allow for fast circulation of fluid or, on the opposite side, can be clogged preventing the flow and creating impervious parts which are not reachable. A fracture can have a portion of its core partially or fully filled and another portion empty.

In addition the fluids present in the underground can carry ions of different types that, under certain thermal conditions, might interact and react forming salts that precipitate and attach to the walls of the void spaces of the porous media. This process tends to reduce the void spaces with a direct impact on the flow properties of the system. We will call these salts "precipitate" while the ions are called "solutes". Conversely, if a precipitate is already present and the environmental conditions are such that it can dissolve we will have an increment of the porosity (i.e., void space) and the creation of ions that can be transported by the liquids. Some reference on this subject are: reactive transport on porous media at pore-scale [12, 22, 30] and at macro-scale [1, 14, 25] with experiment comparison [24]. For an micro to macro upscaling procedure see [31, 32].

In presence of fractures the situation is even more complex. The deposition or dissolution reactions can also take place inside the fractures, substantially altering their physical properties and impacting the global flow properties of the geological system. This work aims to introduce a mathematical model to accurately describe these phenomena with the technique of dimensionality reduction. This technique is rather standard in the treatment of problems with thin interfaces and frequently used in problems involving fractures. Single-phase flow [2, 4, 6, 7, 9, 11, 16, 17, 29, 33, 35, 36], two-phase flow [13, 18, 23], passive transport [3, 10, 19], and poro-elasticity [5, 8, 20, 37] are some of the physical problems which have been successfully modeled with this technique.

This work is organised as follow: in Sect. 2 the mathematical model for flow and reactive transport in a porous media is presented. Section 3 is devoted to the derivation of the reduced model to describe the fracture flow and transport via reduced models. Finally, Sect. 4 contains the conclusion of this work.

2 Reactive Flow

In this section we present the mathematical model which describes the flow in porous media and the transport of several species of ions (solutes). These may react forming a salt and the salt may in turn dissolve forming the ions. We consider two possible reactions: (i) the precipitation or crystallisation where these solutes form a solid part or (ii) a dissolution. For simplicity in the exposition, we consider only one precipitate. For more details see [25–28, 31].

The porous media is described by the domain $\Omega \subset \mathbb{R}^n$, for $n = 2, 3$. We suppose that the porous media is saturated by a single liquid phase, e.g. water, and the ions are transported by its motion. Finally, the fine scale composition of the porous media is such that a Darcy model at macroscopic scale can be applied. Our presentation is indeed given at the macro-scale. For simplicity, the initial time is assumed to be 0.

2.1 Reactive Model

Let us consider several solutes $\{U_i\}_{i=1}^N$ which are transported in the porous media by a liquid phase. As already mentioned, these solutes may react to form a solid part W. The integer N indicates the number of species of ions that are involved in the chemical reactions, which can be written as

$$\sum_{i=1}^{N} \alpha_i^+ U_i \leftrightarrow W + \sum_{i=1}^{N} \alpha_i^- U_i. \tag{1}$$

The terms $\alpha_i^\pm \geq 0$ are the stoichiometric coefficients of the reactions. Each reaction (precipitation and dissolution) is characterized by a reaction constant $\lambda^\pm$, being λ^+ the one associated with the precipitation and λ^- the one associated with the dissolution. We have $\lambda^\pm \geq 0$. We indicate with $\{u_i\}_{i=1}^N$ and w the molar concentration of the species $\{U_i\}_{i=1}^N$ and W, respectively. We have the lower bound $u_i \geq 0$, for all $i = 1, \ldots, N$, as well as $w \geq 0$. We can write the net precipitation rate associated with the reaction (1) in the following way

$$r_w(\{u_i\}_{i=1}^N) = \lambda^+ \prod_{i=1}^{N} u_i^{\alpha_i^+} - \lambda^- \prod_{i=1}^{N} u_i^{\alpha_i^-},$$

the first term being the rate of creation of solid part w and the second term the dissolution rate of the solid part in a unit time.

For simplicity, we suppose only one species of positive ions and one species of negative ions, meaning $N = 2$. Moreover, we assume electrical equilibrium, i.e., number of anions equal to the number of cations, and thus we can have $u_i = u$ for $i = 1, 2$. The previously introduced reaction rate can be simplified as

$$r_w(u) = \lambda^+ u^{\alpha^+} - \lambda^- u^{\alpha^-} \quad \text{with} \quad \alpha^\pm = \alpha_1^\pm + \alpha_2^\pm. \tag{2}$$

We consider that the dissolution of w does not depend on the presence of ions u whereas precipitation involves all the ions present. In formula (2), if we assume that $\alpha_i^- = 0$, for $i = 1, 2$, and (2) becomes

$$r_w(u) = \lambda^- \left(\frac{\lambda^+}{\lambda^-} u^{\alpha^+} - 1 \right).$$

Inspired by the previous relation, we can finally write the more abstract reaction rate law that is considered in this work. In a more compact way, we have

$$r_w(u) = \lambda[r(u) - 1],$$

with $\lambda \geq 0$ a coefficient and $r(u) = u^{\zeta}$, with ζ a positive integer. The previous models suffer of an inconsistency, in fact they do not not vanish in the limit case of $w = 0$ and might create negative values of the quantities involved. To overcome this problem, we reformulate the reaction rate as follows

$$r_w(u, w) = \begin{cases} \lambda[r(u) - 1] & \text{if } r(u) - 1 \geq 0 \\ -\lambda[r(u) - 1] & \text{if } r(u) - 1 < 0 \text{ and } w > 0 \\ 0 & \text{if } r(u) - 1 < 0 \text{ and } w \leq 0 \end{cases}. \tag{3}$$

The first condition models the case of a positive net precipitation rate, i.e., ions precipitate and form the salt w. The second condition requires that the precipitate w is present, i.e. $w > 0$, and allows for its dissolution. The last equation stops the reaction when the dissolution should occur but the precipitate is not present.

The chemical model we are considering is rather general and it does not depend on the fact that the solutes are transported in porous media. However, in our case these reactions occur each spatial point of the domain, and they can alter porosity and permeability of the porous media. The flow and transport properties of the system will be thus altered.

2.2 Porosity and Permeability Model

We assume that the solid matrix is formed by two distinct parts: the precipitate w and the solid inert part that does not react. The latter will be called solid rock. In the absence of precipitate the porous media has a prescribed or reference values of porosity and permeability due to the solid rock, named $\overline{\phi}$ and $\overline{k}$ respectively.

During the flow of chemical species in the porous media, transported by the liquid phase, a reaction may happen and the deposition of new material is assumed to be around the grains of solid rock or on a layer of precipitate already deposited. See [31] for a more detailed discussion. A graphical representation is given in Fig. 1, where we can notice that the deposition of new material alters the flow path in the porous media itself.

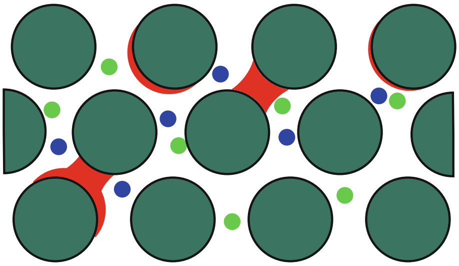


Fig. 1 Graphical representation of a porous media in presence of reactive species. The floating green and blue circles represent the anions and cations flowing in the void space between solid rock grains. The red parts are the deposited material due to the reaction

We can model the change of porosity of porous media by a law which accounts for the dependence on the precipitate concentration as follows

$$\begin{aligned} &\partial_t \phi = -\nu(\phi)\partial_t w \quad t > 0 \\ &\phi(t = 0) = \overline{\phi} \end{aligned} \tag{4}$$

where the precipitate dependent function, which represents the rate of deposition of the solute around the solid rock grains, has the properties

$$\nu \geq 0 \quad \text{and} \quad \phi = 0 \quad \Rightarrow \quad \nu = 0.$$

We notice that when $\phi = 0$ the porous media is occluded and no deposition of new material can take place. Moreover, (4) allows the porosity to increase in presence of the dissolution of precipitate, conversely the porosity decreases when the precipitate is deposited. Other model can be taken into consideration, but to keep the presentation simpler we adopt (4) where $\nu(\phi) = \eta\phi$, with η a positive constant.

Finally, also the permeability of the porous media is influenced by the reaction. In this work, we consider a Kozeny relationship between the porosity and the permeability k, namely

$$k(\phi) = \overline{k}\frac{\phi^\alpha}{\overline{\phi}^\alpha}, \tag{5}$$

with $\alpha > 0$ a rock dependent parameter. In this work we chose $\alpha = 2$. More sophisticated models can be found in, e.g., [21].

2.3 Transport Model

We introduce now the transport model, assuming that the anions and cations are transported in the porous media as passive scalars, meaning that there is not a direct influence of the scalar variable u on the given advective field $\boldsymbol{q}$. In addition, we consider a Fick's law to describe the molecular diffusivity of u in the liquid with a coefficient (or tensor) d. The model we are considering for the solute u is given, in its mixed formulation, by

$$\begin{aligned} &\left.\begin{aligned} &\boldsymbol{\chi} - \boldsymbol{q}u + \phi d \nabla u = \boldsymbol{0} \\ &\partial_t(\phi u) + \nabla \cdot \boldsymbol{\chi} + \phi r_w(u, w) = 0 \end{aligned}\right. && \text{in } \Omega \times \{t > 0\} \\ &\text{tr} u = \hat{u} && \text{on } \Gamma_{in} \times \{t > 0\} \\ &\text{tr} \phi d \nabla u \cdot \boldsymbol{n} = 0 && \text{on } \Gamma_{out} \times \{t > 0\} \\ &\text{tr} \boldsymbol{\chi} \cdot \boldsymbol{n} = 0 && \text{on } \Gamma_N \times \{t > 0\} \\ &u(t = 0) = \overline{u} && \text{in } \Omega \end{aligned} . \tag{6}$$

With χ we have denoted the total flux given by the contributions of advection and diffusion. The boundary $\partial\Omega$ of the porous media is divided into three disjoint parts Γ_{in}, Γ_{out}, and Γ_N such that $\overline{\Gamma_{in}} \cup \overline{\Gamma_{out}} \cup \overline{\Gamma_N} = \overline{\partial\Omega}$. The portion Γ_{in} represents the inflow boundary, with tr$\boldsymbol{q} \cdot \boldsymbol{n} < 0$, where the value of u is prescribed as $\hat{u}$. The part Γ_{out} is where the outflow takes place with tr$\boldsymbol{q} \cdot \boldsymbol{n} > 0$. On Γ_N we prescribe zero flux exchange with the outside, we are here assuming that tr$\boldsymbol{q} \cdot \boldsymbol{n} = 0$ in agreement with the boundary conditions of the Darcy problem, see Sect. 2.4. Other types of boundary conditions can be considered. The outward unit normal of $\partial\Omega$ is indicated with $\boldsymbol{n}$, and the operator tr indicates, in a formal way, a spacial trace operator mapping the variable at the corresponding portion of the boundary $\partial\Omega$. Finally, $\overline{u}$ represents the initial data for the solute.

Problem (6) is an advection-diffusion-reaction equation, which can degenerate due to clogging, i.e. $\phi = 0$ from (4), in some parts of the domain. The reaction term, described by the law (3), is a non-linear and non-smooth function of the solution u and of the precipitate w.

The evolution of the precipitate w follows a similar model of u, with the additional assumption that w does not move in space. All the spatial differential operators are thus removed and we obtain that the model is an ordinary differential equation in each point of Ω, namely

$$\begin{array}{ll} \partial_t(\phi w) - \phi r_w(u, w) = 0 & \text{in } \Omega \times \{t > 0\} \\ w(t = 0) = \overline{w} & \text{in } \Omega \end{array}. \tag{7}$$

The value $\overline{w}$ represents the initial condition of w in Ω. The reaction terms in the two Eqs. (6) and (7) match each other.

2.4 Darcy Model

In this part we introduce the Darcy model and its relation with the previously discussed chemical model. We are interested in computing the Darcy velocity $\boldsymbol{q}$ and the pressure field p in the porous media satisfying the following relations

$$\begin{array}{ll} \left.\begin{array}{l}\boldsymbol{q} = -k(\phi)\nabla p \\ \partial_t\phi + \nabla \cdot \boldsymbol{q} = f\end{array}\right. & \text{in } \Omega \times \{t > 0\} \\ \mathrm{tr} p = \overline{p} & \Gamma_{in} \times \{t > 0\} \\ \mathrm{tr}\boldsymbol{q} \cdot \boldsymbol{n} = \overline{q} & \Gamma_{out} \times \{t > 0\} \\ \mathrm{tr}\boldsymbol{q} \cdot \boldsymbol{n} = 0 & \Gamma_N \times \{t > 0\} \end{array}. \tag{8}$$

The division of the boundary $\partial\Omega$ into parts follows the description given in (6). The boundary value $\overline{p}$ represents the data at the inflow. The value $\overline{q}$ is the outflow flux out of Γ_{out} with the request that $\overline{q} > 0$. The condition on Γ_N is a no flow condition for that portion of boundary. By conservation we obtain that tr$\boldsymbol{q} \cdot \boldsymbol{n} < 0$ on Γ_{in}. Also in this case other type of boundary conditions can be considered, but they should be coherent with the one prescribed in the model (6).

Equation (8) is coupled with the reactive models (6) and (7) via the dependency of porosity and permeability on the solute u and precipitate w.

2.5 The Complete Model

The complete model is a six unknowns model and describes the evolution in time and space of: (i) u solute, (ii) w precipitate, (iii) ϕ porosity, (iv) k permeability, (v) $\boldsymbol{q}$ Darcy velocity, and (vi) p pressure. The equations involved are (6), (7), (4), (5), and (8) respectively for each variable or pair of variables. The resulting system is fully coupled, non-smooth and non-linear with a possible degeneracy due to vanishing porosity and permeability.

3 A Reduced Model for the Fracture

A fracture is a thin object immersed in a porous media, whose aperture is orders of magnitude smaller than any other characteristic size of the problem at hand. The fact that the fracture may exhibit higher or lower permeability with respect to the surrounding porous media increases the problem complexity and requires a proper treatment to obtain an effective and reliable model. The choice adopted here is a reduced model, meaning that the fracture is reduced as an object of lower dimension and new equations and coupling conditions are derived.

In this part, we start by presenting the interface conditions used to couple the porous media and the fracture, being the latter represented as an equi-dimensional object. Then, we present the model reduction procedure to introduce the new model and interface conditions.

3.1 Coupling Conditions for the Equi-dimensional Model

Given a parameter $\varepsilon(t)$, called the fracture aperture, which might change in time due to deposition of dissolution of new material. Following the presentation given in [16] we can define the fracture as the domain $\Omega_\gamma(t)$ given by

$$\Omega_\gamma(t) = \left\{ \boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{x} = \boldsymbol{s} + \xi(t)\boldsymbol{n}, \text{ with } \boldsymbol{s} \in \gamma \text{ and } \xi \in \left(-\frac{\varepsilon(\boldsymbol{s},t)}{2}, \frac{\varepsilon(\boldsymbol{s},t)}{2} \right) \right\}, \tag{9}$$

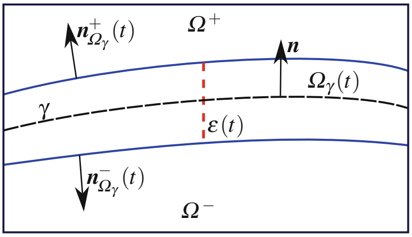


Fig. 2 Equi-dimensional representation of a fracture Ω_γ immersed in a porous media Ω

where γ is a non self-intersecting one-codimensional manifold of class C^2. We have $\varepsilon(\cdot, t) \in C^2(\gamma)$ and we assume that the fracture aperture varies slowly compared to the local coordinate system. The vector $\boldsymbol{n}$ is the normal vector of γ pointing towards one of the sides of the surrounding porous media. This choice of orientation is arbitrary and will not change the following procedure. Finally, to ease the presentation we suppose that the fracture cuts the porous media in two disjoint parts indicated with $+$ and $-$. Extension to more general cases are straightforward. An example is reported in Fig. 2.

Being Ω_γ equi-dimensional with respect to the surrounding porous media Ω it is possible to write the same equations to model the reactive transport as the one discussed in Sect. 2.5 but applied to Ω_γ instead. We will indicate with a subscript if the variable or data is referred to the porous media Ω or to the equi-dimensional representation of the fracture Ω_γ.

In addition to this, interface conditions have to be considered to couple the two problems at their common boundaries. For the transport equation (6), following [19], we have the conservation of the total flux and the continuity of the solute u, meaning

$$\begin{aligned} &\operatorname{tr}\boldsymbol{\chi}_\Omega \cdot \boldsymbol{n}_{\Omega_\gamma} = \operatorname{tr}\boldsymbol{\chi}_{\Omega_\gamma} \cdot \boldsymbol{n}_{\Omega_\gamma} \\ &\operatorname{tr}u_\Omega = \operatorname{tr}u_{\Omega_\gamma} \end{aligned} \qquad \text{on } \partial\Omega \cap \partial\Omega_\gamma, \tag{10}$$

where $\boldsymbol{n}_{\Omega_\gamma}$ is the unit normal of the boundary of Ω_γ pointing from the latter toward Ω. For the Darcy equation (8) across the interfaces we have continuity of the normal component of Darcy velocity $\boldsymbol{q}$ as well as the continuity of the pressure p. Following [15, 29, 34] we obtain

$$\begin{aligned} &\operatorname{tr}\boldsymbol{q}_\Omega \cdot \boldsymbol{n}_{\Omega_\gamma} = \operatorname{tr}\boldsymbol{q}_{\Omega_\gamma} \cdot \boldsymbol{n}_{\Omega_\gamma} \\ &\operatorname{tr}p_\Omega = \operatorname{tr}p_{\Omega_\gamma} \end{aligned} \qquad \text{on } \partial\Omega \cap \partial\Omega_\gamma. \tag{11}$$

Finally, the full equi-dimensional model for porous media-fracture system is given by equations (6), (7), (4), (5), and (8) for both Ω and Ω_γ along with the coupling conditions given by (10) and (11).

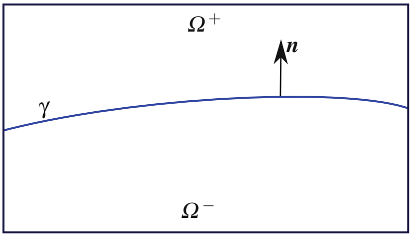


Fig. 3 Hybrid-dimensional representation of a fracture immersed in a porous media

3.2 The Reduced Variables

The model reduction procedure approximates the equi-dimensional representation of the fracture Ω_γ by its centre line γ, and derives new equations to describe the variables in γ and new interface conditions for the coupling with the surrounding porous media. Due to the previously mentioned assumptions on γ, we approximate $\boldsymbol{n}^{\pm}_{\Omega_\gamma}$ with $\pm\boldsymbol{n}$. The representation of $\Omega \subset \mathbb{R}^n$ and γ as co-dimension one object is usually named as hybrid-dimensional. See Fig. 3 as an example.

The new variables defined on γ are defined differently if they are scalar or vector fields. In the former case we define average values as

$$u_\gamma(\boldsymbol{s},t) = \frac{1}{\varepsilon(\boldsymbol{s},t)} \int_{-\frac{\varepsilon(\boldsymbol{s},t)}{2}}^{\frac{\varepsilon(\boldsymbol{s},t)}{2}} u_{\Omega_\gamma}(t) d\boldsymbol{n}(\boldsymbol{s}) \quad \text{and} \quad p_\gamma(\boldsymbol{s}) = \frac{1}{\varepsilon(\boldsymbol{s},t)} \int_{-\frac{\varepsilon(\boldsymbol{s},t)}{2}}^{\frac{\varepsilon(\boldsymbol{s},t)}{2}} p_{\Omega_\gamma}(t) d\boldsymbol{n}(\boldsymbol{s}), \tag{12}$$

where $\boldsymbol{s} \in \gamma$ and the time dependent integrals are done along the direction normal to the fracture γ. For the vector fields $\boldsymbol{\chi}$ and $\boldsymbol{q}$ we need to introduce the following projection matrices along and across the fracture, given by

$$N = \boldsymbol{n} \otimes \boldsymbol{n} \quad \text{and} \quad T = I - N.$$

Now, we can define

$$\boldsymbol{\chi}_\gamma = \int_{-\frac{\varepsilon(\boldsymbol{s},t)}{2}}^{\frac{\varepsilon(\boldsymbol{s},t)}{2}} T(\boldsymbol{s}) \boldsymbol{\chi}_{\Omega_\gamma}(t) d\boldsymbol{n}(\boldsymbol{s}) \quad \text{and} \quad \boldsymbol{q}_\gamma = \int_{-\frac{\varepsilon(\boldsymbol{s},t)}{2}}^{\frac{\varepsilon(\boldsymbol{s},t)}{2}} T(\boldsymbol{s}) \boldsymbol{q}_{\Omega_\gamma}(t) d\boldsymbol{n}(\boldsymbol{s}) \tag{13}$$

which, it is worth to notice, are not average values of fluxes or velocity, but time dependent integrals. We require that, following the idea of [2, 29], the permeability for the fracture (8) is aligned to the local coordinate system, meaning that we have

$$k_{\Omega_\gamma} = k_\gamma T + \kappa N, \tag{14}$$

where k_γ and κ are positive real numbers. Moreover, also the diffusion tensor of Eq. (6) is required to follow the similar request

$$d_{\Omega_\gamma} = d_\gamma T + \delta N,$$

where d_γ and δ are positive real numbers.

Finally, the fracture is initially considered open, meaning $\phi_{\Omega_\gamma} = 1$, and in the reduced model its role will be played by the aperture ε. We will give a specific law for its evolution. This will be part of the discussion in Sect. 3.4.

3.3 Reduced Transport Model

We describe now the procedure to derive the reduced model for the system (6). First of all the first equation of (6) is decomposed in its normal and tangential parts as

$$\begin{aligned} T\boldsymbol{\chi}_{\Omega_\gamma} - T\boldsymbol{q}_{\Omega_\gamma} u_{\Omega_\gamma} + T d_{\Omega_\gamma} \nabla u_{\Omega_\gamma} &= \mathbf{0}, \\ N\boldsymbol{\chi}_{\Omega_\gamma} - N\boldsymbol{q}_{\Omega_\gamma} u_{\Omega_\gamma} + N d_{\Omega_\gamma} \nabla u_{\Omega_\gamma} &= \mathbf{0}. \end{aligned} \tag{15}$$

Now, the tangential equation is integrated in the normal direction $\boldsymbol{n}$ across the fracture. Dropping the dependency on $\boldsymbol{s}$ and t when not needed, we obtain

$$\int_{-\frac{\varepsilon}{2}}^{\frac{\varepsilon}{2}} T\boldsymbol{\chi}_{\Omega_\gamma} d\boldsymbol{n} - \int_{-\frac{\varepsilon}{2}}^{\frac{\varepsilon}{2}} T\boldsymbol{q}_{\Omega_\gamma} u_{\Omega_\gamma} d\boldsymbol{n} + \int_{-\frac{\varepsilon}{2}}^{\frac{\varepsilon}{2}} T d_{\Omega_\gamma} \nabla u_{\Omega_\gamma} d\boldsymbol{n} = \mathbf{0},$$

having $T d_{\Omega_\gamma} = T d_\gamma$ and by assuming small variations along the thickness of the fracture of $\boldsymbol{q}_\gamma$ and u_γ, we get the following expression

$$\boldsymbol{\chi}_\gamma - \boldsymbol{q}_\gamma u_\gamma + \varepsilon d_\gamma \nabla_T u_\gamma = \mathbf{0} \qquad \text{in } \gamma \times \{t > 0\}, \tag{16}$$

where the nabla operator $\nabla_T = T\nabla$ is defined now on the tangent space of the fracture. The second equation of (15) gives the coupling conditions between the fracture γ and the surrounding porous medium, i.e. the sides Ω^+ and Ω^-. The derivation of such conditions requires the integration from the centre line of Ω_γ to its boundary, which is given, for Ω^+, by

$$\int_0^{\frac{\varepsilon}{2}} N\boldsymbol{\chi}_{\Omega_\gamma} \cdot \boldsymbol{n} d\boldsymbol{n} - \int_0^{\frac{\varepsilon}{2}} N\boldsymbol{q}_{\Omega_\gamma} u_{\Omega_\gamma} \cdot \boldsymbol{n} d\boldsymbol{n} + \int_0^{\frac{\varepsilon}{2}} N d_{\Omega_\gamma} \nabla u_{\Omega_\gamma} \cdot \boldsymbol{n} d\boldsymbol{n} = 0.$$

For the last term, we have $N d_{\Omega_\gamma} = N\delta$ and it can be approximated as

$$\int_0^{\frac{\varepsilon}{2}} N\delta \nabla u_{\Omega_\gamma} \cdot \boldsymbol{n} d\boldsymbol{n} \approx \delta(\mathrm{tr} u_{\Omega^+} - u_\gamma).$$

While for the other two terms we consider a first order one-side integration rule, along with the continuity conditions (10) and (11) for its approximation. We get

$$\int_0^{\frac{\varepsilon}{2}} N\boldsymbol{\chi}_{\Omega_\gamma}\cdot \boldsymbol{n}d\boldsymbol{n} - \int_0^{\frac{\varepsilon}{2}} N\boldsymbol{q}_{\Omega_\gamma} u_{\Omega_\gamma}\cdot \boldsymbol{n}d\boldsymbol{n} \approx \frac{\varepsilon}{2}\left(\mathrm{tr}\boldsymbol{\chi}_{\Omega^+}\cdot\boldsymbol{n} - \mathrm{tr}\boldsymbol{q}_{\Omega^+}\cdot\boldsymbol{n}\mathrm{tr}u_{\Omega^+}\right).$$

Finally, we get the coupling condition for the side of the fracture in contact with Ω^+

$$\varepsilon\left(\mathrm{tr}\boldsymbol{\chi}_{\Omega^+}\cdot\boldsymbol{n} - \mathrm{tr}\boldsymbol{q}_{\Omega^+}\cdot\boldsymbol{n}\mathrm{tr}u_{\Omega^+}\right) = 2\delta(u_\gamma - \mathrm{tr}u_{\Omega^+}) \tag{17}$$

For the other side Ω^- the derivation is similar.

The conservation equation, second of (6), is reduced following the same approach presented in [7]. Its integral form is given by

$$\partial_t \int_{\omega(t)} u_{\Omega_\gamma} d\boldsymbol{x} + \int_{\partial\omega(t)} \mathrm{tr}\boldsymbol{\chi}_{\Omega_\gamma}\cdot\boldsymbol{n}_\omega d\boldsymbol{\sigma} + \int_{\omega(t)} r_w(u_{\Omega_\gamma}, w_{\Omega_\gamma})d\boldsymbol{x} = 0 \tag{18}$$

where $\omega(t) = (l_0, l_1)\times(-\varepsilon(t)/2, \varepsilon(t)/2) \subset \Omega_\gamma(t)$ and with $(l_0, l_1)\subset\gamma$. Note that the latter does not depend on time. The vector $\boldsymbol{n}_\omega$ is the outward unit normal of ω. See Fig. 4 for a more detailed representation of the objects involved. The first and third part of the previous equation, omitting the dependency on t, are now given by

$$\int_{l_0}^{l_1}\left(\partial_t\int_{-\frac{\varepsilon}{2}}^{\frac{\varepsilon}{2}} u_{\Omega_\gamma} d\boldsymbol{n} + \int_{-\frac{\varepsilon}{2}}^{\frac{\varepsilon}{2}} r_w(u_{\Omega_\gamma}, w_{\Omega_\gamma})d\boldsymbol{n}\right)d\boldsymbol{s} = \int_{l_0}^{l_1}\partial_t(\varepsilon u_\gamma) + \varepsilon r_w(u_\gamma, w_\gamma)d\boldsymbol{s}$$

The second part of (18) becomes

$$\int_{\partial\omega}\mathrm{tr}\boldsymbol{\chi}_{\Omega_\gamma}\cdot\boldsymbol{n}_\omega d\boldsymbol{\sigma} = \int_{\partial\Omega_\gamma\cap\partial\omega}\mathrm{tr}\boldsymbol{\chi}_{\Omega_\gamma}\cdot\boldsymbol{n}_{\Omega_\gamma}d\boldsymbol{\sigma} + \int_{\partial\omega^+}\mathrm{tr}\boldsymbol{\chi}_{\Omega_\gamma}\cdot\boldsymbol{n}_\omega d\boldsymbol{\sigma} + \\ \int_{\partial\omega^-}\mathrm{tr}\boldsymbol{\chi}_{\Omega_\gamma}\cdot\boldsymbol{n}_\omega d\boldsymbol{\sigma},$$

with $\partial\omega^+ = \{l_1\}\times(-\varepsilon/2, \varepsilon/2)$ and $\partial\omega^- = \{l_0\}\times(-\varepsilon/2, \varepsilon/2)$. Now, setting $|(l_0, l_1)| \to 0$ we obtain the following expressions

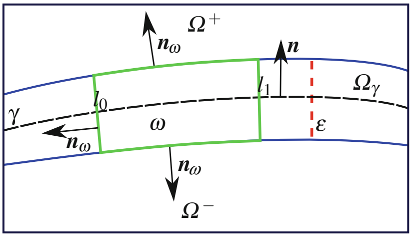


Fig. 4 Equi-dimensional representation of a fracture immersed in a porous media with the control volume ω

$$\lim_{|(l_0,l_1)|\to 0} \frac{1}{|(l_0,l_1)|}\int_{l_0}^{l_1} \partial_t(\varepsilon u_\gamma) + \varepsilon r_w(u_\gamma, w_\gamma) d\boldsymbol{s} = \partial_t(\varepsilon u_\gamma) + \varepsilon r_w(u_\gamma, w_\gamma)$$

$$\lim_{|(l_0,l_1)|\to 0} \frac{1}{|(l_0,l_1)|}\int_{\partial\omega} \mathrm{tr}\boldsymbol{\chi}_{\Omega_\gamma}\cdot \boldsymbol{n}_\omega d\boldsymbol{\sigma} = \mathrm{tr}\boldsymbol{\chi}_{\Omega_\gamma}\cdot \boldsymbol{n}_{\Omega_\gamma}|_{\frac{\varepsilon}{2}} + \mathrm{tr}\boldsymbol{\chi}_{\Omega_\gamma}\cdot \boldsymbol{n}_{\Omega_\gamma}|_{-\frac{\varepsilon}{2}} + \nabla_T\cdot\boldsymbol{\chi}_\gamma$$

by using the continuity conditions (10) the last equation becomes

$$\lim_{|(l_0,l_1)|\to 0} \frac{1}{|(l_0,l_1)|}\int_{\partial\omega} \mathrm{tr}\boldsymbol{\chi}_{\Omega_\gamma}\cdot \boldsymbol{n}_\omega d\boldsymbol{\sigma} = \mathrm{tr}\boldsymbol{\chi}_{\Omega^+}\cdot\boldsymbol{n} - \mathrm{tr}\boldsymbol{\chi}_{\Omega^-}\cdot\boldsymbol{n} + \nabla_T\cdot\boldsymbol{\chi}_\gamma.$$

Finally, the conservation equation for the transport system of the solute is reduced as

$$\begin{array}{ll} \partial_t(\varepsilon u_\gamma) + \nabla_T\cdot\boldsymbol{\chi}_\gamma + \mathrm{tr}\boldsymbol{\chi}_{\Omega^+}\cdot\boldsymbol{n} - \mathrm{tr}\boldsymbol{\chi}_{\Omega^-}\cdot\boldsymbol{n} + \varepsilon r_w(u_\gamma, w_\gamma) = 0 & \text{in } \gamma\times\{t>0\} \\ u_\gamma(t=0) = \overline{u}_\gamma & \text{in } \gamma \end{array}, \tag{19}$$

where $\overline{u}_\gamma$ is the reduced initial condition, given by $\overline{u}_\gamma = \frac{1}{\varepsilon}\int_{-\frac{\varepsilon}{2}}^{\frac{\varepsilon}{2}} \overline{u}_{\Omega_\gamma}$. The reduced boundary conditions for the transport problem are given by the following

$$\begin{array}{ll} \mathrm{tr} u_\gamma = \hat{u}_\gamma & \text{on } (\partial\gamma\cap\Gamma_{in})\times\{t>0\} \\ \mathrm{tr}\varepsilon d_\gamma\nabla_T u_\gamma\cdot\boldsymbol{n} = 0 & \text{on } (\partial\gamma\cap\Gamma_{out})\times\{t>0\}, \\ \mathrm{tr}\boldsymbol{\chi}_\gamma\cdot\boldsymbol{n} = 0 & \text{on } (\partial\gamma\cap\Gamma_N)\times\{t>0\} \end{array} \tag{20}$$

where $\hat{u}_\gamma$ is defined accordingly. We have assumed here that, for example if γ is one-dimensional, a single boundary condition is assigned to each end point $\partial\gamma$.

For the precipitate the derivation of the reduced model is rather easy since no spatial differential operators are involved. From (7) and by considering again the control volume $\omega(t)$, we get

$$\partial_t\int_{\omega(t)} w_{\Omega_\gamma} d\boldsymbol{x} - \int_{\omega(t)} r_w(u_{\Omega_\gamma}, w_{\Omega_\gamma}) d\boldsymbol{x} = 0$$

and proceeding as before we obtain

$$\lim_{|(l_0,l_1)|\to 0} \frac{1}{|(l_0,l_1)|}\int_{l_0}^{l_1} \left(\partial_t\int_{-\frac{\varepsilon}{2}}^{\frac{\varepsilon}{2}} w_{\Omega_\gamma} d\boldsymbol{n} - \int_{-\frac{\varepsilon}{2}}^{\frac{\varepsilon}{2}} r_w(u_{\Omega_\gamma}, w_{\Omega_\gamma}) d\boldsymbol{n}\right) d\boldsymbol{s} =$$

$$\lim_{|(l_0,l_1)|\to 0} \frac{1}{|(l_0,l_1)|}\int_{l_0}^{l_1} \partial_t(\varepsilon w_\gamma) - \varepsilon r_w(u_\gamma, w_\gamma) d\boldsymbol{s} = \partial_t(\varepsilon w_\gamma) - \varepsilon r_w(u_\gamma, w_\gamma).$$

Finally, the following is the reduced model for the precipitate w

$$\begin{array}{ll} \partial_t(\varepsilon w_\gamma) - \varepsilon r_w(u_\gamma, w_\gamma) = 0 & \text{in } \gamma \times \{t > 0\} \\ w_\gamma(t = 0) = \overline{w}_\gamma & \text{in } \gamma \end{array}, \tag{21}$$

where $\overline{w}_\gamma$ is the reduced initial condition for w_γ, given by $\overline{w}_\gamma = \frac{1}{\varepsilon}\int_{-\frac{\varepsilon}{2}}^{\frac{\varepsilon}{2}} \overline{w}_{\Omega_\gamma}$.

3.4 Aperture and Permeability Models

Following the ideas discussed in [26], to derive the variation of the fracture aperture by the deposition or dissolution of the solute, we consider a law similar to the one given for the porosity in (4). However, in this case since the fracture is supposed to be initially empty, free from granular material, we assume that the new material is accumulated or dissolved at the fracture boundary. See Fig. 5 for a graphical representation. We consider again a precipitate dependent law to describe the rate of aperture change, we have

$$\begin{aligned} &\partial_t\varepsilon = -\upsilon(\varepsilon)\partial_t w_\gamma \quad t > 0 \\ &\varepsilon(t = 0) = \overline{\varepsilon} \end{aligned} \tag{22}$$

where the aperture dependent model, which represents the rate of deposition of the solute around the fracture walls, has the following properties

$$\upsilon \geq 0 \quad \text{and} \quad \varepsilon = 0 \quad \Rightarrow \quad \upsilon = 0.$$

We notice that when $\varepsilon = 0$ the fracture is occluded and no deposition of new material takes place. Moreover, (22) allows the aperture to increase in presence of the dissolution of precipitate, conversely the aperture decreases when the precipitate is deposited. Other models can be taken into consideration, but to keep the presentation simpler we adopt (22) where $\upsilon(\varepsilon) = \eta_\gamma\varepsilon$, with η_γ a positive constant. In (22), the value of $\overline{\varepsilon} \geq 0$ represents the initial aperture of the fracture.

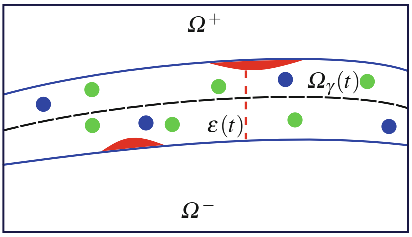


Fig. 5 Equi-dimensional representation of a fracture immersed in a porous media with dynamics of deposition and dissolution due to the chemical reaction. The solutes are the blue and green circles and the precipitate is depicted in red

The fracture permeability, both normal κ and tangential k_γ are now related to the fracture aperture by the cubic law

$$k_\gamma = \overline{k}_\gamma \frac{\varepsilon^2}{\overline{\varepsilon}} \quad \text{and} \quad \kappa = \overline{\kappa} \frac{\varepsilon^2}{\overline{\varepsilon}}. \tag{23}$$

Here $\overline{k}_\gamma$, symmetric and positive defined, and $\overline{\kappa} > 0$ are the reference tangential and normal fracture permeability, respectively.

3.4.1 Reduced Darcy Model

In this part we derive the reduced model for the Darcy system (8) written in the fracture. The steps are rather similar to the one presented for the transport equation with few modifications. The Darcy equation, first of (8) is projected on the tangential and normal directions of the fracture obtaining

$$\begin{aligned} T\boldsymbol{q}_{\Omega_\gamma} + Tk_{\Omega_\gamma} \nabla p_{\Omega_\gamma} &= \boldsymbol{0}, \\ N\boldsymbol{q}_{\Omega_\gamma} + Nk_{\Omega_\gamma} \nabla p_{\Omega_\gamma} &= \boldsymbol{0}. \end{aligned} \tag{24}$$

The first of (24) is now integrated across the normal section of the fracture, along the direction given by $\boldsymbol{n}$. We have

$$\int_{-\frac{\varepsilon}{2}}^{\frac{\varepsilon}{2}} T\boldsymbol{q}_{\Omega_\gamma} d\boldsymbol{n} + \int_{-\frac{\varepsilon}{2}}^{\frac{\varepsilon}{2}} Tk_{\Omega_\gamma} \nabla p_{\Omega_\gamma} d\boldsymbol{n} = \boldsymbol{0}.$$

From the assumption on the permeability (14) we obtain $Tk_{\Omega_\gamma} = Tk_\gamma$. By assuming small variations along the thickness of the fracture of ∇p_{Ω_γ}, we get the following relation

$$\boldsymbol{q}_\gamma + \varepsilon k_\gamma \nabla_T p_\gamma = \boldsymbol{0} \quad \text{in} \quad \gamma \times \{t > 0\}. \tag{25}$$

The second relation in (24) gives the coupling conditions between the fracture and the surrounding porous media for the Darcy problem. The approach is similar to the one already presented for the transport part, we integrate the second of (24) from 0 to $\varepsilon/2$ and we do some approximation of the integrals involved. We start with

$$\int_0^{\frac{\varepsilon}{2}} N\boldsymbol{q}_{\Omega_\gamma} \cdot \boldsymbol{n} + \int_0^{\frac{\varepsilon}{2}} Nk_{\Omega_\gamma} \nabla p_{\Omega_\gamma} \cdot \boldsymbol{n} = \boldsymbol{0}.$$

The first integral is approximated by a one-side integration rule and the coupling conditions (11) are considered to get

$$\int_0^{\frac{\varepsilon}{2}} N\boldsymbol{q}_{\Omega_\gamma}\cdot\boldsymbol{n} \approx \frac{\varepsilon}{2}\mathrm{tr}\boldsymbol{q}_{\Omega^+}\cdot\boldsymbol{n},$$

while recognising that $Nk_{\Omega_\gamma} = N\kappa$, for the second term we obtain

$$\int_0^{\frac{\varepsilon}{2}} Nk_{\Omega_\gamma}\nabla p_{\Omega_\gamma}\cdot\boldsymbol{n} \approx \kappa(\mathrm{tr}p_{\Omega^+} - p_\gamma).$$

The coupling conditions of the reduced model for the Darcy equation, for the side of the fracture in contact with Ω^+, are thus given by

$$\varepsilon\mathrm{tr}\boldsymbol{q}_{\Omega^+}\cdot\boldsymbol{n} = 2\kappa(\mathrm{tr}p_{\Omega^+} - p_\gamma). \tag{26}$$

Also in this case, for the side Ω^- the derivation of the coupling conditions are similar.

Finally, to complete the Darcy system the conservation equation for the fracture has to be reduced. Unlike the previous steps, which are in agreement with the existing literature, see for instance [7, 16, 29], this last step differs from the previous works on model reduction for fractured media because we have to account for a time dependent aperture. We consider again the control volume $\omega(t)$ given as before, and the integral form of the conservation equation is given by

$$\partial_t\int_{\omega(t)} d\boldsymbol{x} + \int_{\partial\omega(t)} \mathrm{tr}\boldsymbol{q}_{\Omega_\gamma}\cdot\boldsymbol{n}_\omega d\boldsymbol{\sigma} = \int_{\omega(t)} f d\boldsymbol{x}. \tag{27}$$

Reminding that $\omega(t) = (l_0, l_1)\times(-\varepsilon(t)/2, \varepsilon(t)/2)$, the first and last terms, dropping the dependence on t, can be expressed as

$$\lim_{|(l_0,l_1)|\to 0}\frac{1}{|(l_0,l_1)|}\int_{l_0}^{l_1}\left(\partial_t\int_{-\frac{\varepsilon}{2}}^{\frac{\varepsilon}{2}} d\boldsymbol{n} - \int_{-\frac{\varepsilon}{2}}^{\frac{\varepsilon}{2}} f d\boldsymbol{n}\right)d\boldsymbol{s} = \partial_t\varepsilon - \varepsilon f_\gamma,$$

where f_γ is the reduced source or sink term expressed by $f_\gamma = \frac{1}{\varepsilon}\int_{-\frac{\varepsilon}{2}}^{\frac{\varepsilon}{2}} f d\boldsymbol{n}$. The second term in (27) becomes

$$\int_{\partial\omega}\mathrm{tr}\boldsymbol{q}_{\Omega_\gamma}\cdot\boldsymbol{n}_\omega d\boldsymbol{\sigma} = \int_{\partial\Omega_\gamma\cap\partial\omega}\mathrm{tr}\boldsymbol{q}_{\Omega_\gamma}\cdot\boldsymbol{n}_\omega d\boldsymbol{\sigma} + \int_{\partial\omega^+}\mathrm{tr}\boldsymbol{q}_{\Omega_\gamma}\cdot\boldsymbol{n}_\omega d\boldsymbol{\sigma} + \int_{\partial\omega^-}\mathrm{tr}\boldsymbol{q}_{\Omega_\gamma}\cdot\boldsymbol{n}_\omega d\boldsymbol{\sigma}.$$

By shrinking the domain ω as $|(l_0, l_1)| \to 0$ the last relation becomes

$$\lim_{|(l_0,l_1)|\to 0}\frac{1}{|(l_0,l_1)|}\int_{\partial\omega}\mathrm{tr}\boldsymbol{q}_{\Omega_\gamma}\cdot\boldsymbol{n}_\omega d\boldsymbol{\sigma} = \mathrm{tr}\boldsymbol{q}_{\Omega_\gamma}\cdot\boldsymbol{n}_\omega|_{\frac{\varepsilon}{2}} + \mathrm{tr}\boldsymbol{q}_{\Omega_\gamma}\cdot\boldsymbol{n}_\omega|_{-\frac{\varepsilon}{2}} + \nabla_T\cdot\boldsymbol{q}_\gamma,$$

and by using the continuity condition at the fracture-porous media boundary (11) we finally get

$$\lim_{|(l_0,l_1)|\to 0} \frac{1}{|(l_0,l_1)|}\int_{\partial\omega} \mathrm{tr}\boldsymbol{q}_{\Omega_\gamma}\cdot\boldsymbol{n}_\omega d\boldsymbol{\sigma} = \mathrm{tr}\boldsymbol{q}_{\Omega^+}\cdot\boldsymbol{n} - \mathrm{tr}\boldsymbol{q}_{\Omega^-}\cdot\boldsymbol{n} + \nabla_T\cdot\boldsymbol{q}_\gamma.$$

To conclude the conservation equation for the Darcy flow is given by

$$\partial_t\varepsilon + \nabla_T\cdot\boldsymbol{q}_\gamma + \mathrm{tr}\boldsymbol{q}_{\Omega^+}\cdot\boldsymbol{n} - \mathrm{tr}\boldsymbol{q}_{\Omega^-}\cdot\boldsymbol{n} = \varepsilon f_\gamma \quad \text{in } \gamma\times\{t>0\}. \tag{28}$$

The reduced boundary conditions for the Darcy problem are given by the following

$$\begin{aligned} &\mathrm{tr}p_\gamma = \overline{p}_\gamma && (\partial\gamma\cap\Gamma_{in})\times\{t>0\}\\ &\mathrm{tr}\boldsymbol{q}_\gamma\cdot\boldsymbol{n} = \overline{q}_\gamma && (\partial\gamma\cap\Gamma_{out})\times\{t>0\},\\ &\mathrm{tr}\boldsymbol{q}_\gamma\cdot\boldsymbol{n} = 0 && (\partial\gamma\cap\Gamma_{N})\times\{t>0\}\end{aligned} \tag{29}$$

with $\overline{p}_\gamma$ and $\overline{q}_\gamma$ being defined accordingly.

3.4.2 The Complete Reduced Model

We can now summarize the full hybrid-dimensional problem, in this case we have six fields for the porous media and seven other fields for the fractures. For the former the reader can refer to the description given in Sect. 2.5 for Ω, which in the latter we have the evolution of: (i) u_γ solute, (ii) w_γ precipitate, (iii) ε aperture, (iv) κ and k_γ normal and tangential permeability, (v) $\boldsymbol{q}_\gamma$ Darcy velocity, and (vi) p_γ pressure.

For the fracture the equations involved are (16), (17), (25), and (20) for u_γ. For w_γ the problem (21), and for ε (22). For the permeabilities κ and k_γ the model given by (23). Finally, for $\boldsymbol{q}_\gamma$ and p_γ the equations (25), (26), (28), and (29).

Finally, it is important to mention that due to the model reduction procedure the aperture is now a time dependent model parameter and not any more a geometrical constraint for the problem.

4 Conclusion

In this work we have presented a reduced model for fluid flow in fractured porous media. The liquid phase flow is governed by the Darcy law and, dissolved in the liquid itself, chemical species (solutes) can react and precipitate forming a salt (or an immobile phase that fills the void spaces). Moreover, the latter can also dissolve to form solutes. The dissolution or precipitation processes can alter the porosity of the porous media, changing thus the Darcy velocity of the liquid. As mentioned, we have assumed that in the porous medium a fracture is present which may dramatically alter the flow properties of the system and thus requires an adequate model to obtain reliable and accurate results. What we have proposed is a reduced model that leads to a hybrid-dimensional framework, where the fracture is one dimensional smaller than the porous medium itself. New equations have been derived to model the physical processes in the fracture as well as the coupling conditions between the fracture itself and the surrounding porous media. The complete set of equations forms a reactive transport model in a fractured porous medium. An extension, which will be part of a future work, is the introduction of a discrete setting for the efficient solution of the proposed mathematical model.

References

1. Agosti, A., Formaggia, L., Scotti, A.: Analysis of a model for precipitation and dissolution coupled with a darcy flux. J. Math. Anal. Appl. **431**(2), 752–781 (2015). https://doi.org/10.1016/j.jmaa.2015.06.003, http://www.sciencedirect.com/science/article/pii/S0022247X15005466
2. Alboin, C., Jaffré, J., Roberts, J.E., Serres, C.: Modeling fractures as interfaces for flow and transport in porous media. In: Fluid Flow and Transport in Porous Media: Mathematical and Numerical Treatment. South Hadley, MA (2001). Contemp. Math. **295**, 13–24. Amer. Math. Soc., Providence, RI (2002)
3. Ambartsumyan, I., Khattatov, E., Nguyen, T., Yotov, I.: Flow and transport in fractured poroelastic media. GEM—Int. J. Geomath. **10**(1), 11 (2019). https://doi.org/10.1007/s13137-019-0119-5
4. Antonietti, P.F., Formaggia, L., Scotti, A., Verani, M., Verzotti, N.: Mimetic finite difference approximation of flows in fractured porous media. ESAIM: M2AN **50**(3), 809–832 (2016). https://doi.org/10.1051/m2an/2015087
5. Berge, R., Berre, I., Keilegavlen, E., Wohlmuth, B.: Finite volume discretization for poroelastic media with fractures modeled by contact mechanics. Tech. Rep. arXiv:1904.11916 [math.NA] (2019)
6. Berre, I., Doster, F., Keilegavlen, E.: Flow in fractured porous media: a review of conceptual models and discretization approaches. Transp. Porous Media **130**(1), 215–236 (2019). https://doi.org/10.1007/s11242-018-1171-6
7. Boon, W.M., Nordbotten, J.M., Yotov, I.: Robust discretization of flow in fractured porous media. SIAM J. Numer. Anal. **56**(4), 2203–2233 (2018). https://doi.org/10.1137/17M1139102
8. Bukac, M., Yotov, I., Zunino, P.: Dimensional model reduction for flow through fractures in poroelastic media. ESAIM: Mathematical Math. Model. Numer. Anal. **51**(4), 1429–1471 (2017). https://doi.org/10.1051/m2an/2016069
9. Chave, F., Di Pietro, D.A., Formaggia, L.: A hybrid high-order method for darcy flows in fractured porous media. SIAM J. Sci. Comput. **40**(2), A1063–A1094 (2018). https://doi.org/10.1137/17M1119500
10. Chave, F., Di Pietro, D.A., Formaggia, L.: A hybrid high-order method for passive transport in fractured porous media. GEM—Int. J. Geomath. **10**(1), 12 (2019). https://doi.org/10.1007/s13137-019-0114-x
11. D'Angelo, C., Scotti, A.: A mixed finite element method for Darcy flow in fractured porous media with non-matching grids. Math. Model. Numer. Anal. **46**(02), 465–489 (2012). https://doi.org/10.1051/m2an/2011148
12. van Duijn, C., Pop, I.S.: Crystal dissolution and precipitation in porous media: pore scale analysis. J. für die reine und angewandte Mathematik (Crelle's J.) **577**, 171–211 (2004). https://doi.org/10.1515/crll.2004.2004.577.171

13. Elyes, A., Jérôme, J., Roberts, J.E.: A 3-D reduced fracture model for two-phase flow in porous media with a global pressure formulation. In: MAMERN VI. Pau, France (2015). https://hal.inria.fr/hal-01119986
14. Emmanuel, S., Berkowitz, B.: Mixing-induced precipitation and porosity evolution in porous media. Adv. Water Resour. **28**(4), 337–344 (2005). https://doi.org/10.1016/j.advwatres.2004.11.010, http://www.sciencedirect.com/science/article/pii/S030917080400212X
15. Faille, I., Fumagalli, A., Jaffré, J., Roberts, J.E.: Model reduction and discretization using hybrid finite volumes of flow in porous media containing faults. Comput. Geosci. **20**(2), 317–339 (2016). https://doi.org/10.1007/s10596-016-9558-3, https://link.springer.com/article/10.1007/s10596-016-9558-3
16. Formaggia, L., Fumagalli, A., Scotti, A., Ruffo, P.: A reduced model for Darcy's problem in networks of fractures. ESAIM: Math. Model. Numer. Anal. **48**, 1089–1116 (2014). https://doi.org/10.1051/m2an/2013132, https://www.esaim-m2an.org/articles/m2an/abs/2014/04/m2an130132/m2an130132.html
17. Fumagalli, A., Keilegavlen, E.: Dual virtual element methods for discrete fracture matrix models. Oil Gas Sci. Technol.—Revue d'IFP Energies nouvelles **74**(41), 1–17 (2019). https://doi.org/10.2516/ogst/2019008
18. Fumagalli, A., Scotti, A.: A numerical method for two-phase flow in fractured porous media with non-matching grids. Adv. Water Resour. **62, Part C**(0), 454–464 (2013). https://doi.org/10.1016/j.advwatres.2013.04.001, https://www.sciencedirect.com/science/article/pii/S0309170813000523 (Computational Methods in Geologic CO_2 Sequestration)
19. Fumagalli, A., Scotti, A.: A reduced model for flow and transport in fractured porous media with non-matching grids. In: Cangiani, A., Davidchack, R.L., Georgoulis, E., Gorban, A.N., Levesley, J., Tretyakov, M.V. (eds.) Numerical Mathematics and Advanced Applications 2011, pp. 499–507. Springer, Berlin, Heidelberg (2013). https://doi.org/10.1007/978-3-642-33134-3_53
20. Ganis, B., Girault, V., Mear, M., Singh, G., Wheeler, M.: Modeling fractures in a poro-elastic medium. Oil Gas Sci. Technol.—Revue d'IFP Energies nouvelles **69**(4), 515–528 (2014)
21. Hommel, J., Coltman, E., Class, H.: Porosity-permeability relations for evolving pore space: a review with a focus on (bio-)geochemically altered porous media. Transp. Porous Media **124**(2), 589–629 (2018). https://doi.org/10.1007/s11242-018-1086-2
22. Hornung, U., Jäger, W., Mikelić, A.: Reactive transport through an array of cells with semi-permeable membranes. ESAIM: Math. Model. Numer. Anal.—Modélisation Mathématique et Analyse Numérique **28**(1), 59–94 (1994)
23. Jaffré, J., Mnejja, M., Roberts, J.E.: A discrete fracture model for two-phase flow with matrix-fracture interaction. Procedia Comput. Sci. **4**, 967–973 (2011). https://doi.org/10.1016/j.procs.2011.04.102, http://www.sciencedirect.com/science/article/pii/S1877050911001608
24. Katz, G.E., Berkowitz, B., Guadagnini, A., Saaltink, M.W.: Experimental and modeling investigation of multicomponent reactive transport in porous media. J. Contam. Hydrol. **120–121**, 27–44 (2011). https://doi.org/10.1016/j.jconhyd.2009.11.002, http://www.sciencedirect.com/science/article/pii/S0169772209001582 (Reactive Transport in the Subsurface: Mixing, Spreading and Reaction in Heterogeneous Media)
25. Knabner, P., van Duijn, C., Hengst, S.: An analysis of crystal dissolution fronts in flows through porous media. Part 1: Compatible boundary conditions. Adv. Water Resour. **18**(3), 171–185 (1995). https://doi.org/10.1016/0309-1708(95)00005-4, http://www.sciencedirect.com/science/article/pii/0309170895000054
26. Kumar, K., van Noorden, T.L., Pop, I.S.: Effective dispersion equations for reactive flows involving free boundaries at the microscale. Multiscale Model. Simul. **9**(1), 29–58 (2011). https://doi.org/10.1137/100804553
27. Kumar, K., Pop, I.S., Radu, F.A.: Convergence analysis of mixed numerical schemes for reactive flow in a porous medium. SIAM J. Numer. Anal. **51**(4), 2283–2308 (2013). https://doi.org/10.1137/120880938
28. Kumar, K., Pop, I.S., Radu, F.A.: Convergence analysis for a conformal discretization of a model for precipitation and dissolution in porous media. Numerische Mathematik **127**(4), 715–749 (2014). https://doi.org/10.1007/s00211-013-0601-1

29. Martin, V., Jaffré, J., Roberts, J.E.: Modeling fractures and barriers as interfaces for flow in porous media. SIAM J. Sci. Comput. **26**(5), 1667–1691 (2005). https://doi.org/10.1137/S1064827503429363
30. van Noorden, T.: Crystal precipitation and dissolution in a thin strip. Eur. J. Appl. Math. **20**(1), 69–91 (2009). https://doi.org/10.1017/S0956792508007651
31. van Noorden, T.L.: Crystal precipitation and dissolution in a porous medium: effective equations and numerical experiments. Multiscale Model. Simul. **7**(3), 1220–1236 (2009). https://doi.org/10.1137/080722096
32. Ray, N., Oberlander, J., Frolkovic, P.: Numerical investigation of a fully coupled micro-macro model for mineral dissolution and precipitation. Comput. Geosci. **23**(5), 1173–1192 (2019). https://doi.org/10.1007/s10596-019-09876-x
33. Sandve, T.H., Berre, I., Nordbotten, J.M.: An efficient multi-point flux approximation method for Discrete Fracture-Matrix simulations. J. Comput. Phys. **231**(9), 3784–3800 (2012). https://doi.org/10.1016/j.jcp.2012.01.023, http://www.sciencedirect.com/science/article/pii/S0021999112000447
34. Schwenck, N., Flemisch, B., Helmig, R., Wohlmuth, B.: Dimensionally reduced flow models in fractured porous media: crossings and boundaries. Comput. Geosci. **19**(6), 1219–1230 (2015). https://doi.org/10.1007/s10596-015-9536-1
35. Scotti, A., Formaggia, L., Sottocasa, F.: Analysis of a mimetic finite difference approximation of flows in fractured porous media. ESAIM: M2AN (2017). https://doi.org/10.1051/m2an/2017028
36. Stefansson, I., Berre, I., Keilegavlen, E.: Finite-volume discretisations for flow in fractured porous media. Transp. Porous Media **124**(2), 439–462 (2018). https://doi.org/10.1007/s11242-018-1077-3
37. Ucar, E., Keilegavlen, E., Berre, I., Nordbotten, J.M.: A finite-volume discretization for deformation of fractured media. Comput. Geosci. **22**(4), 993–1007 (2018). https://doi.org/10.1007/s10596-018-9734-8

Numerical Schemes for Semiconductors Energy-Transport Models

Marianne Bessemoulin-Chatard, Claire Chainais-Hillairet, and Hélène Mathis

Abstract We introduce some finite volume schemes for unipolar energy-transport models. Using a reformulation in dual entropy variables, we can show the decay of a discrete entropy with control of the discrete entropy dissipation.

Keywords Energy-tranport model · Finite volumes · Entropy method

MSC (2010): 65M08, 65M12, 35K20

1 Energy-Transport Models

Presentation

In this article, we are interested in the discretization of unipolar energy-transport models for semiconductor devices. Such models describe the flow of electrons through a semiconductor crystal, influenced by diffusive, electrical and thermal effects. As they have a drift-diffusion form, they remain simpler than hydrodynamic equations or semiconductor Boltzmann equations. As explained for example in [17] (and the references therein), these energy-transport models can be derived from the Boltzmann equation by the moment method.

The unipolar energy-transport system consists in two continuity equations for the electron density ρ_1 and the internal energy density ρ_2, coupled with a Poisson

M. Bessemoulin-Chatard (✉) · H. Mathis
Laboratoire de Mathématiques Jean Leray, Université de Nantes & CNRS UMR 6629,
BP 92208, 44322 Nantes Cedex 3 Nantes, France
e-mail: marianne.bessemoulin@univ-nantes.fr

H. Mathis
e-mail: helene.mathis@univ-nantes.fr

C. Chainais-Hillairet
University of Lille, CNRS, UMR 8524-Laboratoire Paul Painlevé, 59000 Lille, France
e-mail: claire.chainais@univ-lille.fr

R. Klöfkorn et al. (eds.), *Finite Volumes for Complex Applications IX - Methods,
Theoretical Aspects, Examples*, Springer Proceedings in Mathematics & Statistics 323,
https://doi.org/10.1007/978-3-030-43651-3_5

equation describing the electrical potential V. Following the framework adopted in [6], we consider that the electron and energy densities are defined as functions of the entropy variables $u_1 = \mu/T$ and $u_2 = -1/T$ where μ is the chemical potential and T the temperature. We set $u = (u_1, u_2)$.

Let Ω be an open bounded subset of $\mathbb{R}^d$ ($d \geq 1$) describing the geometry of the considered semiconductor device and let $T_{\max} > 0$ be a finite time horizon. The energy transport model writes in $\Omega \times (0, T_{\max})$

$$\partial_t \rho_1(u) + \operatorname{div} J_1 = 0, \tag{1a}$$

$$\partial_t \rho_2(u) + \operatorname{div} J_2 = \nabla V \cdot J_1 + W(u), \tag{1b}$$

$$-\lambda^2 \Delta V = C(x) - \rho_1(u), \tag{1c}$$

where J_1 and J_2 are respectively the electron and energy current densities, $\nabla V \cdot J_1$ corresponds to a Joule heating term and $W(u)$ is an energy relaxation term. The doping profile $C(x)$ describes the fixed charged background and λ is the rescaled Debye length. The electron and energy current densities are given by:

$$J_1 = -L_{11}(u)(\nabla u_1 + u_2 \nabla V) - L_{12}(u) \nabla u_2, \tag{2a}$$

$$J_2 = -L_{21}(u)(\nabla u_1 + u_2 \nabla V) - L_{22}(u) \nabla u_2, \tag{2b}$$

where $\mathbb{L}(u) = (L_{ij}(u))_{1 \leq i,j \leq 2}$ is a symmetric uniformly positive definite matrix.

The system (1)–(2) is supplemented with an initial condition $u_0 = (u_{1,0}, u_{2,0})$ and with mixed boundary conditions. There are Dirichlet boundary conditions on the ohmic contacts and homogeneous Neumann boundary conditions on insulating segments. More precisely, we assume that Ω is an open bounded polygonal (or polyhedral) subset of $\mathbb{R}^d$, such that its boundary $\partial\Omega$ is split into $\partial\Omega = \Gamma^D \cup \Gamma^N$, with $\Gamma^D \cup \Gamma^N = \emptyset$ and $\mathrm{m}_{d-1}(\Gamma^D) > 0$. We denote by $\mathbf{n}$ the normal to $\partial\Omega$ outward Ω. The boundary conditions write

$$u_1 = u_1^D, \ u_2 = u_2^D, \ V = V^D \text{ on } \Gamma^D \times [0, T_{\max}], \tag{3a}$$

$$J_1 \cdot \mathbf{n} = J_2 \cdot \mathbf{n} = \nabla V \cdot \mathbf{n} \quad \text{on } \Gamma^N \times [0, T_{\max}]. \tag{3b}$$

We assume that the Dirichlet boundary conditions u_1^D, u_2^D and V^D do not depend on time and are the traces of some functions defined on the whole domain Ω, still denoted by u_1^D, u_2^D and V^D. Moreover, we assume that $u_2^D < 0$ is constant on Γ^D and that the energy relaxation term $W(u)$ verifies, for all $u \in \mathbb{R}^2$ and $u_2^D < 0$,

$$W(u)(u_2 - u_2^D) \leq 0. \tag{4}$$

The main results on the energy-transport model (1)–(3) are presented in [15]: existence of solutions to the transient system, regularity, uniqueness and existence and uniqueness of steady-states. The main assumptions needed on the function $u \mapsto \rho(u) = (\rho_1(u), \rho_2(u))$ for the existence result are the following:

$$\rho \in W^{1,\infty}(\mathbb{R}^2; \mathbb{R}^2), \tag{5a}$$

$$\exists c_0 > 0 \text{ such that } (\rho(u) - \rho(v)) \cdot (u - v) \geq c_0 |u - v|^2 \quad \text{for } u, v \in \mathbb{R}^2, \tag{5b}$$

$$\exists \chi \in C^1(\mathbb{R}^2; \mathbb{R}) \text{ strictly convex such that } \rho = \nabla_u \chi. \tag{5c}$$

These hypotheses are rather hard to satisfy in the applications (see Sect. 4), as well as the hypothesis on uniform positive definiteness of the diffusion matrix $\mathbb{L}$. Existence results for physically more realistic diffusion matrices (only positive semi-definite) are established in [10, 12] for the stationary model and in [4, 5] for the transient system, but only in the case of data close to thermal equilibrium. More recently, existence of solutions has been proved in a simplified degenerate case, namely for a model with a simplified temperature equation in [16] and for vanishing electric fields (avoiding the coupling with Poisson equation) in [20].

The existence result due to Degond, Génieys and Jüngel [6, 15] is based on a reformulation of the system in terms of dual entropy variables. This reformulation symmetrizes the system and allows to apply an entropy method. Since we are going to adapt the results of [6] to the discrete framework, let us now introduce the system reformulated in terms of dual entropy variables and give the outline of the entropy structure.

The System in Dual Entropy Variables

The key point of the analysis of the primal model (1)–(2) is to use another set of variables which symmetrizes the problem, see [6]. Let us define the so-called dual entropy variables $w = (w_1, w_2)$ (w_1 is an electrochemical potential):

$$w_1 = u_1 + u_2 V, \tag{6a}$$

$$w_2 = u_2. \tag{6b}$$

Through this change of variables, the problem (1)–(2) is equivalent to

$$\partial_t b_1(w, V) + \text{div} I_1(w, V) = 0, \tag{7a}$$

$$\partial_t b_2(w, V) + \text{div} I_2(w, V) = \widetilde{W}(w) - \partial_t V b_1(w, V), \tag{7b}$$

$$-\lambda^2 \Delta V = C - b_1(w, V), \tag{7c}$$

where the function $b(w, V) = (b_1(w, V), b_2(w, V))$ is related to ρ and V by

$$b_1(w, V) = \rho_1(u), \quad b_2(w, V) = \rho_2(u) - V\rho_1(u), \tag{8}$$

and the new energy relaxation term is defined by $\widetilde{W}(w) = W(u)$. Moreover, the symmetrized currents are given by $I_1 = J_1$ and $I_2 = J_2 - V J_1$, which leads to

$$I_1(w,V) = -D_{11}(w,V)\nabla w_1 - D_{12}(w,V)\nabla w_2, \tag{9a}$$

$$I_2(w,V) = -D_{21}(w,V)\nabla w_1 - D_{22}(w,V)\nabla w_2, \tag{9b}$$

where the new diffusion matrix $\mathbb{D}(w,V) = (D_{ij}(w,V))_{1\leq i,j\leq 2}$ is defined by

$$\mathbb{D}(w,V) = \mathbb{P}(V)^T \mathbb{L}(u)\mathbb{P}(V), \quad \text{with } \mathbb{P}(V) = \begin{pmatrix} 1 & -V \\ 0 & 1 \end{pmatrix}. \tag{10}$$

It is therefore clear that the new diffusion matrix $\mathbb{D}$ is also symmetric and uniformly positive definite.

Entropy Structure

We recall in this section the entropy/entropy-dissipation property satisfied by the energy-transport model (1)–(3) established in [6]. The entropy function is defined by

$$S(t) = \int_\Omega \left[\rho(u)\cdot(u-u^D) - (\chi(u) - \chi(u^D))\right] dx - \frac{\lambda^2}{2} u_2^D \int_\Omega |\nabla(V - V^D)|^2 dx. \tag{11}$$

Since $u_2^D < 0$ and χ is a convex function such that $\rho = \nabla_u \chi$, $S(t)$ is nonnegative for all $t \geq 0$.

In addition to the hypotheses already given above, we assume that the Dirichlet boundary conditions are at thermal equilibrium, namely

$$\nabla w_1^D = \nabla w_2^D = 0. \tag{12}$$

Then the entropy function satisfies the following identity:

$$\frac{d}{dt} S(t) = -\int_\Omega (\nabla w)^T \mathbb{D} \nabla w + \int_\Omega W(u)(u_2 - u_2^D) \leq 0. \tag{13}$$

The proof of (13) is given in [6], even for more general boundary conditions.

2 Numerical Schemes

Different kind of numerical schemes have already been designed for the energy-transport systems, essentially for the stationary systems: finite difference schemes in [11, 19], finite element schemes in [7, 14]. We also refer to [3] for DDFV (Discrete Duality Finite Volume) schemes for the evolutive case. Up to our knowledge, there exists no convergence analysis of these numerical schemes. In this paper, we are interested in the design and the analysis of some finite volume schemes for the system (1)–(3), with two-point flux approximations (TPFA) of the numerical fluxes. We pay attention, while building the scheme, on the possibility of adapting the entropy method to the discrete setting. This will be crucial in order to fulfill the convergence analysis of the scheme.

Mesh and Notations

Let $\Delta t > 0$ be the time step and set $t^n = n\Delta t$ for all $n \geq 0$. We now define the mesh of the domain Ω. It is given by a family $\mathscr{T}$ of open polygonal (or polyhedral in 3D) control volumes, a family $\mathscr{E}$ of edges (or faces), and a family $\mathscr{P} = (x_K)_{K\in\mathscr{T}}$ of points. The schemes we will consider are based on two-points flux approximations, so that we assume that the mesh is admissible in the sense of [9, Definition 9.1].

In the set of edges $\mathscr{E}$, we distinguish the interior edges $\sigma = K|L \in \mathscr{E}_{int}$ and the boundary edges $\sigma \in \mathscr{E}_{ext}$. Due to the mixed boundary conditions, we have to distinguish the edges included in Γ^D from the edges included in Γ^N: $\mathscr{E}_{ext} = \mathscr{E}^D \cup \mathscr{E}^N$. For a control volume $K \in \mathscr{T}$, we define $\mathscr{E}_K$ the set of its edges, which is also split into $\mathscr{E}_K = \mathscr{E}_{K,int} \cup \mathscr{E}_K^D \cup \mathscr{E}_K^N$.

In the sequel, we denote by d the distance in $\mathbb{R}^d$ and m the measure in $\mathbb{R}^d$ or $\mathbb{R}^{d-1}$. For all $\sigma \in \mathscr{E}$, we define $\mathrm{d}_\sigma = d(x_K, x_L)$ if $\sigma = K|L \in \mathscr{E}_{int}$ and $\mathrm{d}_\sigma = d(x_K, \sigma)$ if $\sigma \in \mathscr{E}_{ext}$, with $\sigma \in \mathscr{E}_K$. Then the transmissibility coefficient is defined by $\tau_\sigma = \mathrm{m}(\sigma)/\mathrm{d}_\sigma$, for all $\sigma \in \mathscr{E}$.

A finite volume scheme with two-point flux approximation provides, for an unknown v, a vector $\mathbf{v} = (v_K)_{K\in\mathscr{T}} \in \mathbb{R}^\theta$ (with $\theta = \mathrm{Card}(\mathscr{T})$) of approximate values on each cells. We can associate to $\mathbf{v}$ a piecewise constant function, still denoted $\mathbf{v}$. For all $K \in \mathscr{T}$ and all $\sigma \in \mathscr{E}_K$, we define

$$v_{K,\sigma} = \begin{cases} v_L & \text{if } \sigma = K|L \in \mathscr{E}_{int}, \\ v_\sigma^D & \text{if } \sigma \in \mathscr{E}^D, \\ v_K & \text{if } \sigma \in \mathscr{E}^N, \end{cases}$$

and

$$D_{K,\sigma}\mathbf{v} = v_{K,\sigma} - v_K, \qquad D_\sigma \mathbf{v} = |D_{K,\sigma}\mathbf{v}|.$$

Schemes in Primal and Dual Entropy Variables

Our aim is to design a scheme for the energy transport model in the primal entropy variables (1)–(3). This scheme must lead to an equivalent scheme for the system written in the dual entropy variables (7)–(9). Indeed, in this case, it will be possible to apply the entropy method at the discrete level. This step is crucial as it brings *a priori* estimates on the sequences of approximate solutions, leading to compactness results. Moreover, it also permits to prove existence of a solution to the scheme.

One main difficulty in writing a TPFA scheme for the energy-transport model (1)–(3) comes from the approximation of the Joule heating term $\nabla V \cdot J_1$. One possibility would be to apply the technique developed in [1], and further used in [8, 18], to discretize de Joule heating term. However, with such discretization, the rewriting of the scheme in dual entropy variables is not straightforward. Therefore, following [2], we propose an approximation of the Joule heating term which is based on its following reformulation:

$$\nabla V \cdot J_1 = \operatorname{div}(V J_1) - V \operatorname{div} J_1.$$

Let us now turn to the definition of the scheme for the model (1)–(3). Initial and Dirichlet boundary conditions are discretized as usually: $u_{i,K}^0$ is the mean value of $u_{i,0}$ over K for all $K \in \mathscr{T}$ and $i = 1, 2$, $u_{i,\sigma}^D$ and V_σ^D are the mean values of u_i^D for $i = 1, 2$ and V^D for $\sigma \in \mathscr{E}^D$ and we define:

$$u_{1,\sigma}^n = u_{1,\sigma}^D, \quad u_{2,\sigma}^n = u_{2,\sigma}^D, \quad V_\sigma^n = V_\sigma^D, \quad \forall \sigma \in \mathscr{E}^D, \quad \forall n \geq 0. \tag{14}$$

The scheme is backward Euler in time and finite volume in space with a two-point flux approximation. It writes, for all $n \geq 0$, for all $K \in \mathscr{T}$:

$$\mathrm{m}(K)\frac{\rho_{1,K}^{n+1} - \rho_{1,K}^n}{\Delta t} + \sum_{\sigma \in \mathscr{E}_K} \mathscr{F}_{1,K,\sigma}^{n+1} = 0, \tag{15a}$$

$$\begin{aligned} \mathrm{m}(K)\frac{\rho_{2,K}^{n+1} - \rho_{2,K}^n}{\Delta t} + \sum_{\sigma \in \mathscr{E}_K} \mathscr{F}_{2,K,\sigma}^{n+1} &= \mathrm{m}(K) W_K^{n+1} \\ &+ \sum_{\sigma \in \mathscr{E}_K} V_\sigma^{n+1} \mathscr{F}_{1,K,\sigma}^{n+1} - V_K^{n+1} \sum_{\sigma \in \mathscr{E}_K} \mathscr{F}_{1,K,\sigma}^{n+1}, \end{aligned} \tag{15b}$$

$$-\lambda^2 \sum_{\sigma \in \mathscr{E}_K} \tau_\sigma D_{K,\sigma} V^{n+1} = \mathrm{m}(K)(C_K - \rho_{1,K}^{n+1}), \tag{15c}$$

where

$$\rho_{i,K}^{n+1} = \rho_i(u_K^{n+1}), \quad i = 1,\ 2 \text{ and } W_K^{n+1} = W(u_K^{n+1}) \text{ for all } K \in \mathscr{T}.$$

The numerical fluxes are given by

$$\mathscr{F}_{1,K,\sigma}^{n+1} = -\tau_\sigma \left(L_{11,\sigma}^n (D_{K,\sigma} \mathbf{u}_1{}^{n+1} + u_{2,\sigma}^{n+1} D_{K,\sigma} \mathbf{V}^{n+1}) + L_{12,\sigma}^n D_{K,\sigma} \mathbf{u}_2{}^{n+1} \right), \tag{16a}$$

$$\mathscr{F}_{2,K,\sigma}^{n+1} = -\tau_\sigma \left(L_{12,\sigma}^n (D_{K,\sigma} \mathbf{u}_1{}^{n+1} + u_{2,\sigma}^{n+1} D_{K,\sigma} \mathbf{V}^{n+1}) + L_{22,\sigma}^n D_{K,\sigma} \mathbf{u}_2{}^{n+1} \right), \tag{16b}$$

where the matrix $\mathbb{L}_\sigma^n = (L_{ij,\sigma}^n)_{1 \leq i,j \leq n}$ is defined as

$$\mathbb{L}_\sigma^n = \mathbb{L}\left(\frac{u_K^n + u_{K,\sigma}^n}{2}\right) \quad \text{for all } K \in \mathscr{T}, \sigma \in \mathscr{E}_K. \tag{17}$$

At this point, it remains to define V_σ^{n+1} involved in (15b) and $u_{2,\sigma}^{n+1}$ involved in (16) for all $\sigma \in \mathscr{E}$. We will do it later. The choice will be driven by the expected equivalence with a scheme for (7)–(10).

In order to obtain an equivalent scheme for the energy transport system in the dual entropy variables (7)–(10), we apply the change of variables (6), associated with the new functions defined in (8)–(10), to (15) and (16). Let us define for all $K \in \mathscr{T}$, for all $n \geq 0$,

$$w_{1,K}^n = u_{1,K}^n + u_{2,K}^n V_K^n, \qquad w_{2,K}^n = u_{2,K}^n, \tag{18a}$$

$$b_{1,K}^n = \rho_{1,K}^n = b_1(w_K^n, V_K^n), \qquad b_{2,K}^n = \rho_{2,K}^n - \rho_{1,K}^n V_K^n = b_2(w_K^n, V_K^n). \tag{18b}$$

We similarly define $w_{1,\sigma}^D$ and $w_{2,\sigma}^D$ for $\sigma \in \mathscr{E}^D$. From (15a) and (15b), we deduce

$$\mathrm{m}(K)\frac{b_{1,K}^{n+1} - b_{1,K}^n}{\Delta t} + \sum_{\sigma \in \mathscr{E}_K} \mathscr{F}_{1,K,\sigma}^{n+1} = 0,$$

$$\mathrm{m}(K)\frac{b_{2,K}^{n+1} - b_{2,K}^n}{\Delta t} + \sum_{\sigma \in \mathscr{E}_K} \left(\mathscr{F}_{2,K,\sigma}^{n+1} - V_\sigma^{n+1}\mathscr{F}_{1,K,\sigma}^{n+1}\right)$$
$$= \mathrm{m}(K) W_K^{n+1} - \mathrm{m}(K)\frac{V_K^{n+1} - V_K^n}{\Delta t} b_{1,K}^n.$$

It leads to the following scheme for the system written in the dual variables (7):

$$\mathrm{m}(K)\frac{b_{1,K}^{n+1} - b_{1,K}^n}{\Delta t} + \sum_{\sigma \in \mathscr{E}_K} \mathscr{G}_{1,K,\sigma}^{n+1} = 0, \tag{19a}$$

$$\mathrm{m}(K)\frac{b_{2,K}^{n+1} - b_{2,K}^n}{\Delta t} + \sum_{\sigma \in \mathscr{E}_K} \mathscr{G}_{2,K,\sigma}^{n+1} = \mathrm{m}(K)\tilde{W}_K^{n+1} - \mathrm{m}(K)\frac{V_K^{n+1} - V_K^n}{\Delta t} b_{1,K}^n, \tag{19b}$$

$$-\lambda^2 \sum_{\sigma \in \mathscr{E}_K} \tau_\sigma D_{K,\sigma} V^{n+1} = \mathrm{m}(K)(C_K - b_{1,K}^{n+1}), \tag{19c}$$

with

$$\mathscr{G}_{1,K,\sigma}^{n+1} = \mathscr{F}_{1,K,\sigma}^{n+1}, \quad \forall K \in \mathscr{T}, \forall \sigma \in \mathscr{E}_K, \tag{20a}$$

$$\mathscr{G}_{2,K,\sigma}^{n+1} = \mathscr{F}_{2,K,\sigma}^{n+1} - V_\sigma^{n+1}\mathscr{F}_{1,K,\sigma}^{n+1}, \quad \forall K \in \mathscr{T}, \forall \sigma \in \mathscr{E}_K, \tag{20b}$$

and $\tilde{W}_K^{n+1} = W_K^{n+1} = \tilde{W}(w_K^{n+1})$.

The crucial point now is to ensure that the new numerical fluxes $\mathscr{G}_{1,K,\sigma}^{n+1}$, $\mathscr{G}_{2,K,\sigma}^{n+1}$ can be seen as approximations of the currents I_1 and I_2 defined by (9). This means that we want to rewrite the numerical fluxes as

$$\mathscr{G}_{1,K,\sigma}^{n+1} = -\tau_\sigma (\mathrm{D}_{11,\sigma}^* D_{K,\sigma}\mathbf{w}_1^{n+1} + \mathrm{D}_{12,\sigma}^* D_{K,\sigma}\mathbf{w}_2^{n+1}), \tag{21a}$$

$$\mathscr{G}_{2,K,\sigma}^{n+1} = -\tau_\sigma (\mathrm{D}_{21,\sigma}^* D_{K,\sigma}\mathbf{w}_1^{n+1} + \mathrm{D}_{22,\sigma}^* D_{K,\sigma}\mathbf{w}_2^{n+1}), \tag{21b}$$

with the coefficients $(\mathrm{D}_{ij,\sigma}^*)_{1\le i,j\le 2}$ defined such that the associate matrix $\mathbb{D}_\sigma^*$ is symmetric and uniformly positive definite. This property will now depend on the definition of V_σ^{n+1} and $u_{2,\sigma}^{n+1}$, respectively involved in (15b) and (16), for each edge $\sigma \in \mathscr{E}$.

Equivalence of the Schemes in the Primal and Dual Entropy Variables

Proposition 1 *Let us supplement the scheme (15)–(16) with the definition of the* $(V_\sigma^{n+1})_{\sigma\in\mathscr{E},\ n\ge 0}$ *and* $(u_{2,\sigma}^{n+1})_{\sigma\in\mathscr{E},\ n\ge 0}$. *We distinguish two cases:*

- *Case 1: centered scheme. For all* $\sigma \in \mathscr{E}$ *and* $n \ge 0$, *we set:*

$$u_{2,\sigma}^{n+1} = \frac{u_{2,K}^{n+1} + u_{2,K,\sigma}^{n+1}}{2} \quad \textit{and} \quad V_\sigma^{n+1} = \frac{V_K^{n+1} + V_{K,\sigma}^{n+1}}{2}. \tag{22}$$

- *Case 2: upwind scheme. For all* $\sigma \in \mathscr{E}$ *and* $n \ge 0$, *we set:*

$$u_{2,\sigma}^{n+1} = \begin{cases} u_{2,K,\sigma}^{n+1}, & \textit{if } D_{K,\sigma}V^{n+1} > 0, \\ u_{2,K}^{n+1}, & \textit{if } D_{K,\sigma}V^{n+1} \le 0, \end{cases} \quad \textit{and } V_\sigma^{n+1} = \min(V_K^{n+1}, V_{K,\sigma}^{n+1}). \tag{23}$$

Then, in both cases, the scheme (15)–(16) written in the primal entropy variables is equivalent with the scheme (19)–(21) written in the dual entropy variables, provided that

$$\mathbb{D}_\sigma^* = (\mathbb{P}_\sigma^{n+1})^T \mathbb{L}^n \mathbb{P}_\sigma^{n+1} \textit{ with } \mathbb{P}_\sigma^{n+1} = \begin{pmatrix} 1 & -V_\sigma^{n+1} \\ 0 & 1 \end{pmatrix}. \tag{24}$$

Proof Starting from the definition (20) of the numerical fluxes $\mathscr{G}_{1,K,\sigma}^{n+1}$ and $\mathscr{G}_{2,K,\sigma}^{n+1}$, we want to establish (21) with $\mathbb{D}_\sigma^*$ defined by (24).

Let us first notice that, due to the change of variables (18a), we can rewrite $D_{K,\sigma}\mathbf{u}_1{}^{n+1}$ and $D_{K,\sigma}\mathbf{u}_2{}^{n+1}$ for all $K \in \mathscr{T}$ and $\sigma \in \mathscr{E}_K$. It is clear that $D_{K,\sigma}\mathbf{u}_2{}^{n+1} = D_{K,\sigma}\mathbf{w}_2{}^{n+1}$. Moreover, we have

$$\begin{aligned} D_{K,\sigma}\mathbf{u}_1{}^{n+1} &= D_{K,\sigma}\mathbf{w}_1{}^{n+1} - V_K^{n+1}\, D_{K,\sigma}\mathbf{w}_2{}^{n+1} - w_{2,K,\sigma}^{n+1}\, D_{K,\sigma}\mathbf{V}^{n+1}, \\ &= D_{K,\sigma}\mathbf{w}_1{}^{n+1} - V_{K,\sigma}^{n+1}\, D_{K,\sigma}\mathbf{w}_2{}^{n+1} - w_{2,K}^{n+1}\, D_{K,\sigma}\mathbf{V}^{n+1}. \end{aligned}$$

This yields, for Case 1 as well as for Case 2,

$$D_{K,\sigma}\mathbf{u}_1{}^{n+1} = D_{K,\sigma}\mathbf{w}_1{}^{n+1} - V_\sigma^{n+1}\, D_{K,\sigma}\mathbf{w}_2{}^{n+1} - w_{2,\sigma}^{n+1}\, D_{K,\sigma}\mathbf{V}^{n+1},$$

with $w_{2,\sigma}^{n+1} = u_{2,\sigma}^{n+1}$. Therefore, from (16) and (20), we deduce that

$$\begin{aligned}\mathscr{G}_{1,K,\sigma}^{n+1} &= -\tau_\sigma \left(L_{11,\sigma}^n D_{K,\sigma} \mathbf{w}_1{}^{n+1} + (L_{12,\sigma}^n - V_\sigma^{n+1} L_{11,\sigma}^n) D_{K,\sigma} \mathbf{w}_2{}^{n+1} \right), \\ \mathscr{G}_{2,K,\sigma}^{n+1} &= -\tau_\sigma \left((L_{12,\sigma}^n - V_\sigma^{n+1} L_{11,\sigma}^n) D_{K,\sigma} \mathbf{w}_1{}^{n+1} \right) \\ &\qquad\qquad + (L_{22,\sigma}^n - 2V_\sigma^{n+1} L_{12,\sigma}^n + (V_\sigma^{n+1})^2 L_{11,\sigma}^n) D_{K,\sigma} \mathbf{w}_2{}^{n+1} \Big).\end{aligned}$$

This corresponds to (21) with $\mathbb{D}_\sigma^*$ defined by (24). We have shown that the scheme (15) and (16), supplemented either with (22) or (23), implies (19)–(24). Starting from (19)–(24), we similarly get (15) and (16).

3 Discrete Entropy Inequality

In this Section, we establish the discrete counterpart of the decay of the entropy, with the control of its dissipation, (13). The result is stated in Proposition 2.

Main Result

First of all, since the functions u_1^D, u_2^D, V^D are assumed to be defined on the whole domain Ω, we can set

$$(u_{1,K}^D, u_{2,K}^D, V_K^D) = \frac{1}{\mathrm{m}(K)} \int_K (u_1^D(x), u_2^D(x), V^D(x)) dx, \quad \forall K \in \mathscr{T}.$$

Moreover, we remember that u_2^D is a constant function, such that

$$u_{K,2}^D = u_2^D < 0, \quad \forall K \in \mathscr{T}. \tag{25}$$

Let $(u_K^n = (u_{1,K}^n, u_{2,K}^n)^T, V_K^n)_{K \in \mathscr{T}, n \geq 0}$ be a solution to the scheme (14)–(17), supplemented with either (22) or (23). For all $n \geq 0$, we define the discrete entropy functional as follows:

$$\begin{aligned}S^n = \sum_{K \in \mathscr{T}} \mathrm{m}(K) \left[\rho_K^n \cdot (u_K^n - u_K^D) - (\chi(u_K^n) - \chi(u_K^D)) \right] \\ - \frac{\lambda^2}{2} u_2^D \sum_{\sigma \in \mathscr{E}} \tau_\sigma (D_\sigma (\mathbf{V}^n - \mathbf{V}^D))^2.\end{aligned} \tag{26}$$

We recall that $\rho_K^n = \rho(u_K^n) = (\rho_1(u_K^n), \rho_2(u_K^n))^T$ and that ρ is related to χ by (5c). Therefore, S^n is nonnegative for all $n \geq 0$.

Proposition 2 (Discrete entropy dissipation) *Assume (4), (5), (25) and let* $(u_K^n = (u_{1,K}^n, u_{2,K}^n)^T, V_K^n)_{K\in\mathcal{T},n\geq 0}$ *be a solution to the scheme (14)–(17), supplemented with either (22) or (23). The discrete entropy satisfies the following inequality: for all* $n \geq 0$*,*

$$\begin{aligned}\frac{S^{n+1}-S^n}{\Delta t} \leq &- \sum_{\sigma\in\mathcal{E}} \tau_\sigma (D_{K,\sigma}\mathbf{w}^{n+1})^T \mathbb{D}_\sigma^* D_{K,\sigma}\mathbf{w}^{n+1} \\ &+ \sum_{K\in\mathcal{T}} \mathrm{m}(K)\, W_K^{n+1}(w_{2,K}^{n+1} - w_{2,K}^D) \leq 0,\end{aligned} \tag{27}$$

where $D_{K,\sigma}\mathbf{w}^{n+1} = (D_{K,\sigma}\mathbf{w}_1^{n+1}, D_{K,\sigma}\mathbf{w}_2^{n+1})^T$.

Proof Using the definition (26) of the discrete entropy, one has

$$S^{n+1} - S^n = A + B, \tag{28}$$

where

$$\begin{aligned}A = \sum_{K\in\mathcal{T}} \mathrm{m}(K)\Big(&\rho_K^{n+1}\cdot(u_K^{n+1}-u_K^D) - (\chi(u_K^{n+1}) - \chi(u_K^D)) \\ &- \rho_K^n\cdot(u_K^n - u_K^D) + (\chi(u_K^n) - \chi(u_K^D))\Big),\end{aligned} \tag{29}$$

$$B = -\frac{\lambda^2}{2}u_2^D \sum_{\sigma\in\mathcal{E}} \tau_\sigma \Big[(D_\sigma(\mathbf{V}^{n+1}-\mathbf{V}^D))^2 - (D_\sigma(\mathbf{V}^n - \mathbf{V}^D))^2\Big]. \tag{30}$$

We first consider the term A. As χ is a convex function such that $\rho = \nabla_u\chi$, leading to $\rho_K^n = \nabla_u\chi(u_K^n)$, we have:

$$\chi(u_K^{n+1}) - \chi(u_K^n) - \rho_K^n\cdot(u_K^{n+1} - u_K^n) \geq 0.$$

This yields

$$A \leq \sum_{K\in\mathcal{T}} m(K)(\rho_K^{n+1} - \rho_K^n)\cdot(u_K^{n+1} - u_K^D). \tag{31}$$

We now address the term B. Since $(a^2 - b^2)/2 \leq a(a-b)$, for all $a, b \in \mathbb{R}$, and $u_2^D \leq 0$, we get:

$$B \leq -\lambda^2 u_2^D \sum_{\sigma\in\mathcal{E}} \tau_\sigma D_{K,\sigma}(\mathbf{V}^{n+1} - \mathbf{V}^D)\, D_{K,\sigma}(\mathbf{V}^{n+1} - \mathbf{V}^n).$$

A discrete integration by part leads to

$$B \leq \lambda^2 u_2^D \sum_{K\in\mathcal{T}} (V_K^{n+1} - V_K^D)\left(\sum_{\sigma\in\mathcal{E}_K} \tau_\sigma D_{K,\sigma}(\mathbf{V}^{n+1} - \mathbf{V}^n)\right).$$

Using the scheme for the Poisson equation (15c), we obtain

$$B \leq u_2^D \sum_{K\in\mathcal{T}} \mathrm{m}(K)(V_K^{n+1} - V_K^D)(\rho_{1,K}^{n+1} - \rho_{1,K}^n). \tag{32}$$

From (28), (31) and (32), we deduce:

$$\begin{aligned} S^{n+1} - S^n \leq \sum_{K\in\mathcal{T}} \mathrm{m}(K)(\rho_{1,K}^{n+1} - \rho_{1,K}^n)\big((u_{1,K}^{n+1} - u_{1,K}^D) + u_2^D(V_K^{n+1} - V_K^D)\big) \\ + \sum_{K\in\mathcal{T}} \mathrm{m}(K)(\rho_{2,K}^{n+1} - \rho_{2,K}^n)(u_{2,K}^{n+1} - u_{2,K}^D). \end{aligned} \tag{33}$$

Using the primal scheme (15a), (15b), the inequality (33) becomes

$$\frac{S^{n+1} - S^n}{\Delta t} \leq C + D + \sum_{K\in\mathcal{T}} m(K) W_K^{n+1}(u_{2,K}^{n+1} - u_{2,K}^D), \tag{34}$$

with

$$\begin{aligned} C = -\sum_{K\in\mathcal{T}} \left(\sum_{\sigma\in\mathcal{E}_K} \mathcal{F}_{1,K,\sigma}^{n+1}\right) \left[(u_{1,K}^{n+1} - u_{1,K}^D) + V_K^{n+1}(u_{2,K}^{n+1} - u_{2,K}^D)\right] \\ - u_2^D \sum_{K\in\mathcal{T}} \left(\sum_{\sigma\in\mathcal{E}_K} \mathcal{F}_{1,K,\sigma}^{n+1}\right)(V_K^{n+1} - V_K^D), \end{aligned}$$

$$D = -\sum_{K\in\mathcal{T}} \left(\sum_{\sigma\in\mathcal{E}_K} \mathcal{F}_{2,K,\sigma}^{n+1} - V_\sigma^{n+1} F_{1,K,\sigma}^{n+1}\right)(u_{2,K}^{n+1} - u_{2,K}^D).$$

Using the change of variables (18a), the relations (20) on the numerical fluxes written in the primal and dual entropy variables and the hypothesis (25), we get

$$\begin{aligned} C &= -\sum_{K\in\mathcal{T}} \left(\sum_{\sigma\in\mathcal{E}_K} \mathcal{G}_{1,K,\sigma}^{n+1}\right)(w_{1,K}^{n+1} - w_{1,K}^D), \\ D &= -\sum_{K\in\mathcal{T}} \left(\sum_{\sigma\in\mathcal{E}_K} \mathcal{G}_{2,K,\sigma}^{n+1}\right)(w_{2,K}^{n+1} - w_{2,K}^D). \end{aligned} \tag{35}$$

Accounting for the boundary conditions, we conclude by a discrete integration by parts which gives (27):

$$\begin{aligned}\frac{S^{n+1}-S^{n}}{\Delta t} \le \sum_{\sigma\in\mathscr{E}} \mathscr{G}_{1,K,\sigma}^{n+1} D_{K,\sigma}\mathbf{w}_1^{n+1} + \sum_{\sigma\in\mathscr{E}} \mathscr{G}_{2,K,\sigma}^{n+1} D_{K,\sigma}\mathbf{w}_2^{n+1} \\ + \sum_{K\in\mathscr{T}} \mathrm{m}(K) W_K^{n+1}(w_{2,K}^{n+1} - w_{2,K}^{D}).\end{aligned} \tag{36}$$

The formulation (21) of the numerical fluxes $\mathscr{G}_{i,K,\sigma}^{n+1}$ permits to rewrite

$$\begin{aligned}\sum_{\sigma\in\mathscr{E}} \mathscr{G}_{1,K,\sigma}^{n+1} D_{K,\sigma}\mathbf{w}_1^{n+1} + \sum_{\sigma\in\mathscr{E}} \mathscr{G}_{2,K,\sigma}^{n+1} D_{K,\sigma}\mathbf{w}_2^{n+1} \\ = -\sum_{\sigma\in\mathscr{E}} \tau_\sigma \begin{pmatrix} D_{K,\sigma}\mathbf{w}_1^{n+1} \\ D_{K,\sigma}\mathbf{w}_2^{n+1}\end{pmatrix}^{T} \mathbb{D}_\sigma^{*} \begin{pmatrix} D_{K,\sigma}\mathbf{w}_1^{n+1} \\ D_{K,\sigma}\mathbf{w}_2^{n+1}\end{pmatrix}.\end{aligned} \tag{37}$$

From (36) and (37), we deduce (27). The hypothesis (4) on the energy relaxation term and the positive definiteness of the matrices $\mathbb{D}_\sigma$ ensure the nonpositivity of the right-hand-side in (27) and the decay of the discrete entropy.

Consequences

From Proposition 2, we deduce the uniform bound: $S^n \le S^0$ for all $n \ge 0$. The control of the dissipation writes

$$\sum_{n=0}^{N}\sum_{\sigma\in\mathscr{E}} \tau_\sigma (D_{K,\sigma}\mathbf{w}^{n+1})^T \mathbb{D}_\sigma^{*} D_{K,\sigma}\mathbf{w}^{n+1} \le S^0.$$

This yields a discrete $L^2(0, T_{\max}, H^1)$ estimates on w_1 and w_2. But, following the ideas of [6, 17], we may obtain other *a priori* estimates on the solution. They permit first to prove the existence of a solution to the scheme, thanks to a topological degree argument, and second to show the compactness of the sequence of approximate solutions leading to the convergence of the scheme. The existence result and the convergence analysis will be detailed in a forthcoming paper.

4 Numerical Experiments

For the numerical experiments, we consider the unipolar energy-transport model under Boltzmann statistics, as in [6, 17]. It is based on the following definitions of the densities $\rho_i(u)$, $i = 1, 2$:

$$\begin{cases} \rho_1(u) = \left(-\dfrac{1}{u_2}\right)^{3/2} \exp(u_1), \\ \rho_2(u) = \dfrac{3}{2}\left(-\dfrac{1}{u_2}\right)^{5/2} \exp(u_1). \end{cases} \tag{38}$$

so that $\rho(u) = \nabla_u \chi(u)$ with $\chi(u) = (-u_2)^{-3/2} \exp(u_1)$.

The diffusion matrix $\mathbb{L}(u) = (L_{ij}(u))_{1\le i,j\le 2}$ actually depends on u under the following form [17]:

$$\mathbb{L} = c_o \rho_1(u) T^{1/2-\beta} \begin{pmatrix} 1 & (2-\beta)T \\ (2-\beta)T & (3-\beta)(2-\beta)T^2 \end{pmatrix}, \tag{39}$$

where $c_0 > 0$ is a constant (and we recall that $T = -1/u_2$). The usual values of β are $1/2$, corresponding to the Chen model, and 0, corresponding to the Lyumkis model [17]. The matrix $\mathbb{L}(u)$ is symmetric positive definite.

Presentation of the Test Case

We consider a test case of a 2-D n^+nn^+ silicon diode, uniform in one space direction, already introduced in [3, 7, 13]. It is a simple model for the channel of a MOS transistor. The adopted model is the Chen model ($\beta = 1/2$ in (39)). Additional test cases will be given in a forthcoming paper.

The domain is $\Omega = (0, l_x) \times (0, l_y)$ with $l_x = 0.6\,\mu\text{m}$ and $l_y = 0.2\,\mu\text{m}$. The channel length is $0.4\,\mu\text{m}$, see Fig. 1.

The numerical values of the physical parameters for a silicon diode are given in Table 1. The doping profile is

$$\begin{aligned} C = C_m &= 5 \times 10^{17} cm^{-3} \quad \text{in the } n^+ \text{ region}, \\ C = C_m &= 2 \times 10^{15} cm^{-3} \quad \text{in the } n \text{ region}. \end{aligned}$$

The boundary conditions are

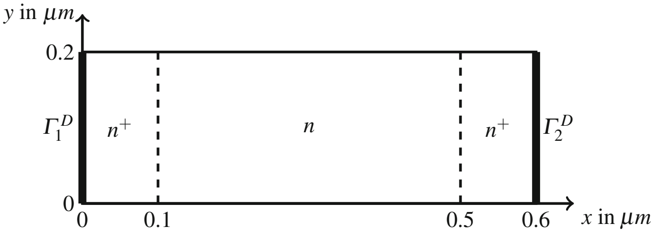


Fig. 1 Geometry of the n^+nn^+ ballistic diode

Table 1 Physical parameters

Parameter	Physical meaning	Numerical value
q	Elementary charge	10^{-19} As
ε	Permittivity constant	10^{-12}AsV^{-1}cm^{-1}
μ_0	Low field mobility	1.5×10^3cm^{2}V^{-1}s^{-1}
U_T	Thermal voltage at $T_0 = 300\,\mathrm{K}$	0.0259 V
τ_0	Energy relaxation time	0.4×10^{-12} s

$$\begin{aligned}
&V = 1.5\mathrm{V} \text{ on } \Gamma_1^D \text{ and } V = 0 \text{ on } \Gamma_2^D,\\
&u_2 = -1/T_0, \text{ with } T_0 = 300K, \text{ on } \Gamma_1^D \cup \Gamma_2^D,\\
&\rho_1(u) = C_m \text{ on } \Gamma_1^D \cup \Gamma_2^D,
\end{aligned}$$

the latest giving the boundary condition for u_1 according to (38). The initial conditions for u_1 and u_2 are constant and equal to the boundary conditions.

The function W reads

$$W(u) = c_1\rho_1(u) - c_2\rho_2(u),$$

with

$$c_1 = \frac{3}{2}\frac{l_x^2}{\tau_0\mu_0 U_T}, \quad c_2 = \frac{l_x^2}{\tau_0\mu_0 U_T},$$

and the scaling ensures that the Debye length is

$$\lambda^2 = \frac{\varepsilon U_T}{q l_x^2 C_m}.$$

Numerical Results

We use an admissible mesh made of 896 triangles. Figure 2 presents the results obtained by the scheme (15) and (16) in the centered case (22). The results are plotted for the final time $T_{\mathrm{final}} = 1$s, as the equilibrium state is reached. Although the discretization is fully implicit, it is necessary to use an adaptative time step during the first iterations, in order to allow the convergence of the Newton's method. As expected, the computed quantities are almost uniform in one space direction. Moreover one observes the expected hot electron effect in the channel, which compares with the results given in [3, 7, 13].

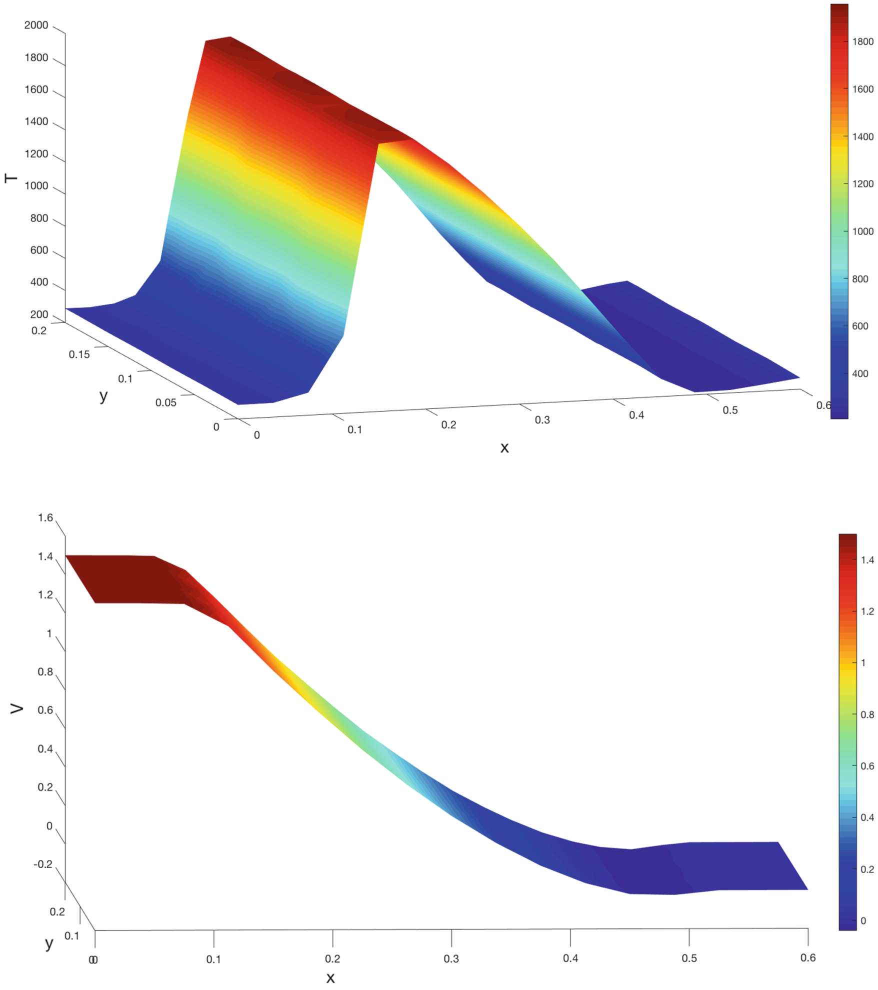


Fig. 2 2-D $n^{+}nn^{+}$ diode: temperature (above) and electrostatic potential (below)

References

1. Bradji, A., Herbin, R.: Discretization of coupled heat and electrical diffusion problems by finite-element and finite-volume methods. IMA J. Numer. Anal. **28**(3), 469–495 (2008)
2. Calgaro, C., Colin, C., Creusé, E.: A combined finite volume—finite element scheme for a low-Mach system involving a Joule term. AIMS Math. **5**(1), 311–331 (2019)
3. Chainais-Hillairet, C.: Discrete duality finite volume schemes for two-dimensional drift-diffusion and energy-transport models. Internat. J. Numer. Methods Fluids **59**(3), 239–257 (2009)
4. Chen, L., Hsiao, L.: The solution of Lyumkis energy transport model in semiconductor science. Math. Methods Appl. Sci. **26**(16), 1421–1433 (2003)

5. Chen, L., Hsiao, L., Li, Y.: Large time behavior and energy relaxation time limit of the solutions to an energy transport model in semiconductors. J. Math. Anal. Appl. **312**(2), 596–619 (2005)
6. Degond, P., Génieys, S., Jüngel, A.: A system of parabolic equations in nonequilibrium thermodynamics including thermal and electrical effects. J. Math. Pures Appl. (9), **76**(10), 991–1015 (1997)
7. Degond, P., Jüngel, A., Pietra, P.: Numerical discretization of energy-transport models for semiconductors with nonparabolic band structure. SIAM J. Sci. Comput. **22**(3), 986–1007 (2000)
8. Doan, D.H., Fischer, A., Fuhrmann, J., Glitzky, A., Liero, M.: Drift-diffusion simulation of s-shaped current-voltage relations for organic semiconductor devices. Working paper or preprint. http://www.wias-berlin.de/preprint/2630/wias_preprints_2630.pdf (2019)
9. Eymard, R., Gallouët, T., Herbin, R.: Finite volume methods. Handbook of Numerical Analysis, vol. VII, pp. 713–1020. North-Holland, Amsterdam, (2000)
10. Fang, W., Ito, K.: Existence of stationary solutions to an energy drift-diffusion model for semiconductor devices. Math. Models Methods Appl. Sci. **11**(5), 827–840 (2001)
11. Fournié, M.: Numerical discretization of energy-transport model for semiconductors using high-order compact schemes. Appl. Math. Lett. **15**(6), 721–726 (2002)
12. Griepentrog, J.A.: An application of the implicit function theorem to an energy model of the semiconductor theory. ZAMM Z. Angew. Math. Mech. **79**(1), 43–51 (1999)
13. Holst, S., Jüngel, A., Pietra, P.: A mixed finite-element discretization of the energy-transport model for semiconductors. SIAM J. Sci. Comput. **24**(6), 2058–2075 (2003)
14. Holst, S., Jüngel, A., Pietra, P.: An adaptive mixed scheme for energy-transport simulations of field-effect transistors. SIAM J. Sci. Comput. **25**(5), 1698–1716 (2004)
15. Jüngel, A.: Regularity and uniqueness of solutions to a parabolic system in nonequilibrium thermodynamics. Nonlinear Anal. **41**(5–6, Ser. A: Theory Methods), 669–688 (2000)
16. Jüngel, A., Pinnau, R., Röhrig, E.: Existence analysis for a simplified transient energy-transport model for semiconductors. Math. Methods Appl. Sci. **36**(13), 1701–1712 (2013)
17. Jüngel, A.: Quasi-hydrodynamic semiconductor equations. In: Progress in Nonlinear Differential Equations and their Applications, vol. 41. Birkhäuser Verlag, Basel (2001)
18. Kantner, M.: Generalized Scharfetter-Gummel schemes for electro-thermal transport in degenerate semiconductors using the Kelvin formula for the Seebeck coefficient. J. Comput. Phys. **402**, 109091 (2020)
19. Romano, V.: 2D numerical simulation of the MEP energy-transport model with a finite difference scheme. J. Comput. Phys. **221**(2), 439–468 (2007)
20. Zamponi, N., Jüngel, A.: Global existence analysis for degenerate energy-transport models for semiconductors. J. Differ. Equ. **258**(7), 2339–2363 (2015)

Theoretical Aspects

Compatible Discrete Operator Schemes for the Steady Incompressible Stokes and Navier–Stokes Equations

Jérôme Bonelle, Alexandre Ern, and Riccardo Milani

Abstract We extend the Compatible Discrete Operator (CDO) schemes to the steady incompressible Stokes and Navier–Stokes equations. The main features of the CDO face-based schemes are recalled: a hybrid velocity discretization with degrees of freedom at faces and cells, a stabilized velocity gradient reconstruction defined on the face-based subcell pyramids, and a discrete pressure attached to the mesh cells. We introduce a discrete divergence operator that will account for the velocity-pressure coupling, and a hybrid discretization of the convection term. The results of several benchmark test cases validate the framework.

Keywords CDO · Stokes · Navier–Stokes · Structure-preserving schemes

MSC (2010) 65N08 · 76D05 · 76M12

1 Introduction

The Compatible Discrete Operator (CDO) schemes provide a low-order framework which is part of the so-called mimetic or structure-preserving methods. One of the main advantages of the CDO schemes is that they can handle polytopal, nonmatching (cf. left part of Fig. 1) and deformed meshes. Taking advantage of a dual mesh,

R. Milani (✉)
EDF R&D and ENPC (CERMICS) and INRIA Paris, 6 Quai Watier,
78400 Chatou, France
e-mail: riccardo.milani@edf.fr

J. Bonelle
EDF R&D, 6 Quai Watier, 78400 Chatou, France
e-mail: jerome.bonelle@edf.fr

A. Ern
ENPC (CERMICS) and INRIA Paris, 6-8 Av. Blaise Pascal,
77455 Marne-la-Vallée, France
e-mail: alexandre.ern@enpc.fr

R. Klöfkorn et al. (eds.), *Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples*, Springer Proceedings in Mathematics & Statistics 323,
https://doi.org/10.1007/978-3-030-43651-3_6

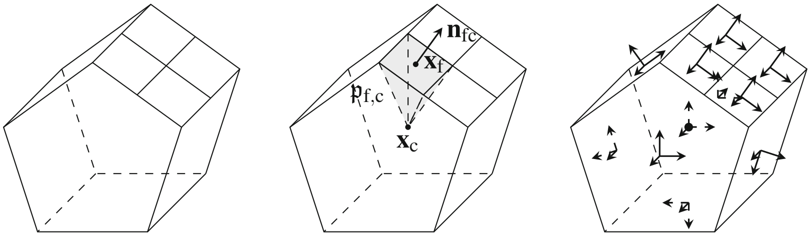


Fig. 1 Example of cell compatible with CDO. **Left** cell with hanging nodes. **Center** cell with one of its subpyramids $\mathfrak{p}_{f,c}$. **Right** cell with velocity (arrows) and pressure (circle) DoFs

invisible to the end user, discrete differential operators are carefully designed to satisfy conservation laws and properties typical of their continuous counterparts. This special treatment results in optimal order of convergence in space [3] (first order for the reconstructed gradient and second for the original variable) and a cell-wise and fully parallelizable building stage ensures good performances of the overall method. Thanks to its flexibility, the CDO framework allows to define the main problem variables on different mesh entities, according to their physical nature. Hence, one can choose to use a cell-, vertex- [3], edge- [8] or face-based [2] scheme.

Here, the Stokes and Navier–Stokes equations (NSE) are discretized by means of face-based CDO (CDO-Fb) schemes. In this case, the velocity is defined at faces and cells, and the pressure is defined at cells only. CDO-Fb was introduced initially for the Poisson problem [2] and its key ingredient is a stabilized subcell gradient reconstruction, which can be bridged to the one used in the Hybrid Mixed Mimetic (HMM) framework [12] and to a generalization of the Crouzeix-Raviart framework [10] (GCR). A divergence operator is derived from this gradient and it is the tool on which the velocity-pressure coupling hinges. Finally, the discretization of the convection term is inspired by the lowest-order case of the Hybrid High-Order (HHO($k = 0$)) method [9]. The Stokes problem in its curl formulation has been already treated by means of CDO with vertex- and cell-based schemes [4] but here we retain the face-based one.

Let $\mathscr{D} \subset \mathbb{R}^d$, $d = 2, 3$, be a bounded connected polyhedral domain and denote by $\partial\mathscr{D}$ its boundary. We consider the following model problem:

$$-\nu\Delta\mathbf{u} + \chi(\mathbf{u}\cdot\nabla)\mathbf{u} + \nabla p = \mathbf{f}, \qquad \text{in } \mathscr{D} \tag{1a}$$

$$\nabla\cdot\mathbf{u} = 0, \qquad \text{in } \mathscr{D} \tag{1b}$$

where $\nu > 0$ is the viscosity, and $\chi = 0$ for the Stokes equations or $\chi = 1$ for the NSE. For the sake of simplicity, homogeneous Dirichlet boundary conditions (BC) are considered. The pressure is uniquely defined by requiring that $\int_{\mathscr{D}} p = 0$.

2 Space Discretization

A mesh discretizing $\mathscr{D}$ is a finite collection $\mathrm{C} := \{\mathrm{c}\}$ of nonempty, disjoint, open, polytopal elements of $\mathbb{R}^d$, $d = 2, 3$, usually referred to as cells c. The faces f are assumed to be planar and are gathered in the set F which may be subdivided in two disjoint sets: $\mathrm{F}^b := \{\mathrm{f} \mid \mathrm{f} \subset \partial\mathscr{D}\}$ collects the boundary faces and $\mathrm{F}^i := \mathrm{F} \setminus \mathrm{F}^b$ the interfaces. One associates with each face f a normal vector $\mathbf{n}_\mathrm{f}$: if $\mathrm{f} \in \mathrm{F}^b$, $\mathbf{n}_\mathrm{f}$ points outward $\mathscr{D}$ and, if $\mathrm{f} \in \mathrm{F}^i$, the direction is chosen arbitrarily. For a mesh entity $\mathrm{z} = \mathrm{c}, \mathrm{f}$, $\mathbf{x}_\mathrm{z}$ denotes its barycenter and $|\mathrm{z}|$ its measure. Consider now a generic cell c. Define the set of faces of the cell c as $\mathrm{F}_\mathrm{c} := \{\mathrm{f} \in \mathrm{F} \mid, \mathrm{f} \subset \partial\mathrm{c}\}$. For every $\mathrm{f} \in \mathrm{F}_\mathrm{c}$, $\mathbf{n}_\mathrm{fc} := \pm\mathbf{n}_\mathrm{f}$ is the normal vector to the face f pointing outward c, the sign depends on the direction chosen for $\mathbf{n}_\mathrm{f}$. The subpyramid obtained by joining the vertices of f to the barycenter $\mathbf{x}_\mathrm{c}$ of the cell (cf. the central part of Fig. 1) is denoted by $\mathfrak{p}_{\mathrm{f,c}}$.

2.1 Discrete Functional Spaces and Differential Operators

Given a generic mesh entity $\mathrm{z} = \mathrm{c}, \mathrm{f}$, $\mathbb{P}^0(\mathrm{z}) \equiv \mathbb{R}$ denotes the scalar-valued, zero-th order polynomials defined on z. We denote with $\pi_\mathrm{z}\colon L^1(\mathrm{z}) \to \mathbb{P}^0(\mathrm{z}) \equiv \mathbb{R}$ the L^2-projection (average): for all $s \in L^1(\mathrm{z})$, $\pi_\mathrm{z}(s) = \int_\mathrm{z} s / |\mathrm{z}|$. For vector-valued functions $\mathbf{s} \in [L^1(\mathrm{z})]^d$ the projection is applied component-wise: $\pi_\mathrm{z}(\mathbf{s}) := (\pi_\mathrm{z}(s_i))_{i=1,\dots,d}$.

In the CDO-Fb framework the velocity is hybrid, meaning that it has cell- and face-based degrees of freedom (DoFs). Hence, the global velocity space is

$$\widehat{\mathbf{U}}_\mathrm{h} := \bigtimes_{\mathrm{c}\in\mathrm{C}} [\mathbb{P}^0(\mathrm{c})]^d \times \bigtimes_{\mathrm{f}\in\mathrm{F}} [\mathbb{P}^0(\mathrm{f})]^d \,. \tag{2}$$

An element of $\widehat{\mathbf{U}}_\mathrm{h}$ is denoted by $\widehat{\mathbf{u}}_\mathrm{h} := \big((\mathbf{u}_\mathrm{c})_{\mathrm{c}\in\mathrm{C}}, (\mathbf{u}_\mathrm{f})_{\mathrm{f}\in\mathrm{F}}\big)$, where, for a generic $\mathrm{z} = \mathrm{c}, \mathrm{f}$, $\mathbf{u}_\mathrm{z}$ is the z-based DoF. Notice that the value at the interfaces is uniquely defined. The velocity DoFs associated with a cell c are denoted $\widehat{\mathbf{u}}_\mathrm{c} := \big(\mathbf{u}_\mathrm{c}, (\mathbf{u}_\mathrm{f})_{\mathrm{f}\in\mathrm{F}_\mathrm{c}}\big) \in \widehat{\mathbf{U}}_\mathrm{c} := [\mathbb{P}^0(\mathrm{c})]^d \times \bigtimes_{\mathrm{f}\in\mathrm{F}_\mathrm{c}} [\mathbb{P}^0(\mathrm{f})]^d$. The pressure is defined at cells only: $P_\mathrm{h} := \bigtimes_{\mathrm{c}\in\mathrm{C}} P_\mathrm{c} \ni p_\mathrm{h} := (p_\mathrm{c})_{\mathrm{c}\in\mathrm{C}}$, where $P_\mathrm{c} := \mathbb{P}^0(\mathrm{c})$. In order to take into consideration the velocity BC and the constraint on the pressure average, one also needs $\widehat{\mathbf{U}}_{\mathrm{h},0} := \big\{\widehat{\mathbf{u}}_\mathrm{h} \in \widehat{\mathbf{U}}_\mathrm{h} \mid \mathbf{u}_\mathrm{f} = 0 \forall \mathrm{f} \in \mathrm{F}^b\big\}$, $P_{\mathrm{h},*} := \big\{p_\mathrm{h} \in P_\mathrm{h} \mid \sum_{\mathrm{c}\in\mathrm{C}} |\mathrm{c}|\, p_\mathrm{c} = 0\big\}$. The right part of Fig. 1 gives an example of local velocity and pressure DoFs for a cell.

2.1.1 Discrete Velocity Gradient and Divergence

For each cell $\mathrm{c} \in \mathrm{C}$, the discrete local gradient G_c is piecewise constant on the pyramid partition $\big\{\mathfrak{p}_{\mathrm{f,c}}\big\}_{\mathrm{f}\in\mathrm{F}_\mathrm{c}}$ (cf. central part of Fig. 1) and is defined as follows: G_c: $\widehat{\mathbf{U}}_\mathrm{c} \to [\mathbb{P}^0(\big\{\mathfrak{p}_{\mathrm{f,c}}\big\}_{\mathrm{f}\in\mathrm{F}_\mathrm{c}})]^{d\times d}$ such that for all $\mathrm{f} \in \mathrm{F}_\mathrm{c}$

$$\mathsf{G}_\mathrm{c}(\widehat{\mathbf{u}}_\mathrm{c})_{|\mathfrak{p}_{\mathrm{f,c}}} := \mathsf{G}^0_\mathrm{c}(\widehat{\mathbf{u}}_\mathrm{c}) + \beta \frac{|\mathrm{f}|}{|\mathfrak{p}_{\mathrm{f,c}}|} \left((\mathbf{u}_\mathrm{f} - \mathbf{u}_\mathrm{c}) - \mathsf{G}^0_\mathrm{c}(\widehat{\mathbf{u}}_\mathrm{c})\,(\mathbf{x}_\mathrm{f} - \mathbf{x}_\mathrm{c}) \right) \otimes \mathbf{n}_\mathrm{fc}\ , \qquad (3)$$

where $\beta > 0$ is an arbitrary stability parameter and $\mathsf{G}^0_\mathrm{c}(\widehat{\mathbf{u}})$ is a $\mathbb{P}_0$-consistent gradient, constant inside each cell and defined as $\mathsf{G}^0_\mathrm{c}(\widehat{\mathbf{u}}_\mathrm{c}) := 1/\,|\mathrm{c}| \sum_{\mathrm{f}\in\mathrm{F_c}} |\mathrm{f}|\,(\mathbf{u}_\mathrm{f} - \mathbf{u}_\mathrm{c}) \otimes \mathbf{n}_\mathrm{fc}$. The definition (3) is the vector-valued version of the gradient introduced in [2]. In the numerical tests, we will use $\beta = 1$, which recovers the GCR framework [10]; the choice $\beta = 1/\sqrt{d}$ gives the HMM one [12].

For each cell $\mathrm{c} \in \mathrm{C}$, the discrete velocity divergence $D_\mathrm{c} \colon \widehat{\mathbf{U}}_\mathrm{c} \to \mathbb{P}^0(\mathrm{c})$ is defined as follows

$$D_\mathrm{c}(\widehat{\mathbf{u}}_\mathrm{c}) := \mathrm{trace}\left(\mathsf{G}^0_\mathrm{c}(\widehat{\mathbf{u}}_\mathrm{c})\right) = \frac{1}{|\mathrm{c}|} \sum_{\mathrm{f}\in\mathrm{F_c}} |\mathrm{f}|\,\mathbf{u}_\mathrm{f} \cdot \mathbf{n}_\mathrm{fc}\ . \qquad (4)$$

Notice that only the face-based DoFs are used (since faces are planar). The discrete velocity divergence is the tool on which the velocity-pressure coupling hinges. This divergence operator can be found also in the HMM framework [11].

2.1.2 Discrete Advection Scheme

The design of the advection scheme is inspired by HHO($k = 0$) [9]. We aim at discretizing the classical advective trilinear form such that $\int_\mathscr{D} ((\mathbf{w}\cdot\nabla)\mathbf{u})\cdot\mathbf{v}$. Given $\widehat{\mathbf{u}}_\mathrm{h}, \widehat{\mathbf{v}}_\mathrm{h}, \widehat{\mathbf{w}}_\mathrm{h} \in \widehat{\mathbf{U}}_{\mathrm{h},0}$, we use

$$\begin{aligned} a_\mathrm{adv}(\widehat{\mathbf{w}}_\mathrm{h}; \widehat{\mathbf{u}}_\mathrm{h}, \widehat{\mathbf{v}}_\mathrm{h}) := {} & \frac{1}{2} \sum_{\mathrm{c}\in\mathrm{C}} \sum_{\mathrm{f}\in\mathrm{F_c}} |\mathrm{f}|\,(\mathbf{w}_\mathrm{f}\cdot\mathbf{n}_\mathrm{fc})(\mathbf{u}_\mathrm{f} - \mathbf{u}_\mathrm{c})(\mathbf{v}_\mathrm{f} + \mathbf{v}_\mathrm{c}) \\ & + \theta^\mathrm{upw} \sum_{\mathrm{f}\in\mathrm{F}^i} \sum_{\mathrm{c}\in\mathrm{C_f}} |\mathrm{f}|\,|\mathbf{w}_\mathrm{f}\cdot\mathbf{n}_\mathrm{f}|\,(\mathbf{u}_\mathrm{f} - \mathbf{u}_\mathrm{c})(\mathbf{v}_\mathrm{f} - \mathbf{v}_\mathrm{c})\ , \end{aligned} \qquad (5)$$

where $\theta^\mathrm{upw} := 1$ in one wants a stabilization by upwinding, or $\theta^\mathrm{upw} := 0$ for a centered scheme. Suppose, for now, that $\theta^\mathrm{upw} = 0$. One has:

$$a_\mathrm{adv}(\widehat{\mathbf{w}}_\mathrm{h}; \widehat{\mathbf{u}}_\mathrm{h}, \widehat{\mathbf{v}}_\mathrm{h}) + a_\mathrm{adv}(\widehat{\mathbf{w}}_\mathrm{h}; \widehat{\mathbf{v}}_\mathrm{h}, \widehat{\mathbf{u}}_\mathrm{h}) = -\sum_{\mathrm{c}\in\mathrm{C}} |\mathrm{c}|\,D_\mathrm{c}(\widehat{\mathbf{w}}_\mathrm{c})\mathbf{u}_\mathrm{c}\cdot\mathbf{v}_\mathrm{c} + \sum_{\mathrm{f}\in\mathrm{F}^b} |\mathrm{f}|\,(\mathbf{w}_\mathrm{f}\cdot\mathbf{n}_\mathrm{f})\mathbf{u}_\mathrm{f}\cdot\mathbf{v}_\mathrm{f} \qquad (6)$$

obtained by using (4) and by discarding the internal face-defined DoFs since they sum to zero. The boundary DoFs are kept in order to better show that (6) is the discrete counterpart of a known integral-by-parts result. Plugging $\widehat{\mathbf{v}}_\mathrm{h} = \widehat{\mathbf{u}}_\mathrm{h}$ into (6) one obtains

$$a_\mathrm{adv}(\widehat{\mathbf{w}}_\mathrm{h}; \widehat{\mathbf{u}}_\mathrm{h}, \widehat{\mathbf{u}}_\mathrm{h}) = -\frac{1}{2} \sum_{\mathrm{c}\in\mathrm{C}} |\mathrm{c}|\,D_\mathrm{c}(\widehat{\mathbf{w}}_\mathrm{c})\mathbf{u}_\mathrm{c}^2\ . \qquad (7)$$

Supposing there exists $\mu > 0$ such that $-1/2\, D_{\mathrm{c}}(\widehat{\mathbf{w}}_{\mathrm{c}}) \geq \mu$ for all $\mathrm{c} \in \mathrm{C}$ (this is a discrete counterpart of the well-known stability hypothesis of the continuous advection problem), then (7) proves the coercivity of $a_{\mathrm{adv}}(\widehat{\mathbf{w}}_{\mathrm{h}}; \cdot, \cdot)$.

2.2 Discrete Bilinear Form

The discrete counterpart of problem (1a) stemming from the CDO-Fb scheme writes: Find $(\widehat{\mathbf{u}}_{\mathrm{h}},\ p_{\mathrm{h}}) \in \widehat{\mathbf{U}}_{\mathrm{h},0} \times P_{\mathrm{h},*}$ such that, $\forall \widehat{\mathbf{v}}_{\mathrm{h}} \in \widehat{\mathbf{U}}_{\mathrm{h},0}$ and $\forall q_{\mathrm{h}} \in P_{\mathrm{h},*}$

$$\sum_{\mathrm{c}\in\mathrm{C}} \int_{\mathrm{c}} \{\nu \mathsf{G}_{\mathrm{c}}(\widehat{\mathbf{u}}_{\mathrm{c}}) : \mathsf{G}_{\mathrm{c}}(\widehat{\mathbf{v}}_{\mathrm{c}}) - p_{\mathrm{c}} D_{\mathrm{c}}(\widehat{\mathbf{v}}_{\mathrm{c}})\} + \chi a_{\mathrm{adv}}(\widehat{\mathbf{u}}_{\mathrm{h}}; \widehat{\mathbf{u}}_{\mathrm{h}}, \widehat{\mathbf{v}}_{\mathrm{h}}) = \sum_{\mathrm{c}\in\mathrm{C}} \int_{\mathrm{c}} \mathbf{f} \cdot \mathbf{v}_{\mathrm{c}} \ , \tag{8a}$$

$$\sum_{\mathrm{c}\in\mathrm{C}} -D_{\mathrm{c}}(\widehat{\mathbf{u}}_{\mathrm{c}}) q_{\mathrm{c}} = 0 \ . \tag{8b}$$

The Stokes problem ($\chi = 0$ in (1a)) has been analyzed in [11].

A static condensation procedure eliminating the cell-based velocity DoFs can be performed in order to reduce the size of the global system, which thus becomes $d\,\mathrm{Card}(\mathrm{F}) + \mathrm{Card}(\mathrm{C})$. The discarded DoFs are recovered after the solving stage, as a post-processing.

3 Numerical Results

The proposed framework is validated on four test cases, two for the Stokes equations (in 2D and 3D), and two for the NSE (both in 2D). When considering the latter, the nonlinear equations are solved by Picard iterations, and the stopping criterion is evaluated using the cell-based, discrete L^2-norm of the increment, namely $\left\|\widehat{\mathbf{u}}_{\mathrm{h}}^{k} - \widehat{\mathbf{u}}_{\mathrm{h}}^{k-1}\right\|_{\mathrm{C}} / \left\|\widehat{\mathbf{u}}_{\mathrm{h}}^{k-1}\right\|_{\mathrm{C}} < \varepsilon$, where $\|\widehat{\mathbf{u}}_{\mathrm{h}}\|_{\mathrm{C}}^2 := \sum_{\mathrm{c}\in\mathrm{C}} |\mathrm{c}|\, \|\mathbf{u}_{\mathrm{c}}\|_2^2$. When computing the errors, this velocity norm is considered, as well as the norm of the velocity gradient $\|\widehat{\mathbf{u}}_{\mathrm{h}}\|_{\mathsf{G},\mathrm{C}}^2 := \sum_{\mathrm{c}\in\mathrm{C}} |\mathrm{c}|\, \|\mathsf{G}(\widehat{\mathbf{u}}_{\mathrm{c}})\|_2^2$ and the discrete L^2-norm of the pressure$\|p_{\mathrm{h}}\|_{\mathrm{C}}^2 := \sum_{\mathrm{c}\in\mathrm{C}} |\mathrm{c}|\, p_{\mathrm{c}}^2$. The resulting error norms used in the analysis are:

$$\mathrm{erru} := \frac{\|\widehat{\mathbf{u}}_{\mathrm{h}} - \widehat{\pi}_{\mathrm{h}}(\mathbf{u})\|_{\mathrm{C}}}{\|\widehat{\pi}_{\mathrm{h}}\mathbf{u}\|_{\mathrm{C}}} \ , \ \mathrm{errgu} := \frac{\|\widehat{\mathbf{u}}_{\mathrm{h}} - \widehat{\pi}_{\mathrm{h}}(\mathbf{u})\|_{\mathsf{G},\mathrm{C}}}{\|\widehat{\pi}_{\mathrm{h}}(\mathbf{u})\|_{\mathsf{G},\mathrm{C}}} \ , \ \mathrm{errp} := \frac{\|p_{\mathrm{h}} - \pi_{\mathrm{h}}(p)\|_{\mathrm{C}}}{\|\pi_{\mathrm{h}}(p)\|_{\mathrm{C}}} \ , \tag{9}$$

where $\widehat{\pi}_{\mathrm{h}}(\mathbf{u}) := ((\pi_{\mathrm{c}}(\mathbf{u}))_{\mathrm{c}\in\mathrm{C}},\ (\pi_{\mathrm{f}}(\mathbf{u}))_{\mathrm{f}\in\mathrm{F}})$ and $\pi_{\mathrm{h}}(p) := (\pi_{\mathrm{c}}(p))_{\mathrm{c}\in\mathrm{C}}$. Let nuu (resp. npu) stand for the number of velocity (pressure) unknowns. They will be used to evaluate the orders of convergence in space.

We will use the CDO implementation available via *Code_Saturne* [1], an open-source multi-purpose CFD solver developed at EDF R&D. The computations have

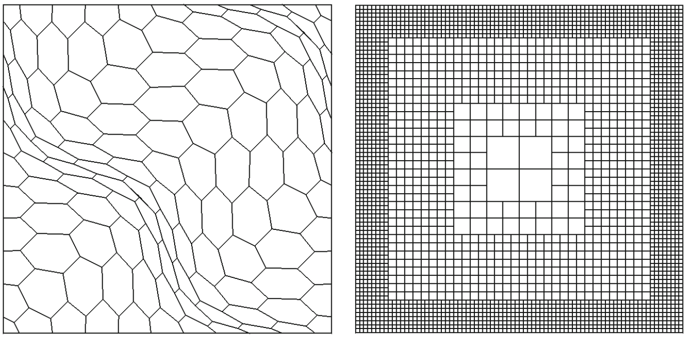


Fig. 2 Examples of 2D meshes. **Left** polygonal. **Right** progressively refined cartesian

been performed on an octa-core, Intel i7 laptop with 32GB RAM using PETSc and MUMPS libraries to solve the linear systems.

3.1 Stokes Equations

Two test cases are considered for the Stokes equations ($\chi = 0$ in (1a)).

2D Bercovier–Engelman test case It is proposed in the test case 2.1 of the benchmark [7]. The sequence of Cartesian meshes (denoted by Hn where n is the number of segments an edge of the domain is divided into) from [7] and a 2D polygonal family (similarly denoted by Pn, cf. left part of Fig. 2) have been considered. The results are collected in Table 1.

3D Taylor–Green vortex This test case corresponds to Sect. 2.2 of the benchmark [7]. The meshes used were the Cartesian (Hn) and prismatic with triangular bases (PrTn) sequences proposed in [7], and one composed of tetrahedra (Tn, the refine-

Table 1 Errors for the 2D Bercovier—Engelman test case. Cartesian and polygonal meshes

Mesh	nuu	npu	errgu	Order	erru	Order	errp	Order
H32	4224	1024	9.15×10^{-4}	–	7.71×10^{-4}	–	1.06×10^{-1}	–
H64	16640	4096	3.16×10^{-4}	1.55	1.93×10^{-4}	2.02	2.87×10^{-2}	1.98
H128	66048	16384	1.35×10^{-4}	1.24	4.82×10^{-5}	2.01	7.36×10^{-3}	1.99
H256	263168	65536	6.41×10^{-5}	1.07	1.21×10^{-5}	2.01	1.85×10^{-3}	1.99
P10	720	121	9.99×10^{-2}	–	3.33×10^{-2}	–	2.63×10^{0}	–
P20	2640	441	5.81×10^{-2}	0.83	9.07×10^{-3}	2.00	7.42×10^{-1}	1.96
P30	5760	961	4.07×10^{-2}	0.91	4.07×10^{-3}	2.05	3.36×10^{-1}	2.03
P40	10080	1681	3.13×10^{-2}	0.94	2.29×10^{-3}	2.05	1.91×10^{-1}	2.02

Table 2 Errors for the 3D Taylor–Green Vortex. Cartesian, tetrahedral and prismatic meshes

Mesh	nuu	npu	errgu	Order	erru	Order	errp	Order
H4	720	64	4.36×10^{-1}	–	3.18×10^{-1}	–	4.83×10^{-1}	–
H8	5184	512	2.60×10^{-1}	0.79	1.05×10^{-1}	1.69	1.49×10^{-1}	1.70
H16	39168	4096	1.36×10^{-1}	0.96	2.82×10^{-2}	1.95	3.95×10^{-2}	1.92
H32	304128	32768	6.91×10^{-2}	1.00	7.18×10^{-3}	2.00	1.00×10^{-2}	1.98
T6	15090	2383	3.39×10^{-1}	–	8.99×10^{-2}	–	1.45×10^{-1}	–
T12	117552	19064	1.76×10^{-1}	0.96	2.39×10^{-2}	1.94	5.65×10^{-2}	1.36
T24	927744	152512	8.86×10^{-2}	1.00	6.07×10^{-3}	1.99	2.49×10^{-2}	1.18
PrT10	16200	2000	3.12×10^{-1}	–	9.18×10^{-2}	–	1.64×10^{-1}	–
PrT20	124800	16000	1.67×10^{-1}	0.92	2.72×10^{-2}	1.79	6.60×10^{-2}	1.32
PrT30	415800	54000	1.13×10^{-1}	0.97	1.28×10^{-2}	1.88	3.96×10^{-2}	1.26
PrT40	979200	128000	8.54×10^{-2}	0.99	7.40×10^{-3}	1.92	2.81×10^{-2}	1.20

ment is achieved by dividing each tetrahedra into 8 subtetrahedra). The results are collected in Table 2.

3.2 Navier–Stokes Equations

Two test cases are considered for the Navier–Stokes equations ($\chi = 1$ in (1a)).

Burggraf flow It consists in a manufactured polynomial solution of the 2D NSE presented in [6]. The centered scheme was considered ($\theta^{\mathrm{upw}} = 0$ in (5)). The viscosity is $\nu = 1/100$. About 15 Picard iterations were needed to reach the prescribed tolerance $\varepsilon = 10^{-7}$. Two sequences of meshes have been considered: the Cartesian one (Hn) from [7], and one composed of nonmatching squares (HRn, cf. right part of Fig. 2), obtained by progressively refining the Cartesian meshes. The results are collected in Table 3.

Table 3 Errors for the 2D Burggraf flow. Cartesian and refined cartesian meshes

Mesh	nuu	npu	errgu	Order	erru	Order	errp	Order
H32	4224	1024	2.40×10^{-1}	–	1.73×10^{-2}	–	1.33×10^{-2}	–
H64	16640	4096	1.20×10^{-1}	1.01	4.35×10^{-3}	2.01	3.39×10^{-3}	1.98
H128	66048	16384	6.01×10^{-2}	1.00	1.09×10^{-3}	2.00	8.53×10^{-4}	1.99
H256	263168	65536	3.01×10^{-2}	1.00	2.73×10^{-4}	2.02	2.14×10^{-4}	1.99
HR80	12984	3124	1.74×10^{-1}	–	2.19×10^{-2}	–	2.38×10^{-2}	–
HR160	50960	12496	8.85×10^{-2}	0.99	5.95×10^{-3}	1.90	6.59×10^{-3}	1.85
HR320	201888	49984	4.44×10^{-2}	1.00	1.51×10^{-3}	1.99	1.69×10^{-3}	1.96

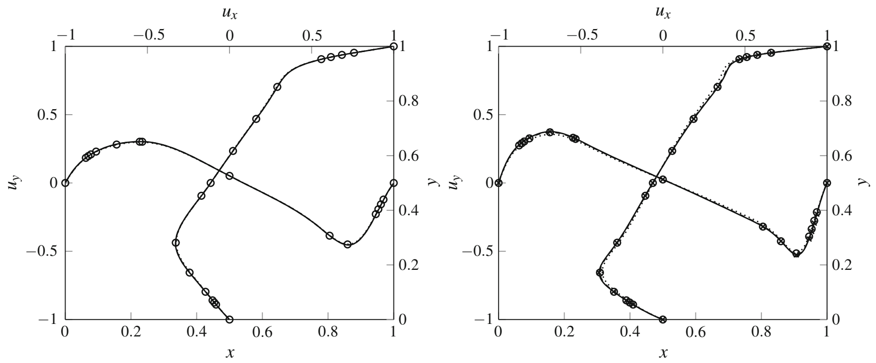


Fig. 3 Lid-driven cavity, vertical and horizontal velocity profiles at the axis of symmetry. Data: CDO H127 (*dotted line*), CDO H255 (*dashed line*), CDO H511 (*solid line*), [13] (*circle*), [5] (*cross*). **Left** $\nu = 1/400$. **Right** $\nu = 1/1000$

2D lid-driven cavity It is proposed in the test case 6 of the benchmark [7]. Two values of the viscosity have been considered: $\nu = 1/400,\ 1/1000$. Computations have been run on Cartesian meshes with edges divided into 127, 255, and 511 segments. The centered scheme was considered ($\theta^{\text{upw}} = 0$ in (5)). The prescribed tolerance for the Picard iterations is $\varepsilon = 10^{-7}$, less than 25 iterations were needed for $\nu = 1/400$ and less than 30 for 1/1000. In Fig. 3, one can find the plots of the computed vertical and horizontal velocity profiles on the symmetry axes for three Cartesian meshes as well as those from Refs. [5, 13]. Some computations have been run with an upwind scheme ($\theta^{\text{upw}} = 1$ in (5)) for the advection term, and the results on the velocity profiles were less accurate on the coarser meshes than those obtained with the centered one.

References

1. Archambeau, F., Méchitoua, N., Sakiz, M.: Code_Saturne: a finite volume code for turbulent flows—industrial applications. Int. J. Finite **1**(1) (2004)
2. Bonelle, J.: Compatible discrete operator schemes on polyhedral meshes for elliptic and Stokes equations. Ph.D. thesis, Université Paris-Est (2014)
3. Bonelle, J., Ern, A.: Analysis of compatible discrete operator schemes for elliptic problems on polyhedral meshes. ESAIM Math. Model. Numer. Anal. **48**(2), 553–581 (2014)
4. Bonelle, J., Ern, A.: Analysis of compatible discrete operator schemes for the Stokes equations on polyhedral meshes. IMA J. Numer. Anal. **35**(4), 1672–1697 (2015)
5. Botella, O., Peyret, R.: Benchmark spectral results on the lid-driven cavity flow. Comput. Fluids **27**(4), 421–433 (1998)
6. Burggraf, O.R.: Analytical and numerical studies of the structure of steady separated flows. J. Fluid Mech. **24**(1), 113–151 (1966)
7. Cancès, C., Omnes, P. (eds.): Finite Volumes for Complex Applications VIII—Methods and Theoretical Aspects, Springer Proceedings in Mathematics & Statistics, vol. 199. Springer International Publishing, Lille, France (2017)

8. Cantin, P., Ern, A.: Vertex-based compatible discrete operator schemes on polyhedral meshes for advection-diffusion equations. Comput. Methods Appl. Math. **16**(2), 187–212 (2016)
9. Di Pietro, D.A., Droniou, J., Ern, A.: A discontinuous-skeletal method for advection-diffusion-reaction on general meshes. SIAM J. Numer. Anal. **53**(5), 2135–2157 (2015)
10. Di Pietro, D.A., Lemaire, S.: An extension of the Crouzeix-Raviart space to general meshes with application to quasi-incompressible linear elasticity and Stokes flow. Math. Comput. **84**(291), 1–31 (2015)
11. Droniou, J., Eymard, R., Feron, P.: Gradient schemes for Stokes problem. IMA J. Numer. Anal. **36**(4), 1636–1669 (2015)
12. Droniou, J., Eymard, R., Gallouët, T., Herbin, R.: A unified approach to mimetic finite difference, hybrid finite volume and mixed finite volume methods. Math. Model. Methods Appl. Sci. **20**(2), 265–295 (2010)
13. Ghia, U., Ghia, K.N., Shin, C.T.: High-Re solutions for incompressible flow using the Navier–Stokes equations and a multigrid method. J. Comput. Phys. **48**(3), 387–411 (1982)

On the Significance of Pressure-Robustness for the Space Discretization of Incompressible High Reynolds Number Flows

Alexander Linke and Christian Merdon

Abstract Only recently, strong gradient fields in the momentum balance of incompressible flows have been identified as a common major source for numerical errors in flow solvers. The novel notion of pressure-robustness denotes those space discretizations that behave in a robust manner with respect to strong gradient fields. This contribution elaborates on certain advantages of pressure-robust solvers versus standard solvers: (i) the asymptotic convergence rate of pressure-robust solvers may be reached on much coarser grids than for standard solvers; (ii) certain preasymptotic convergence-rates may be provably suboptimal for standard solvers; thus, low-order pressure-robust solver can outperform high-order classical solvers on coarse grids. Last but not least, the contribution explains how strong gradient fields develop in complex incompressible flows.

Keywords Incompressible Navier–Stokes equations · Pressure-robustness · High Reynolds numbers · Vortex-dominated flows

MSC (2010) 76D05 · 65N12 · 65N30

1 Introduction

Strong gradient fields in the momentum balance have been identified in recent years as a common source for numerical errors for standard solvers of the incompressible Navier–Stokes equations

A. Linke (✉) · C. Merdon
Weierstrass Institute for Applied Analysis and Stochastics,
Mohrenstraße 39, 10117 Berlin, Germany
e-mail: alexander.linke@wias-berlin.de

C. Merdon
e-mail: christian.merdon@wias-berlin.de

R. Klöfkorn et al. (eds.), *Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples*, Springer Proceedings in Mathematics & Statistics 323,
https://doi.org/10.1007/978-3-030-43651-3_7

$$\begin{aligned} \mathbf{u}_t - \nu\Delta\mathbf{u} + (\mathbf{u}\cdot\nabla)\,\mathbf{u} + \nabla p &= \mathbf{f}, \\ \nabla\cdot\mathbf{u} &= 0, \end{aligned} \tag{1}$$

see [8, 11, 12]. In the following, the notion *standard solvers* denotes classical discretely inf-sup stable space discretizations like the conforming Taylor–Hood element [6]. However, all the arguments in this contribution can be extended to many other space discretizations like the nonconforming Crouzeix–Raviart element [4] or standard Discontinuous Galerkin discretizations [14]; also extensions to space discretizations that are not discretely inf-sup stable, but apply some kind of pressure-stabilization [7] can be shown to fit into this framework.

1.1 Pressure-Robustness and Velocity-Equivalence

The importance of gradient field forces in (1) is quickly explained. Assuming that $(\mathbf{u}, p)$ solves (1) for a forcing $\mathbf{f}$, then $(\mathbf{u}, p + \phi)$ will solve (1) for the modified forcing $\mathbf{f} + \nabla\phi$,—where $\nabla\phi$ is an arbitrary gradient field. Thus, the forcings $\mathbf{f}$ and $\mathbf{f} + \nabla\phi$ are in a sense *velocity-equivalent* [5], i.e.,

$$\mathbf{f} \simeq \mathbf{f} + \nabla\phi, \tag{2}$$

since they induce the very same velocity solution $\mathbf{u}$. In other words, the additional forcing $\nabla\phi$ will be balanced completely by a modified pressure gradient $\nabla(p + \phi)$.

This purely mathematical observation allows for a couple of distinct physical regimes, i.e., special solutions of (1), where gradient forces play a dominant role in the Navier–Stokes momentum balance. Well-known is the hydrostatic regime—which is definitely not a complex flow—, assuming homogeneous Dirichlet velocity boundary conditions. For a gradient field forcing $\mathbf{f} = \nabla\phi$, where ϕ denotes an arbitrary gravitational potential, the hydrostatic solution of (1) is just given by:

$$\begin{aligned} \mathbf{u} &= \mathbf{0}, \\ p &= \phi + \text{const}, \end{aligned} \tag{3}$$

where "const" denotes some constant assuring a unique solution of the problem.

The following examples demonstrates the behavior of certain standard solvers with a different (formal) convergence order for a series of hydrostatic problems with varying complexity of the gravitational potential ϕ and compares their accuracy with a lowest-order pressure-robust solver.

Resembling hydrostatics in a glass of water—, consider the incompressible Stokes problem $-\Delta\mathbf{u} + \nabla p = \nabla\phi$, $\nabla\cdot\mathbf{u} = 0$ with homogeneous Dirichlet velocity boundary conditions, where it holds $\phi(x, y) = y^k$ with $k = 1, 2, 4, 8$. The exact velocity solution reads $\mathbf{u} = 0$, and the numerical results calculated on a unstructured grid are depicted in Fig. 1.

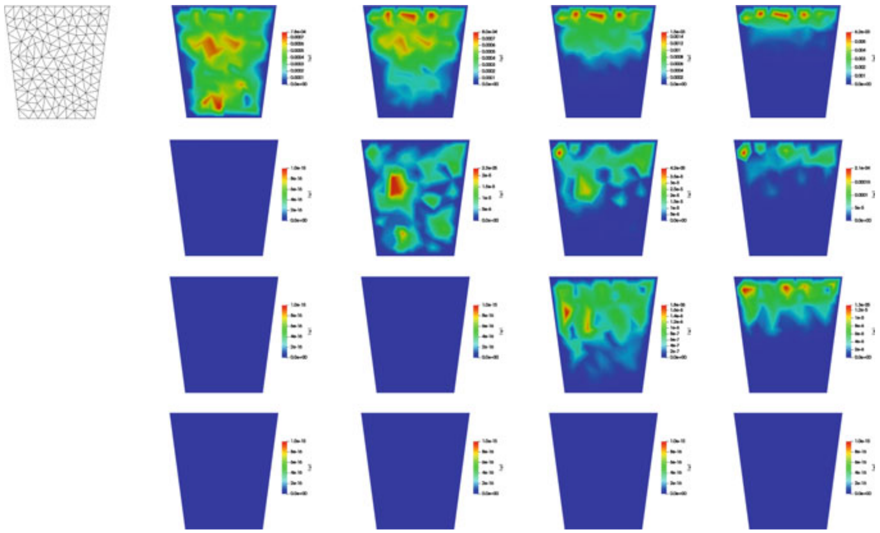


Fig. 1 Numerical errors in water glass calculations for the Bernardi–Raugel method (first row), Taylor–Hood method (second row), cubic Taylor–Hood method (third row) and a modified, pressure-robust Bernardi–Raugel method (last row) for a pressure force with potential $k = 1$ (first column), $k = 2$ (second column), $k = 4$ (third column) and $k = 8$ (last column). The underlying grid for all computations is shown in the upper left corner

Although the numerical example is rather simple, it is nonetheless telling something important: in the first three rows, the results of standard solvers are presented: a first-order Bernardi–Raugel element, a second and a third-order Taylor–Hood element, which are all standard solvers in the sense above.

As a rule of thumb, we can infer from the results: high-order standard solvers are in general more accurate than low-order standard solvers, i.e., high-order standard solvers are able to balance gradient fields $\nabla\phi$ up to a higher polynomial degree. However, in the last row, the numerical results for the first-order pressure-robust, modified Bernardi–Raugel element [10, 12] are presented, and they all deliver the correct velocity hydrostatic velocity solution $\mathbf{u} = \mathbf{0}$. Thus, low-order pressure-robust solvers can be as accurate as high-order standard solvers, whenever strong gradient fields dominate the Navier–Stokes momentum balance. Since the hydrostatic case is rather trivial, we will argue in the following that dominant gradient fields generally appear in typical high Reynolds number flows with $\mathbf{f} = \mathbf{0}$. Thus, also in high Reynolds number flows low-order pressure-robust methods can be competetive with high-order standard solvers, on preasymptotic grids [5].

1.2 Pressure-Robustness and Vorticity Equation

An intuitive understanding for the distinctive significance of gradient field forces in the incompressible Navier–Stokes momentum balance derives from the famous div-curl-problem [2, 8]. A sufficiently smooth vector field $\mathbf{w}$ on a sufficiently smooth, simply-connected domain Ω with a finite number of closed, disjoint surfaces can be determined by prescribing its divergence $\nabla \cdot \mathbf{w}$, its curl $\nabla \times \mathbf{w}$ and some boundary data for the normal component $\mathbf{w} \cdot \mathbf{n}$ at the boundary $\partial\Omega$ of the domain, i.e., $\mathbf{w}$ is determined by

$$\begin{aligned} \nabla \cdot \mathbf{w} &= g, && \mathbf{x} \in \Omega, \\ \nabla \times \mathbf{w} &= \chi, && \mathbf{x} \in \Omega, \\ \mathbf{w} \cdot \mathbf{n} &= b, && \mathbf{x} \in \partial\Omega, \end{aligned} \tag{4}$$

where as assumptions for the data $\chi \in C^1(\Omega)$ and $\nabla \cdot \chi = 0$ and $\int_\Omega g \, dx = \int_{\partial\Omega} b \, \mathrm{dS}$ hold.

Applying this knowledge to the velocity solution $\mathbf{u}$ of the incompressible Navier–Stokes equation (1), we recognize that two of three quantities are known: it holds $\nabla \cdot \mathbf{u} = 0$ and the boundary data of $\mathbf{u}$ is also known, assuming homogeneous Dirichlet velocity boundary conditions. The only missing information about $\mathbf{u}$ is its vorticity

$$\omega := \nabla \times \mathbf{u}. \tag{5}$$

An evolution for the vorticity equation can be formally derived by applying the curl operator to (1), yielding formally

$$\omega_t - \nu \Delta \omega + (\mathbf{u} \cdot \nabla)\, \omega - (\omega \cdot \nabla)\, \mathbf{u} = \nabla \times \mathbf{f}. \tag{6}$$

This vorticity equation tells us that all gradient parts—in the sense of the Helmholtz decomposition [8]—of the forces $\mathbf{f}$, $(\mathbf{u} \cdot \nabla)\, \mathbf{u}$, $-\nu \Delta \mathbf{u}$, and $\mathbf{u}_t$ do not contain any information for the determination of $\mathbf{u}$. These gradient parts rather determine ∇p. So, an important issue in the space discretization of the incompressible Navier–Stokes equations is to derive a space discretization, where the vorticity equation is discretized implicitly in an accurate manner, such that

$$\nabla \times \nabla \psi = \mathbf{0} \quad \text{for arbitrary (!) gradient fields } \nabla \psi. \tag{7}$$

This is the very goal of pressure-robust space discretizations.

1.3 Pressure-Robustness and H(div)-Conforming FEM Spaces

The statement (7) concerns the strong formulation of (1): For a finite element or Discontinuous Galerkin method, one has to translate (7) into a weak setting, with appropriate test functions. Therefore, we multiply (7) with a test function $\mathbf{v}$ with compact support and integrate over the domain, yielding:

$$0 = \int \mathbf{v} \cdot (\nabla \times \nabla\psi)\ \mathrm{dx} = \int (\nabla \times \mathbf{v}) \cdot \nabla\psi\, \mathrm{dx}. \tag{8}$$

Due to the vector calculus identity $\nabla \cdot (\nabla \times \mathbf{v}) = \mathbf{0}$, the identity (7) is nothing else than the $\mathbf{L}^2$-orthogonality of divergence-free vector fields and gradient fields, which is at the basis of the Helmholtz decomposition [8]. Further, every divergence-free vector field $\mathbf{v} \in \mathbf{H}(\mathrm{div})$ with a vanishing normal component $\mathbf{v} \cdot \mathbf{n} = 0$ along $\partial\Omega$ is $\mathbf{L}^2$-orthogonal to arbitrary gradient fields $\nabla\psi$ with $\psi \in H^1$ due to

$$\int \nabla\psi \cdot \mathbf{v}\, dx = -\int \psi \nabla \cdot \mathbf{v}\, dx = 0. \tag{9}$$

This powerful $\mathbf{L}^2$-orthogonality can be exploited to construct pressure-robust discretizations [9]. For example, a standard finite element discretization for the incompressible Stokes problem can be made pressure-robust by replacing the spatial discretization of the right hand side via

$$\int \mathbf{f} \cdot \mathbf{v}_h\, \mathrm{dx} \to \int \mathbf{f} \cdot \mathbf{I}_h \mathbf{v}_h\, \mathrm{dx}. \tag{10}$$

Here, $\mathbf{I}_h$ denotes a (locally defined) interpolation operator, mapping vector-valued finite element functions to $\mathbf{H}(\mathrm{div})$-conforming vector fields such that it holds

$$\nabla \cdot (\mathbf{I}_h \mathbf{v}_h) = \mathrm{div}_h \mathbf{v}_h. \tag{11}$$

Especially discretely divergence-free vector fields are mapped to divergence-free ones in the sense of $\mathbf{H}(\mathrm{div})$ [1].

2 How Do Strong Gradient Field Forces Develop in High Reynolds Number Flows Incompressible Flows?

Regarding the incompressible Euler equations

$$\begin{aligned} \mathbf{u}_t + (\mathbf{u} \cdot \nabla)\, \mathbf{u} + \nabla p &= \mathbf{f}, \\ \nabla \cdot \mathbf{u} &= 0, \end{aligned} \tag{12}$$

one recognizes that even in the case $\mathbf{f} = \mathbf{0}$, the material derivative of the Euler equations is a gradient field, i.e.,

$$\frac{D\mathbf{u}}{\mathrm{dt}} := \mathbf{u}_t + (\mathbf{u} \cdot \nabla)\,\mathbf{u} = -\nabla p. \tag{13}$$

This leads to the important observation that, e.g., in high Reynolds number Navier–Stokes flows with obstacles, the material derivative is approximately a nontrivial (!) and complicated gradient field, see [5]. Moreover, (13) tells us something interesting: a quadratic, velocity-dependent term on the left hand side is balanced by a linear pressure-dependent term. Thus, the pressure p is in general more complicated to approximate by, e.g., piecewise polynomial functions, than the velocity $\mathbf{u}$ [5].

This holds in general, since fundamental objects of fluid mechanics like vortices, vortex rings or vortex filaments can be generated as Galilean-invariant transformations of steady solutions of the incompressible Euler equations, fulfilling

$$\begin{aligned} (\mathbf{u} \cdot \nabla)\,\mathbf{u} + \nabla p &= \mathbf{0}, \\ \nabla \cdot \mathbf{u} &= 0, \end{aligned} \tag{14}$$

see [3] for examples. In the language of (2), standing vortices, vortex rings, …have a strong convection term that is velocity-equivalent to $\mathbf{0}$, thus there is no dominant convection. In [5], the notion of *pseudo-dominant convection* is introduced for such situations.

This observation is highly important for the issue of convection stabilisation. Here, we remark that convection stabilization like an upwind mechanism is only necessary for the divergence-free part (in the sense of the Helmholtz decomposition) of $(\mathbf{u} \cdot \nabla)\,\mathbf{u}$. We remark that any standing steady solution of (14), will get an additional divergence-free part, when its reference system is transformed by a Galilei transformation $\mathbf{u} \to \mathbf{u} + \mathbf{w}$, where $\mathbf{w}$ denotes a constant vector field [5].

3 Numerical Example—The Chorin Vortex

This section illustrates numerically some of the theoretical considerations above. Consider the Chorin vortex

$$\begin{aligned} \mathbf{u}(x, y, t) &= (-\cos(n\pi x)\sin(n\pi y), \sin(n\pi x)\cos(n\pi y))^T\, e^{-2n^2\pi^2\nu^t} \\ p(x, y, t) &= -\frac{1}{4}\,(\cos(2n\pi x) + \cos(2n\pi y))\, e^{-4n^2\pi^2\nu^t}. \end{aligned}$$

for $n = 2$ on the unit square.

First, it is studied for the time-dependent Stokes equations $\mathbf{u}_t - \nu\Delta\mathbf{u} + \nabla p = \mathbf{f}$, $\nabla \cdot \mathbf{u} = 0$, and later for the time-dependent Navier–Stokes equations with $\mathbf{f} = \mathbf{0}$, i.e., the Chorin vortex is a so-called *exact solution* of the incompressible Navier–Stokes equations. In the Stokes case, the right hand side $\mathbf{f}$ contains a dominant gradient field, while in the Navier–Stokes case the nonlinear term $(\mathbf{u} \cdot \nabla)\mathbf{u}$ itself is a dominant gradient field. Thus, the Chorin vortex problem in the high Reynolds number Navier–

Stokes setting is an example of (strong) pseudo-dominant convection in the sense above.

In the following numerical examples, for the space discretization we employ finite element spaces $\mathbf{X}_h \subset \mathbf{H}_0^1$ for the discrete velocities, $Q_h \subset L_0^2$ for the discrete pressures. Additional, we will employ a $\mathbf{H}(\text{div})$ velocity reconstruction operator $\mathbf{I}_h : \mathbf{X}_h \to \mathbf{H}(\text{div})$ as mentioned in (10) and (11). Then, the above space discretization is written as follows: search for $(\mathbf{u}_h, p_h) \in \mathbf{X}_h \times Q_h$ such that it holds for all $(\mathbf{v}_h, q_h) \in \mathbf{X}_h \times Q_h$

$$\begin{aligned}(\mathbf{I}_h \dot{\mathbf{u}}_h, \mathbf{I}_h \mathbf{v}_h) + \nu(\nabla \mathbf{u}_h, \nabla \mathbf{v}_h) + ((\mathbf{u}_h \cdot \nabla)\mathbf{u}_h, \mathbf{I}_h \mathbf{v}_h) - (p_h, \nabla \cdot \mathbf{v}_h) &= (\mathbf{f}, \mathbf{I}_h \mathbf{v}_h) \\ (\nabla \cdot \mathbf{v}_h, q_h) &= 0.\end{aligned} \tag{15}$$

Using $\mathbf{I}_h = id$ denotes a Galerkin scheme, while for $\mathbf{I}_h \neq id$ the scheme leaves the classical framework of Galerkin schemes.

In the numerical examples, we compare the classical (Galerkin) Bernardi–Raugel element, a-pressure-robust variant of it [10, 12] and the classical (Galerkin) cubic Taylor–Hood element P_3-P_2.

Thus, for the classical cubic Taylor–Hood element, we specify $\mathbf{X}_h = (P_3)^2$, $Q_h = P_2$ and $\mathbf{I}_h = id$.

For the Bernardi–Raugel element, we specify $Q_h = P_0$ and $\mathbf{X}_h = (P_1)^2 \oplus$ {normal-weighted face bubbles}, where the ($\mathbf{H}^1$-conforming) face bubbles are elementwise given by $\{\lambda_1\lambda_2\mathbf{n}_3 \oplus \lambda_2\lambda_3\mathbf{n}_1 \oplus \lambda_3\lambda_1\mathbf{n}_2\}$ making the discretization discretely inf-sup stable. For the classical (Galerkin) Bernardi–Raugel element, we choose $\mathbf{I}_h = id$. For the modified, pressure-robust Bernardi–Raugel element, the only change in the scheme is that we employ $\mathbf{I}_h = \mathbf{I}_h^{\text{BDM}_1}$, mapping $\mathbf{X}_h$ elementwise via the BDM_1 standard interpolation to $\mathbf{H}(\text{div})$.

In the following, the errors $\mathbf{e} := \mathbf{u} - \mathbf{u}_h$ are measured in the L^2-norm at final time $T = 0.01$ and the accumulated H^1-seminorm

$$\|\nabla \mathbf{e}\|^2_{L^2(\Omega\times[0,T])} := \tau \sum_{n=1}^{N} \frac{1}{2} \left(\|\nabla \mathbf{e}(t_{n-1})\|^2_{L^2(\Omega)} + \|\nabla \mathbf{e}(t_n)\|^2_{L^2(\Omega)} \right).$$

The time integration is performed by a Crank-Nicolson scheme with time step $\tau = 10^{-4}$ until $T = 0.01$.

In Table 1, the classical (Galerkin) Bernardi–Raugel scheme is considered for the Stokes problem. For $\nu = 1$, one finds the optimal rate 2 in the $\mathbf{L}^2(T)$ norm, but for $\nu = 10^{-6}$ the convergence rate breaks down preasymptotically to 0!

Similarly, in Table 3, the classical (Galerkin) cubic Taylor–Hood scheme is considered for the Stokes problem. For $\nu = 1$, one finds the optimal rate 4 in the $\mathbf{L}^2(T)$ norm, but for $\nu = 10^{-6}$ the convergence rate breaks down preasymptotically to 2!

Table 1 Errors and convergence rates for the classical Bernardi–Raugel for the Stokes case

(classical Bernardi–Raugel $\nu = 1$)					(classical Bernardi–Raugel $\nu = 10^{-6}$)				
ndof	$\|\mathbf{e}(T)\|_{L^2(\Omega)}$	rate	$\|\nabla\mathbf{e}\|_{L^2(\Omega\times[0,T])}$	rate	ndof	$\|\mathbf{e}(T)\|_{L^2(\Omega)}$	rate	$\|\nabla\mathbf{e}\|_{L^2(\Omega\times[0,T])}$	rate
478	$1.056 \cdot 10^{-2}$		$1.516 \cdot 10^{-1}$		478	$2.290 \cdot 10^{-2}$		$2.191 \cdot 10^{-1}$	
1822	$2.737 \cdot 10^{-3}$	1.95	$7.662 \cdot 10^{-2}$	0.98	1822	$1.710 \cdot 10^{-2}$	0.42	$1.389 \cdot 10^{-1}$	0.66
7114	$6.856 \cdot 10^{-4}$	2.00	$3.841 \cdot 10^{-2}$	1.00	7114	$1.755 \cdot 10^{-2}$	−0.04	$1.837 \cdot 10^{-1}$	−0.40
28114	$1.711 \cdot 10^{-4}$	2.00	$1.922 \cdot 10^{-2}$	1.00	28114	$1.793 \cdot 10^{-2}$	−0.03	$3.535 \cdot 10^{-1}$	−0.94
111778	$4.284 \cdot 10^{-5}$	2.00	$9.619 \cdot 10^{-3}$	1.00	111778	$1.806 \cdot 10^{-2}$	−0.01	$7.060 \cdot 10^{-1}$	−1.00

Table 2 Errors and convergence rates for the pressure-robust modified Bernardi–Raugel for the Stokes case

(modified Bernardi–Raugel $\nu = 1$)					(modified Bernardi–Raugel $\nu = 10^{-6}$)				
ndof	$\|\mathbf{e}(T)\|_{L^2(\Omega)}$	rate	$\|\nabla\mathbf{e}\|_{L^2(\Omega\times[0,T])}$	rate	ndof	$\|\mathbf{e}(T)\|_{L^2(\Omega)}$	rate	$\|\nabla\mathbf{e}\|_{L^2(\Omega\times[0,T])}$	rate
478	$1.187 \cdot 10^{-2}$		$1.598 \cdot 10^{-1}$		478	$1.997 \cdot 10^{-2}$		$2.168 \cdot 10^{-1}$	
1822	$3.123 \cdot 10^{-3}$	1.93	$8.240 \cdot 10^{-2}$	0.96	1822	$4.936 \cdot 10^{-3}$	2.02	$1.070 \cdot 10^{-1}$	1.02
7114	$7.801 \cdot 10^{-4}$	2.00	$4.184 \cdot 10^{-2}$	0.98	7114	$1.210 \cdot 10^{-3}$	2.03	$5.330 \cdot 10^{-2}$	1.01
28114	$1.834 \cdot 10^{-4}$	2.09	$2.152 \cdot 10^{-2}$	0.96	28114	$2.939 \cdot 10^{-4}$	2.04	$2.670 \cdot 10^{-2}$	1.00
111778	$4.877 \cdot 10^{-5}$	1.91	$1.094 \cdot 10^{-2}$	0.98	111778	$8.698 \cdot 10^{-5}$	1.76	$1.401 \cdot 10^{-2}$	0.93

Actually, these results for the Galerkin Bernardi–Raugel and the Galerkin cubic Taylor–Hood element are confirmed and proved in [13]. There, it is shown that preasymptotically for $\nu \ll 1$, the order in the $\mathbf{L}^2$ norm equals the formal order of the discrete pressure space, which is 0 for Bernardi–Raugel and 2 for the cubic Taylor–Hood element.

However, one sees in Table 2 that for the (pressure-robust) modified Bernardi–Raugel this breakdown of the convergence for $\nu \ll 1$ does not happen, and the modified pressure-robust Bernardi–Raugel element converges for $\nu \ll 1$ with order 2 in the $\mathbf{L}^2$ norm (Table 3). Thus, a pressure-robust first order element has the same convergence order as a cubic, classical Taylor–Hood element, preasymptotically!

For the full Navier–Stokes Chorin vortex, the results are essentially very similar, see also [13]. This demonstrates that dominant gradient fields in the transient Navier-Stokes problem at high Reynolds number can lead to a severe degradation of the convergence order of space discretizations that are not pressure-robust (Tables 4 and 5). In the transient Navier–Stokes Chorin vortex this gradient field is given by $(\mathbf{u} \cdot \nabla)\mathbf{u}$. Thus, low-order pressure-robust discretizations can be competitive against higher-order standard solvers on preasymptotic grids [13].

Table 3 Errors and convergence rates for the classical cubic Taylor–Hood method for the Stokes case

(classical cubic Taylor–Hood $\nu = 1$)					(classical cubic Taylor–Hood $\nu = 10^{-6}$)				
ndof	$\|\mathbf{e}(T)\|_{L^2(\Omega)}$	rate	$\|\nabla\mathbf{e}\|_{L^2(\Omega\times[0,T])}$	rate	ndof	$\|\mathbf{e}(T)\|_{L^2(\Omega)}$	rate	$\|\nabla\mathbf{e}\|_{L^2(\Omega\times[0,T])}$	rate
1479	$6.037 \cdot 10^{-4}$		$4.514 \cdot 10^{-3}$		1479	$2.663 \cdot 10^{-3}$		$1.319 \cdot 10^{-2}$	
5683	$7.001 \cdot 10^{-5}$	3.11	$9.535 \cdot 10^{-4}$	2.24	5683	$6.900 \cdot 10^{-4}$	1.95	$5.690 \cdot 10^{-3}$	1.21
22275	$6.078 \cdot 10^{-6}$	3.53	$1.614 \cdot 10^{-4}$	2.56	22275	$1.903 \cdot 10^{-4}$	1.86	$2.801 \cdot 10^{-3}$	1.02
88195	$1.019 \cdot 10^{-6}$	2.58	$2.789 \cdot 10^{-5}$	2.53	88195	$4.613 \cdot 10^{-5}$	2.04	$1.329 \cdot 10^{-3}$	1.07
350979	$8.802 \cdot 10^{-7}$	0.21	$4.902 \cdot 10^{-6}$	2.51	350979	$1.159 \cdot 10^{-5}$	1.99	$6.591 \cdot 10^{-4}$	1.01

Table 4 Errors and convergence rates for the classical and modified Bernardi–Raugel method for the Navier–Stokes case

(classical Bernardi–Raugel $\nu = 10^{-4}$)					(modified Bernardi–Raugel $\nu = 10^{-4}$)				
ndof	$\|\mathbf{e}(T)\|_{L^2(\Omega)}$	rate	$\|\nabla\mathbf{e}\|_{L^2(\Omega\times[0,T])}$	rate	ndof	$\|\mathbf{e}(T)\|_{L^2(\Omega)}$	rate	$\|\nabla\mathbf{e}\|_{L^2(\Omega\times[0,T])}$	rate
478	$2.271 \cdot 10^{-2}$		$2.191 \cdot 10^{-1}$		478	$2.024 \cdot 10^{-2}$		$2.176 \cdot 10^{-1}$	
1822	$1.685 \cdot 10^{-2}$	0.43	$1.378 \cdot 10^{-1}$	0.67	1822	$6.595 \cdot 10^{-3}$	1.62	$1.107 \cdot 10^{-1}$	0.98
7114	$1.713 \cdot 10^{-2}$	−0.02	$1.795 \cdot 10^{-1}$	−0.38	7114	$2.983 \cdot 10^{-3}$	1.14	$6.552 \cdot 10^{-2}$	0.76
28114	$1.646 \cdot 10^{-2}$	0.06	$3.291 \cdot 10^{-1}$	−0.87	28114	$5.878 \cdot 10^{-4}$	2.34	$3.157 \cdot 10^{-2}$	1.05
111778	$1.316 \cdot 10^{-2}$	0.32	$5.473 \cdot 10^{-1}$	−0.73	111778	$1.095 \cdot 10^{-4}$	2.42	$1.506 \cdot 10^{-2}$	1.07

Table 5 Errors and convergence rates for the classical cubic Taylor–Hood method for the Navier–Stokes case

(classical Taylor–Hood $\nu = 10^{-4}$)					(classical cubic Taylor–Hood $\nu = 10^{-4}$)				
ndof	$\|\mathbf{e}(T)\|_{L^2(\Omega)}$	rate	$\|\nabla\mathbf{e}\|_{L^2(\Omega\times[0,T])}$	rate	ndof	$\|\mathbf{e}(T)\|_{L^2(\Omega)}$	rate	$\|\nabla\mathbf{e}\|_{L^2(\Omega\times[0,T])}$	rate
631	$1.001 \cdot 10^{-2}$		$3.875 \cdot 10^{-2}$		1479	$2.458 \cdot 10^{-3}$		$1.289 \cdot 10^{-2}$	
2375	$3.135 \cdot 10^{-3}$	1.68	$2.086 \cdot 10^{-2}$	0.89	5683	$6.793 \cdot 10^{-4}$	1.86	$5.628 \cdot 10^{-3}$	1.20
9211	$9.719 \cdot 10^{-4}$	1.69	$1.316 \cdot 10^{-2}$	0.66	22275	$1.732 \cdot 10^{-4}$	1.97	$2.599 \cdot 10^{-3}$	1.11
36275	$2.830 \cdot 10^{-4}$	1.78	$7.930 \cdot 10^{-3}$	0.73	88195	$4.041 \cdot 10^{-5}$	2.10	$1.183 \cdot 10^{-3}$	1.14
143971	$9.642 \cdot 10^{-5}$	1.55	$5.578 \cdot 10^{-3}$	0.51	350979	$7.541 \cdot 10^{-6}$	2.42	$4.569 \cdot 10^{-4}$	1.37

References

1. Ahmed, N., Linke, A., Merdon, C.: Towards pressure-robust mixed methods for the incompressible Navier-Stokes equations. Comput. Methods Appl. Math. **18**(3), 353–372 (2018)
2. Auchmuty, G., Alexander, J.C.: L^2-well-posedness of 3D div-curl boundary value problems. Quart. Appl. Math. **63**(3), 479–508 (2005)
3. Chorin, A.J., Marsden, J.E.: A mathematical Introduction to Fluid Mechanics. *Texts in Applied Mathematics*, vol. 4, 3d edn. Springer, New York (1993)

4. Crouzeix, M., Raviart, P.A.: Conforming and nonconforming finite element methods for solving the stationary Stokes equations. I. Rev. Française Automat. Informat. Recherche Opérationnelle Sér. Rouge **7**(R-3), 33–75 (1973)
5. Gauger, N.R., Linke, A., Schroeder, P.W.: On high-order pressure-robust space discretisations, their advantages for incompressible high Reynolds number generalised Beltrami flows and beyond. SMAI J. Comput. Math. **5**, 89–129 (2019)
6. Girault, V., Raviart, P.A.: Finite Element Methods for Navier-Stokes Equations: Theory and Algorithms. vol. 5 of Springer Series in Computational Mathematics. Springer, Berlin (1980)
7. John, V.: Finite element methods for incompressible flow problems, Springer Series in Computational Mathematics, vol. 51. Springer, Cham (2016)
8. John, V., Linke, A., Merdon, C., Neilan, M., Rebholz, L.G.: On the divergence constraint in mixed finite element methods for incompressible flows. SIAM Rev. **59**(3), 492–544 (2017)
9. Linke, A.: On the role of the Helmholtz decomposition in mixed methods for incompressible flows and a new variational crime. Comput. Methods Appl. Mech. Eng. **268**, 782–800 (2014)
10. Linke, A., Matthies, G., Tobiska, L.: Robust arbitrary order mixed finite element methods for the incompressible Stokes equations with pressure independent velocity errors. ESAIM Math. Model. Numer. Anal. **50**(1), 289–309 (2016)
11. Linke, A., Merdon, C.: On velocity errors due to irrotational forces in the Navier-Stokes momentum balance. J. Comput. Phys. **313**, 654–661 (2016)
12. Linke, A., Merdon, C.: Pressure-robustness and discrete Helmholtz projectors in mixed finite element methods for the incompressible Navier-Stokes equations. Comput. Methods Appl. Mech. Eng. **311**, 304–326 (2016)
13. Linke, A., Rebholz, L.G.: Pressure-induced locking in mixed methods for time-dependent (Navier-)Stokes equations. J. Comput. Phys. **388**, 350–356 (2019)
14. Rivière, B.: Discontinuous Galerkin Methods for Solving Elliptic and Parabolic Equations, Frontiers in Applied Mathematics, vol. 35. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA (2008)

Well-Balanced Discretisation for the Compressible Stokes Problem by Gradient-Robustness

Alexander Linke and Christian Merdon

Abstract Based on the novel concept of gradient-robustness a well-balanced and provably convergent scheme for the compressible Stokes equations is discussed. Gradient-robustness means that arbitrary gradient fields in the momentum balance are correctly balanced by the discrete pressure gradient if there is enough mass in the system to compensate the force. For low Mach numbers the scheme degenerates to a recent inf-sup stable and pressure-robust discretisation for the incompressible Stokes equations. Numerical examples illustrate the properties for nearly-hydrostatic low Mach number flows also for nonlinear equations of state.

Keywords Compressible Stokes equations · Mixed finite element methods · Gradient-robustness · Well-balanced schemes

MSC (2010) 76D07 · 65N30 · 65N12

1 Introduction

This paper studies a novel well-balanced discretisation scheme for the barotropic compressible (nonlinear) Stokes problem based on gradient-robustness. The scheme is similar to the Crouzeix-Raviart finite element-finite volume scheme of [5] where the continuity equation is discretised by some upwind finite volume technique to ensure non-negativity and the mass constraint of the piecewise-constant discrete density ρ_h. The main important difference is a modified discretization of the right-hand side inspired by certain *pressure-robust schemes* for the incompressible Stokes equations, see e.g. [7, 8]. This modification maps discretely divergence-free test

A. Linke · C. Merdon (✉)
Weierstrass Institute for Applied Analysis and Stochastics,
Mohrenstraße 39, 10117 Berlin, Germany
e-mail: christian.merdon@wias-berlin.de

A. Linke
e-mail: alexander.linke@wias-berlin.de

R. Klöfkorn et al. (eds.), *Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples*, Springer Proceedings in Mathematics & Statistics 323,
https://doi.org/10.1007/978-3-030-43651-3_8

functions to pointwise divergence-free ones by some reconstruction operator Π and so improves the balancing of divergence-free and irrotational gradient forces to gain much more accuracy in nearly hydrostatic situations. Moreover, using the conforming Bernardi–Raugel finite element instead of the nonconforming Crouzeix–Raviart finite element for the velocity-pressure pair allows for application of the stress tensor $\boldsymbol{\sigma}$ and an easier convergence proof of the scheme in [1].

1.1 The Steady Compressible Stokes Equations

Given some force fields $(\mathbf{f}, \mathbf{g}) \in \mathbf{L}^2(\Omega) \times \mathbf{L}^\infty(\Omega)$ on some Lipschitz domain $\Omega \subset \mathbb{R}^d$ (where $d \in \{2, 3\}$), total mass $M > 0$ and Lamé parameters $0 < \mu \in \mathbb{R}$ and $-2\mu < \lambda \in \mathbb{R}$, we seek some velocity field $\mathbf{u}$, a pressure field p and a non-negative density $\rho \geq 0$ with $\int_\Omega \rho \, dx = M$ such that

$$\begin{aligned} -\nabla \cdot \boldsymbol{\sigma} + \nabla p &= \mathbf{f} + \rho \mathbf{g}, \\ \operatorname{div}(\rho \mathbf{u}) &= 0 \\ p &= \varphi(\rho) := c\rho^\gamma. \end{aligned} \tag{1}$$

For simplicity, homogeneous Dirichlet velocity boundary conditions are assumed to close the system. The friction term is modeled as in linear elasticity by

$$\boldsymbol{\sigma} = 2\mu \boldsymbol{\varepsilon}(\mathbf{u}) + \lambda(\nabla \cdot \mathbf{u})\boldsymbol{I}, \tag{2}$$

where $\boldsymbol{\varepsilon}(\mathbf{u}) := \frac{1}{2}(\nabla \mathbf{u} + (\nabla \mathbf{u})^T)$, denotes the symmetric part of the gradient, compare e.g. with [3]. The equation of state function realises some power law with exponent $\gamma \geq 1$ and some $c > 0$ related to the speed of sound in the fluid that (in a dimensionless setting) may model the squared inverse of the Mach number.

Its weak form seeks $(\mathbf{u}, p, \rho) \in \mathbf{H}_0^1(\Omega) \times L^2(\Omega) \times L^{2\gamma}(\Omega)$ with

$$\begin{aligned} a_1(\mathbf{u}, \mathbf{v}) + a_2(\mathbf{u}, \mathbf{v}) + b(p, \mathbf{v}) &= F(\mathbf{v}) + G(\rho, \mathbf{v}) \text{ for all } \mathbf{v} \in \mathbf{H}_0^1(\Omega), \\ c(\rho, u, \phi) &= 0 \text{ for all } \phi \in W^{1,\infty}(\Omega), \end{aligned} \tag{3}$$

according to e.g. [2] where

$$\begin{aligned} a_1(\mathbf{u}, \mathbf{v}) &:= 2\mu \int_\Omega \boldsymbol{\varepsilon}(\mathbf{u}) : \boldsymbol{\varepsilon}(\mathbf{v}) \, dx, & a_2(\mathbf{u}, \mathbf{v}) &:= \lambda \int_\Omega \operatorname{div}(\mathbf{u}) \operatorname{div}(\mathbf{v}), \\ b(p, \mathbf{u}) &:= -\int_\Omega p \operatorname{div}(\mathbf{u}) \, dx, & c(\rho, \mathbf{u}, \phi) &:= \int_\Omega \rho \mathbf{u} \cdot \nabla \phi \, dx, \\ F(\mathbf{v}) &:= \int_\Omega \mathbf{f} \cdot \mathbf{v} \, dx, & G(\rho, \mathbf{v}) &:= \int_\Omega \rho \mathbf{g} \cdot \mathbf{v} \, dx. \end{aligned}$$

1.2 Divergence-Free Part and Helmholtz Projector

Consider the Sobolev space $\mathbf{V} := \mathbf{H}_0^1(\Omega)$ that has an orthogonal split into the divergence-free functions and its orthogonal complement

$$\begin{aligned}\mathbf{V}^0 &= \{\mathbf{v} \in \mathbf{V} : \nabla \cdot \mathbf{v} = 0\},\\ \mathbf{V}^\perp &= \{\mathbf{v} \in \mathbf{V} : (\boldsymbol{\varepsilon}(\mathbf{v}), \boldsymbol{\varepsilon}(\mathbf{w})) = 0 \quad \text{for all } \mathbf{w} \in \mathbf{V}^0\}\end{aligned}$$

(where $(\cdot, \cdot)$ denotes the L^2 scalarproduct). Hence, any $\mathbf{u} \in \mathbf{V}$ can be split into

$$\mathbf{u} = \mathbf{u}^0 + \mathbf{u}^\perp \qquad \text{for some } \mathbf{u}^0 \in \mathbf{V}^0 \text{ and } \mathbf{u}^\perp \in \mathbf{V}^\perp.$$

For test functions $\mathbf{v}^0 \in \mathbf{V}^0$ in Eq. (3), one obtains

$$2\mu(\boldsymbol{\varepsilon}(\mathbf{u}^0), \boldsymbol{\varepsilon}(\mathbf{v}^0)) = (\mathbf{f} + \rho\mathbf{g}, \mathbf{v}^0) = (\mathbb{P}(\mathbf{f} + \rho\mathbf{g}), \mathbf{v}^0). \tag{4}$$

Here, $\mathbb{P}$ denotes the Helmholtz–Hodge projector characterized by projecting into the space of divergence-free vector fields

$$\mathbf{L}_\sigma^2 = \{\mathbf{v} \in \mathbf{L}^2(\Omega) : (\mathbf{v}, \nabla\phi) = 0 \ \text{ for all } \phi \in H^1(\Omega)\},$$

i.e. $\mathbb{P} : \mathbf{L}^2(\Omega) \to \mathbf{L}_\sigma^2$ and $(\mathbb{P}(\mathbf{f}), \mathbf{w}) = (\mathbf{f}, \mathbf{w})$ for all $\mathbf{w} \in \mathbf{L}_\sigma^2$, see [7] for details.

Thus, if ρ is fixed or $\mathbf{g} = \mathbf{0}$, the *divergence-free part* $\mathbf{u}^0$ of the solution $\mathbf{u}$ of the compressible problem fulfills the *linear incompressible Stokes equations* (4) and hence balances $\mathbb{P}(\mathbf{f} + \rho\mathbf{g})$. Also note, that in particular it holds

$$\mathbb{P}(\nabla\phi) = \mathbf{0} \qquad \text{for all } \phi \in H^1(\Omega),$$

which means that gradient forces in $\mathbf{f} + \rho\mathbf{g}$ can only be balanced by the pressure p and not by the divergence-free part $\mathbf{u}^0$. Gradient-robust schemes are concerned with the preservation of this correct balancing.

2 Well-Balanced Bernardi–Raugel Finite Element—Finite Volume Method

This section explains our modified Bernardi–Raugel finite element-finite volume scheme to discretise the weak form of the compressible Stokes system (3).

2.1 Notation and Upwind Divergence

Given some regular triangulation $\mathscr{T}$ of the domain with nodes $\mathscr{N}$ and faces $\mathscr{F}$, the velocity space of the Bernardi–Raugel finite element is given by

$$\mathbf{V}_h := (\mathbf{P}_1(\mathscr{T}) \oplus \mathscr{B}(\mathscr{F})) \cap \mathbf{V}$$

where $\mathbf{P}_1(\mathscr{T})$ denotes the set of piecewise affine polynomials and $\mathscr{B}(\mathscr{F})$ denotes the set of normal-weighted face bubbles. The discrete density and pressure fields are discretised by piecewise constants, i.e.

$$Q_h := P_0(\mathscr{T}).$$

Moreover, below Π denotes the lowest-order Brezzi–Douglas–Marini standard interpolation that preserves the discrete divergence of the test function, i.e.

$$b(\Pi \mathbf{v}_h, q_h) = b(\mathbf{v}_h, q_h) \quad \text{for all } \mathbf{v}_h \in \mathbf{V}_h,\ q_h \in Q_h.$$

Given a simplex $T \in \mathscr{T}$, the subset $\mathscr{F}(T)$ of $\mathscr{F}$ denotes the faces along the boundary of T. With this, the upwind discretisation $\mathrm{div}_{\mathrm{upw}}(\rho_h \mathbf{u}_h) \in P_0(\mathscr{T})$ of $\mathrm{div}(\rho_h \mathbf{u}_h)$ is defined on all $T \in \mathscr{T}$ by

$$\mathrm{div}_{\mathrm{upw}}(\rho_h \mathbf{u}_h)|_T := \frac{1}{|T|} \sum_{F \in \mathscr{F}(T)} u_{T,F}^{+} \rho_h|_T - u_{T,F}^{-} \rho_h|_{T_{\mathrm{neighbour}}} = \frac{1}{|T|} \sum_{F \in \mathscr{F}(T)} \rho_F^{\mathrm{upw}} u_{T,F},$$

where $u_{T,F} = \int_F \mathbf{u}_h \cdot \mathbf{n}_T \, ds$ is the integral over the face F in outer normal direction of the simplex T and $u_{K,F}^{+} \geq 0$ and $u_{T,F}^{-} \geq 0$ is the positive and negative part, respectively. Hence, $\rho_F^{\mathrm{upw}} := \rho_h|_T$ if $u_{T,F} > 0$ and $\rho_F^{\mathrm{upw}} := \rho_h|_{T_{\mathrm{neighbour}}}$ else for $F = \partial T \cap \partial T_{\mathrm{neighbour}}$. This gives rise to the (singular) matrix

$$D_{jk} := \mathrm{div}_{\mathrm{upw}}(\chi_j \mathbf{u}_h)|_{T_k} \tag{5}$$

where χ_j is the characteristic function of $T_j \in \mathscr{T}$. In order to determine a suitable but non-unique density that solves

$$\mathrm{div}_{\mathrm{upw}}(\rho_h \mathbf{u}_h) = 0 \quad \Leftrightarrow \quad D\rho_h = 0$$

we suggest some pseudo-time stepping with the implicit Euler method to preserve the non-negativity and mass constraints. Key motivation is that D plus any positive diagonal matrix results in some M-matrix with positive inverse, see [1] for details. Here, M_{Q_h} is the mass matrix matrix of Q_h.

2.2 An Iterative Algorithm

The previous statements motivate the following algorithm that intends to discretise (3) by some iterative pseudo time stepping.

Algorithm 1 *Given initial values $\mathbf{u}_h^0$ and ρ_h^0, some time step $\tau > 0$ and termination tolerance tol > 0, perform the following loop (start with $n = 1$) until convergence:*

1. *Update matrix D according to (5) (with $\mathbf{u}_h = \mathbf{u}_h^{n-1}$) and find $\rho_h^n \in Q_h$ such that*

$$(M_{Q_h} + \tau D)\rho_h^n = M_{Q_h}\rho_h^{n-1}. \tag{6}$$

2. *Update the pressure according to the equation of state, i.e.*

$$p_h^n := \varphi(\rho_h^n). \tag{7}$$

3. *Find $\boldsymbol{u}_h^n \in \boldsymbol{V}_h$ that satisfies the momentum equation for all $\boldsymbol{v}_h \in \boldsymbol{V}_h$, i.e.,*

$$a_1(\boldsymbol{u}_h^n, \boldsymbol{v}_h) + a_2(\Pi\boldsymbol{u}_h^n, \Pi\boldsymbol{v}_h) = F(\Pi\boldsymbol{v}_h) + G(\rho_h^n, \Pi\boldsymbol{v}_h) - b(p_h^n, \boldsymbol{v}_h). \tag{8}$$

4. *Compute residuals of stationary momentum and continuity equations, i.e.*

$$\begin{aligned} res := \|a_1(\boldsymbol{u}_h^n, \bullet) + a_2(\Pi\boldsymbol{u}_h^n, \Pi\bullet) - F(\Pi\bullet) + G(\rho_h^n, \Pi\bullet) - b(p_h^n, \bullet)\|_{l^2} \\ +|\mathrm{div}_{upw}(\rho_h^n\boldsymbol{u}_n)| \end{aligned}$$

5. *Stop if res $<$ tol, otherwise increase n by one and restart loop at 1.*

The triplet $(\mathbf{u}_h, p_h, \rho_h) := (\mathbf{u}_h^n, p_h^n, \rho_h^n) \in \mathbf{V}_h \times Q_h \times Q_h$ denotes a discrete solution of the 'modified' Bernardi–Raugel scheme for the compressible Stokes equations. If Π is replaced by $\Pi = 1$, we call it a solution of the 'classical' Bernardi–Raugel scheme. Both schemes are used in some comparisons of the final section.

Remark 1 Although some fixpoint argument ensures the existence of a solution, see [1] for details, it cannot be guaranteed that the algorithm converges. However, in the presented examples convergence was always observed for small enough τ. Moreover, convergence of the discrete solutions to an exact solution of the compressible Stokes equations is nontrivial and, so far, has only been shown for the isothermal case in [1], or for special schemes on structured meshes like the MAC-scheme [6].

Remark 2 One may use a solution of the incompressible Stokes equations (with $\rho \equiv M/|\Omega|$ in (1)) as an initial guess. See [1] for a more elaborate choice that chooses the density according to the pressure of this incompressible Stokes problem.

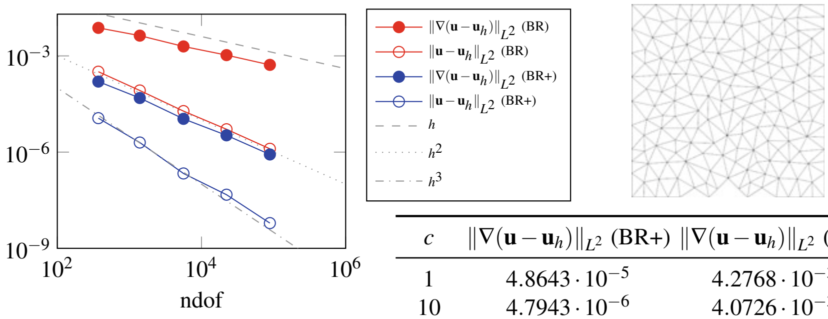
3 Numerical Examples

This section demonstrates the features of the suggested scheme, in particular the well-balanced properties with respect to gradient forces in the momentum balance.

3.1 No-Flow Over Mountains

This example studies the compressible Stokes problem on the domain depicted in Fig. 1 with a non-flat symmetric bottom topography (the large mountain has height 0.1 and width 0.2, the two smaller ones have height 0.05 and width 0.15) and the rest of the boundary. The forces in the right-hand side read $\mathbf{f} = \mathbf{0}$ and $\mathbf{g} = (0, -1)^T$, which leads to the expected no-flow solution $\mathbf{u} = \mathbf{0}$ and some complicated (stratified) density. Note that stratified flows above a non-flat bottom topography are an important research question in oceanography and meteorology.

Figure 1 shows the velocity error of the no-flow velocity for the classical and the modified Bernardi–Raugel method for $\gamma = 1.4$ and $M = 1$. Observe, that the errors are much smaller for the modified well-balanced scheme. For large c the error converges towards zero as expected by gradient-robustness. In fact, for $c \to \infty$, $\rho_h \to$ const and hence also $\rho_h \mathbf{g}$ converges to a gradient of some potential which is not seen by a gradient-robust discretisation. Also, the convergence order with respect to $h \approx \text{ndof}^{-1/2}$ for the modified scheme is one order larger than expected, i.e. cubic convergence for the L^2-error and quadratic convergence for the L^2 gradient error. The classical scheme 'only' converges with the expected rates. This tiny example



c	$\|\nabla(\mathbf{u}-\mathbf{u}_h)\|_{L^2}$ (BR+)	$\|\nabla(\mathbf{u}-\mathbf{u}_h)\|_{L^2}$ (BR)
1	$4.8643 \cdot 10^{-5}$	$4.2768 \cdot 10^{-3}$
10	$4.7943 \cdot 10^{-6}$	$4.0726 \cdot 10^{-3}$
100	$4.7870 \cdot 10^{-7}$	$4.0619 \cdot 10^{-3}$
1000	$4.7863 \cdot 10^{-8}$	$4.0609 \cdot 10^{-3}$
10000	$4.7854 \cdot 10^{-9}$	$4.0608 \cdot 10^{-3}$

Fig. 1 Errors of the modified Bernardi–Raugel method (BR+) and the classical Bernardi–Raugel method (BR) in the no-flow over mountain scenario. Left Plot: convergence history for $c = 1$. Right Plot: Triangulation. Table: Errors on a fixed mesh with ndof = 1394 for different values of c

raises the hope that the novel modified scheme is an interesting discretisation for more complicated low Mach number flows and multi-physics applications.

3.2 Convergence Study and Pressure-Robustness

This example on the unit square examines the exact solution

$$\mathbf{u} := \operatorname{curl}(x^2(x-1)^2y^2(y-1)^2)/\rho, \qquad p = \varphi(\rho) := c\varrho^{\gamma}$$

for $\gamma = 1.4$, $\mu \in \{1, 10^{-3}\}$ and $\lambda = -2\mu/3$. Assuming a quadratic density $\rho := 1 + (y^2 - 1/3)/c$ with $\int_\Omega \rho\, dx = M := 1$, the right-hand side functions are chosen such that $(\mathbf{u}, p, \rho)$ is a solution of the compressible Stokes system with

$$\mathbf{f} := -2\mu\varepsilon(\mathbf{u}) - \frac{\mu}{3}\nabla(\operatorname{div}\mathbf{u}) + \nabla p - \rho(0,-1)^T, \qquad \mathbf{g} := (0,-1)^T.$$

Figure 2 shows convergence histories for $\mu = 1$ and $c = 1$ versus $c = 100$. Both schemes show a very similar behaviour and close to optimal convergence rates, so the modifications do not harm for moderate μ. In Fig. 3 the results for $\mu = 10^{-3}$ show a much different picture. Here, the errors of the modified scheme are much smaller by a factor of about $1/\mu$. This behaviour is related to the lack of pressure-robustness of the classical Bernardi–Raugel method which is healed by the suggested modifications. This example proves that this is also an important aspect in the discretisation of compressible flows.

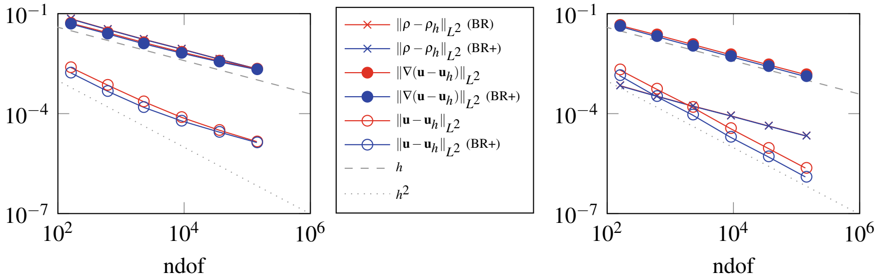


Fig. 2 Convergence histories for the modified method and classical method for $\gamma = 1.4$ and $\mu = 1$ and $c = 1$ (left), $c = 100$ (right)

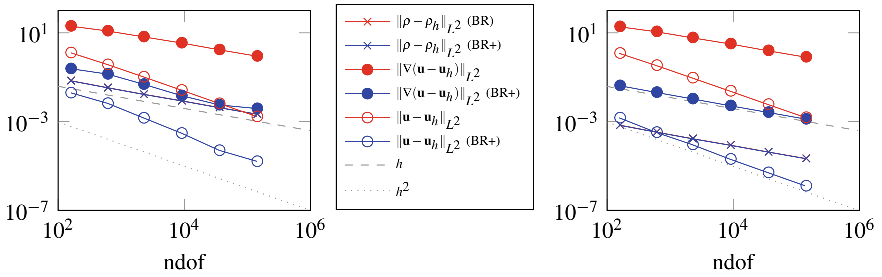


Fig. 3 Convergence histories for the modified method (BR+) and classical method (BR) for $\gamma = 1.4$ and $\mu = 10^{-3}$ and $c = 1$ (left), $c = 100$ (right)

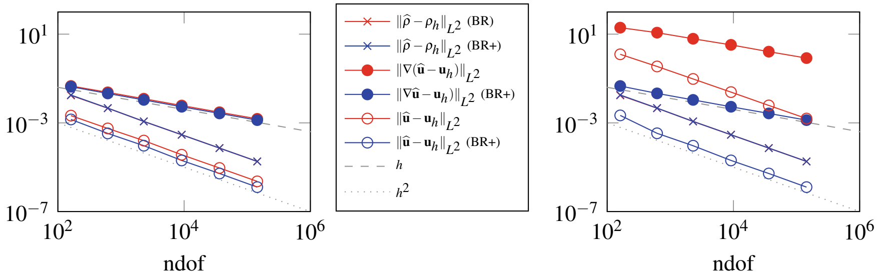


Fig. 4 Convergence histories towards the incompressible Stokes solution for the modified method (BR+) and the classical method (BR) for $\gamma = 1.4$ and $\mu = 1$ (left) and $\mu = 10^{-3}$ (right) and $c \approx 1/h^2$

3.3 Asymptotic Convergence to Stokes System

Inspired by Feireisl et al. [4] the last experiment demonstrates that the scheme is asymptotic preserving in the sense that it converges to the solution of the incompressible Stokes system for $c \to \infty$ and $h \to 0$. To do so, consider the data of the last experiment and c is now coupled to the mesh-width h by $c \approx h^{-2}$ (in fact we start with $c = 16$ on the initial mesh and multiply by 4 after each refinement). For the limit $c = \infty$ it is to be expected that both schemes converge to their parental schemes for the incompressible Stokes problem, which is again locking-free only in case of the 'modified' scheme. Figure 4 plots the errors with respect to the solution of the incompressible Stokes system $\widehat{\mathbf{u}}$ and $\widehat{\rho} \equiv 1$. Another interesting observation is that ρ converges quadratically to $\widehat{\rho}$. This was also observed in [4].

References

1. Akbas, M., Gallouet, T., Gassmann, A., Linke, A., Merdon, C.: A gradient-robust well-balanced scheme for the compressible isothermal stokes problem (2019)
2. Eymard, R., Gallouët, T., Herbin, R., Latché, J.C.: A convergent finite element-finite volume scheme for the compressible Stokes problem II: The isentropic case. Math. Comp. **79**(270), 649–675 (2010)
3. Feireisl, E.: Mathematical models of incompressible fluids as singular limits of complete fluid systems. Milan J. Math. **78**(2), 523–560 (2010)
4. Feireisl, E., Lukáčová-Medvidová, M., Nečasová, v., Novotný, A., She, B.: Asymptotic preserving error estimates for numerical solutions of compressible Navier-Stokes equations in the low Mach number regime. Multiscale Model. Simul. **16**(1), 150–183 (2018). https://doi.org/10.1137/16M1094233
5. Gallouët, T., Herbin, R., Latché, J.C.: A convergent finite element-finite volume scheme for the compressible Stokes problem. Part I: the isothermal case. Math. Comp. **78**(267), 1333–1352 (2009)
6. Gallouët, T., Herbin, R., Latché, J.C., Maltese, D.: Convergence of the MAC scheme for the compressible stationary Navier–Stokes equations. Math. Comp. **87**(311), 1127–1163 (2018)
7. John, V., Linke, A., Merdon, C., Neilan, M., Rebholz, L.G.: On the divergence constraint in mixed finite element methods for incompressible flows. SIAM Rev. **59**(3), 492–544 (2017)
8. Linke, A., Merdon, C.: Pressure-robustness and discrete Helmholtz projectors in mixed finite element methods for the incompressible Navier–Stokes equations. Comput. Methods Appl. Mech. Engrg. **311**, 304–326 (2016)

A Second Order Consistent MAC Scheme for the Shallow Water Equations on Non Uniform Grids

T. Gallouët, R. Herbin, J.-C. Latché, and Y. Nasseri

Abstract We propose in this paper a formally second order scheme for the numerical simulation of the shallow water equations in two space dimensions, based on the so-called Marker-And-Cell (MAC) staggered discretization on non uniform grids. For the space discretization, we use a MUSCL-like scheme for the convection operators while the pressure gadient is centered; time discretization is performed with the Heun scheme. The scheme preserves the positivity of the water height and "lake at rest" steady states. Its consistency in the Lax-Wendroff sense is proven.

Keywords Shallow water · Finite volumes · Heun scheme · Staggered grid

MSC (2010) 65M08 · 76B99

1 Introduction

Let Ω be an open bounded domain of $\mathbb{R}^2$ and let $T > 0$. The shallow water equations with topography over the space and time domain $\Omega \times (0, T)$ read:

$$\partial_t h + \mathrm{div}(h\boldsymbol{u}) = 0 \qquad \text{in } \ \Omega \times (0, T), \tag{1a}$$

T. Gallouët · R. Herbin · Y. Nasseri (✉)
Aix Marseille Université, CNRS, Centrale Marseille, I2M, Marseille, France
e-mail: youssouf.nasseri@univ-amu.fr

T. Gallouët
e-mail: thierry.gallouet@univ-amu.fr

R. Herbin
e-mail: raphaele.herbin@univ-amu.fr

J.-C. Latché
Institut de Sûreté et de Radioprotection Nucléaire (IRSN),
Fontenay-aux-Roses, France
e-mail: jean-claude.latche@irsn.fr

R. Klöfkorn et al. (eds.), *Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples*, Springer Proceedings in Mathematics & Statistics 323,
https://doi.org/10.1007/978-3-030-43651-3_9

$$\partial_t(h\boldsymbol{u}) + \operatorname{div}(h\boldsymbol{u}\otimes\boldsymbol{u}) + \nabla(\frac{1}{2}gh^2) + gh\nabla z = 0 \qquad \text{in } \Omega\times(0,T), \tag{1b}$$

$$\boldsymbol{u}\cdot\boldsymbol{n} = 0 \qquad \text{on } \partial\Omega\times(0,T), \tag{1c}$$

$$h(\boldsymbol{x},0) = h_0(\boldsymbol{x}), \quad \boldsymbol{u}(\boldsymbol{x},0) = \boldsymbol{u}_0(\boldsymbol{x}) \qquad \text{in } \Omega. \tag{1d}$$

where t stands for the time, g is the acceleration of gravity and z the (given) topography, supposed to be continuous with respect to space variables, and independent from the time. For the sake of simplicity, only impermeability conditions are considered; initial conditions h_0 and $\boldsymbol{u}_0$ are such that $h_0 \geq 0$. These equations solve the water height h and the velocity $\boldsymbol{u}$.

We propose in this paper a formally second order extension of a staggered scheme based on the so-called Marker and Cell (MAC) space discretization on non uniform Cartesian grids (often referred to, in the context of shallow-water equations on uniform grids, as the Arakawa-C scheme, after the seminal paper [1]). This scheme may be considered as a higher order extension of an existing first order scheme based on the MAC scheme [3, 6, 7], and staggered schemes on unstructured meshes may be found in [4]. As in [7], numerical fluxes are obtained in a simple way, with an upwinding with respect of the flow speed only, implemented here thanks to a MUSCL-like procedure. Time-stepping is performed with the Heun scheme. This scheme enjoys some stability properties: the water height is shown to be non-negative and the "lake at rest" steady state is preserved. The scheme consistency, in the Lax-Wendroff sense, is shown under rather mild assumptions, namely the boundedness of the approximate solutions and their convergence in $L^1(\Omega\times(0,T))$ thanks to tools developed in [5]; in particular, in contrast with [7], no stability in BV-norms needs to be supposed.

This paper is organized as follows. The scheme is introduced in Sect. 2, and consistency results are given in Sect. 3. Finally, Sect. 4 presents some numerical experiments.

2 The Numerical Scheme

Mesh and notations – Let Ω be a connected subset of $\mathbb{R}^2$ consisting of a union of non uniform rectangles whose edges are assumed to be orthogonal to the canonical basis vectors, denoted by $(\boldsymbol{e}^{(1)}, \boldsymbol{e}^{(2)})$. A discretization $(\mathfrak{M}, \mathcal{E})$ of Ω with a staggered rectangular grid (or MAC grid), involves a primal grid $\mathfrak{M}$ which consists in a conforming structured partition of Ω in rectangles, possibly non uniform. A generic cell of this grid is denoted by K, and its mass center by $\boldsymbol{x}_K$. The scalar unknowns (water height and pressure) are associated to this mesh. The set of all the edges of this mesh is denoted by $\mathcal{E}$, with $\mathcal{E} = \mathcal{E}_{\text{int}} \cup \mathcal{E}_{\text{ext}}$, where $\mathcal{E}_{\text{int}}$ (resp. $\mathcal{E}_{\text{ext}}$) denotes the set of edges that lie in the interior (resp. on the boundary) of the domain. The set of edges (resp. the internal and boundary edges) that are orthogonal to $\boldsymbol{e}^{(i)}$ is denoted by $\mathcal{E}^{(i)}$ (resp. $\mathcal{E}^{(i)}_{\text{int}}$ and $\mathcal{E}^{(i)}_{\text{ext}}$), for $i = 1, 2$. For $\sigma \in \mathcal{E}_{\text{int}}$, we write $\sigma = K|L$ if $\sigma = \partial K \cap \partial L$.

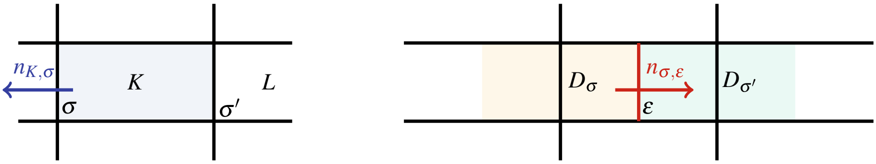


Fig. 1 Notations for control volumes and edges—left: primal mesh, right: dual mesh for the first component of the velocity

A dual cell D_σ associated to an edge $\sigma \in \mathscr{E}$ is defined as follows:

- if $\sigma = K|L \in \mathscr{E}_{\rm int}$ then $D_\sigma = D_{K,\sigma} \cup D_{L,\sigma}$, where $D_{K,\sigma}$ (resp. $D_{L,\sigma}$) is the half-part of K (resp. L) adjacent to σ (see Fig. 1);
- if $\sigma \in \mathscr{E}_{\rm ext}$ is adjacent to the cell K, then $D_\sigma = D_{K,\sigma}$.

For each velocity component i, the domain Ω is thus partitioned into dual cells: $\Omega = \cup_{\sigma \in \mathscr{E}^{(i)}} D_\sigma$. The ith partition is referred to as the ith dual mesh, associated to the ith velocity component, in a sense which is clarified below. The set of the edges of the ith dual mesh is denoted by $\widetilde{\mathscr{E}}^{(i)}$ (note that these edges may be orthogonal to any vector of the basis of $\mathbb{R}^2$ and not only to $\boldsymbol{e}^{(i)}$). The dual edge separating two dual cells D_σ and $D_{\sigma'}$ is denoted by $\epsilon = \sigma|\sigma'$.

The sets of edges of a primal cell K (resp. of a dual cell D_σ) is denoted by $\mathscr{E}(K)$ (resp. $\widetilde{\mathscr{E}}(D_\sigma)$). The vector $\boldsymbol{n}_{K,\sigma}$ stands for the unit normal vector to σ outward K.

The size $\delta_{\mathfrak{M}}$ of the mesh is defined by: $\delta_{\mathfrak{M}} = \max\limits_{K \in \mathfrak{M}} \operatorname{diam}(K)$ and its regularity $\eta_{\mathfrak{M}}$ is given by:

$$\eta_{\mathfrak{M}} = \max\Big\{ \frac{|\sigma|}{|\tau|},\ \sigma \in \mathscr{E}^{(i)},\ \tau \in \mathscr{E}^{(j)},\ i, j = 1, 2,\ i \neq j \Big\}, \tag{2}$$

where $|\cdot|$ stands for the one (or two) dimensional measure of a subset of $\mathbb{R}$ (or $\mathbb{R}^2$).

The discrete unknowns for the ith component of the velocity are associated to the ith dual mesh and are denoted by $(u_{i,\sigma})_{\sigma \in \mathscr{E}^{(i)}}$. The scalar unknowns (discrete water height and topography) are associated to the primal cells and are denoted respectively by $(h_K)_{K \in \mathfrak{M}}$ and $(z_K)_{K \in \mathfrak{M}}$.

Description of the scheme – Let us consider a partition $0 = t_0 < t_1 < \cdots < t_N = T$ of the time interval $(0, T)$, which we suppose uniform, and let $\delta t = t_{n+1} - t_n$ for $n = 0, 1, \ldots, N-1$ be the (constant) time step. The time integration is performed by the second order Heun scheme (which falls in the class of Runge-Kutta schemes), the step n of which may be written as follows:

h^n and $\boldsymbol{u}^n$ being known,

First step – Compute $\tilde{h}^{n+1}$ and $\tilde{u}_i^{n+1}$, $i = 1, 2$, by :

$$\tilde{h}_K^{n+1} = h_K^n - \delta t \operatorname{div}_K(h^n \boldsymbol{u}^n), \quad \forall K \in \mathfrak{M} \tag{3a}$$

$$\tilde{h}_{D_\sigma}^{n+1}\ \tilde{u}_{i,\sigma}^{n+1} = h_{D_\sigma}^n u_{i,\sigma}^n - \delta t\, \mathscr{F}_{D_\sigma}(h^n, u_i^n), \quad \forall \sigma \in \mathscr{E}_{\rm int}^{(i)} \tag{3b}$$

Second step – Compute $\hat{h}^{n+1}$ and $\hat{u}_i^{n+1},\ i=1,2,$ by :

$$\hat{h}_K^{n+1} = \tilde{h}_K^{n+1} - \delta t \operatorname{div}_K(\tilde{h}^{n+1}\tilde{\boldsymbol{u}}^{n+1}), \quad \forall K \in \mathfrak{M} \tag{3c}$$

$$\hat{h}_{D_\sigma}^{n+1}\ \hat{u}_{i,\sigma}^{n+1} = \tilde{h}_{D_\sigma}^{n+1}\tilde{u}_{i,\sigma}^{n+1} - \delta t\, \mathscr{F}_{D_\sigma}(\tilde{h}^{n+1}, \tilde{u}_i^{n+1}), \quad \forall \sigma \in \mathscr{E}_{\text{int}}^{(i)} \tag{3d}$$

Last step – Compute h^{n+1} and $u_i^{n+1}, i=1,2$ by:

$$h_K^{n+1} = \frac{1}{2}\,(h_K^n + \hat{h}_K^{n+1}), \quad \forall K \in \mathfrak{M} \tag{3e}$$

$$h_{D_\sigma}^{n+1}\ u_{i,\sigma}^{n+1} = \frac{1}{2}\big(h_{D_\sigma}^n u_{i,\sigma}^n + \hat{h}_{D_\sigma}^{n+1}\ \hat{u}_{i,\sigma}^{n+1}\big), \quad \forall \sigma \in \mathscr{E}_{\text{int}}^{(i)} \tag{3f}$$

where $\mathscr{F}_{D_\sigma}(h, u_i) = \operatorname{div}_{D_\sigma}(h\boldsymbol{u}\, u_i) + \frac{1}{2}g\,(\eth_i h^2)_\sigma + g h_{\sigma,c}\,(\eth_i z)_\sigma$, and following [6], the water height h_{D_σ} at the face σ, used in (3d) and (3f), is defined by:

$$|D_\sigma|\, h_{D_\sigma} = |D_{K,\sigma}|\, h_K + |D_{L,\sigma}|\, h_L, \text{ for } \sigma = K|L \in \mathscr{E}(K). \tag{4}$$

Note that, to cope with impermeability conditions, the momentum balance equation is not written on the boundary dual cells, and the velocity (in fact, the normal velocity, due to the arrangement of the unknowns) on the boundary edges is just set to zero.

We now define the discrete divergence and gradient operators involved in these relations.

Discrete divergence and gradient operators—The discrete divergence operator div_K on the primal mesh is defined by:

$$|K| \operatorname{div}_K(h\boldsymbol{u}) = \sum_{\sigma\in\mathscr{E}(K)} |\sigma|\, \boldsymbol{F}_\sigma \cdot \boldsymbol{n}_{K,\sigma}, \text{ with } \boldsymbol{F}_\sigma = h_\sigma\, \boldsymbol{u}_\sigma,$$

where $\boldsymbol{u}_\sigma = u_{i,\sigma}\ \boldsymbol{e}^{(i)}$ for $\sigma \in \mathscr{E}^{(i)},\ i=1,2,$ and h_σ is approximated by a second order MUSCL-like interpolation, which may be qualified as "with respect to the particle velocity" in the sense that we apply the technology standard for transport operators (thus disregarding the wave structure of the problem). The ith component of the discrete gradient operator toward an edge $(\eth_i\ \cdot)_\sigma$ applied to the pressure is defined as the transpose of the discrete divergence operator (i.e. the operator obtained by setting $h=1$ in the $\operatorname{div}_K(h\boldsymbol{u})$ operator defined in the previous relation):

$$(\eth_i h^2)_\sigma = (h_K + h_L)\,\frac{|\sigma|}{|D_\sigma|}\,(h_L - h_K)\,\boldsymbol{n}_{K,\sigma}, \text{ for } \sigma = K|L,\ 1 \le i \le d.$$

The same definition holds for the gradient of the topography, and, in this term, the quantity $h_{\sigma,c}$ is defined by

$$h_{\sigma,c} = \frac{1}{2}\,(h_K + h_L). \tag{5}$$

Since no momentum balance equation is written on the boundary edges, a definition of the gradients on these edges is not required.

Discrete divergence, convection operator—The discrete divergence operator on the dual mesh div_{D_σ} is given by:

$$|D_\sigma|\ \mathrm{div}_{D_\sigma}(hu_i\boldsymbol{u}) = \sum_{\epsilon\in\widetilde{\mathscr{E}}^{(i)}(D_\sigma)} |\epsilon|\ \boldsymbol{G}_\epsilon \cdot \boldsymbol{n}_{\sigma,\epsilon},\ \text{with}\ \boldsymbol{G}_\epsilon = \boldsymbol{F}_\epsilon\ u_{i,\epsilon}, \tag{6}$$

where $\boldsymbol{F}_\epsilon$ is the numerical mass flux through ϵ outward D_σ and $u_{i,\epsilon}$ is approximated also by a MUSCL interpolation scheme with respect to $\boldsymbol{F}_\epsilon$. The expression of $\boldsymbol{F}_\epsilon$ depends on the location of the dual edge $\epsilon = \sigma|\sigma'$. Two cases occur:

(i) the normal vector to ϵ is collinear with the normal vector to σ (and σ'), in which case ϵ is included in a primal cell;

(ii) the normal vector to ϵ is perpendicular to the normal vector to σ (and σ'), in which case ϵ results from the union of two half primal edges which we denote by τ and τ'. In both cases, the mass flux $\boldsymbol{F}_\epsilon$ is an average of primal mass fluxes:

$$\boldsymbol{F}_\epsilon = \begin{cases} \dfrac{1}{|\epsilon|}\ (\dfrac{1}{2}|\sigma|\ \boldsymbol{F}_\sigma + \dfrac{1}{2}|\sigma'|\ \boldsymbol{F}_{\sigma'}),\ \text{for case } (i), \\ \dfrac{1}{|\epsilon|}\ (\dfrac{1}{2}|\tau|\ \boldsymbol{F}_\tau + \dfrac{1}{2}|\tau'|\ \boldsymbol{F}_{\tau'}),\ \text{for case } (ii). \end{cases} \tag{7}$$

The definitions (4) and (7) ensure that a discrete mass balance (for the dual water height and the dual mass fluxes) holds on the dual cells, which enables to derive a discrete balance equation for any convex function of the velocity, as well as a discrete kinetic energy balance [7].

3 Stability and Consistency

First of all we verify that, under a CFL restriction, the scheme (3) preserves the positivity of the water height and the "lake at rest" steady state.

Lemma 1 (Preservation of the positivity and the "lake at rest" steady states) *Let* $1 \le n \le N-1$, *let* $(h^n_K,\ \boldsymbol{u}^n_\sigma)_{K\in\mathfrak{M},\ \sigma\in\mathscr{E}}$ *be the approximate solution to the scheme (3), assume that* $h^n_K \ge 0$, *for all* $K \in \mathfrak{M}$ *and assume that the time step satisfies the following condition:*

$$2\delta t \le \min\Bigg[\frac{|K|}{\sum\limits_{\sigma\in\mathscr{E}(K)} |\sigma|\ |\boldsymbol{u}^n_\sigma|}, \frac{|K|}{\sum\limits_{\sigma\in\mathscr{E}(K)} |\sigma|\ |\tilde{\boldsymbol{u}}^n_\sigma|}\Bigg]. \tag{8}$$

Then $h_K^{n+1} \geq 0$ for all $K \in \mathfrak{M}$. Furthermore when $h_K^n + z_K = C$ and $\boldsymbol{u}_\sigma^n = 0$, with C a real number, then $h_K^{n+1} + z_K = C$ and $\boldsymbol{u}_\sigma^{n+1} = 0$.

Before stating the global weak consistency of the scheme, some definitions are needed, as well as some assumed estimates.

Let $(\mathfrak{M}^{(m)}, \mathscr{E}^{(m)})_{m\in\mathbb{N}}$ be a sequence of meshes in the sense introduced in the above section and let $(h^{(m)}, \boldsymbol{u}^{(m)})_{m\in\mathbb{N}}$ be the associated sequence of solutions of the scheme (3).

Assumed estimates—First of all, we assume that $h^{(m)} > 0, \forall m \in \mathbb{N}$; this latter condition can be obtained under the CFL condition (8). Furthermore, we suppose that the water height $h^{(m)}$, its inverse and the velocity are uniformly bounded, i.e.that there exists a $C > 0$ such that for $m \in \mathbb{N}$ and $0 \leq n < N^{(m)}$:

$$1/C < (h^{(m)})_K^n \leq C, \ \forall K \in \mathfrak{M}^{(m)}, \tag{9a}$$

$$|(\boldsymbol{u}^{(m)})_\sigma^n| \leq C, \quad \forall \sigma \in \mathscr{E}^{(m)}. \tag{9b}$$

Finally, we need the following CFL-like assumption on the mesh:

$$\exists\, C \in \mathbb{R}_+ \text{ such that } \forall m \in \mathbb{N}, \delta t^{(m)} \leq C\ \delta_{\mathfrak{M}^{(m)}}. \tag{10}$$

Theorem 1 (Weak consistency of the scheme) *Let $(\mathfrak{M}^{(m)}, \mathscr{E}^{(m)})_{m\in\mathbb{N}}$ be a sequence of meshes such that $\delta_{\mathfrak{M}^{(m)}} \to 0$ as $m \to +\infty$; assume that there exists $\eta > 0$ such that $\eta_{\mathfrak{M}^{(m)}} \leq \eta$ for any $m \in \mathbb{N}$ (with $\eta_{\mathfrak{M}^{(m)}}$ defined by (2)). Let $(h^{(m)}, \boldsymbol{u}^{(m)})_{m\in\mathbb{N}}$ be the associated sequence of solutions to the scheme (3), and assume that (9a), (9b) and (10) hold. Finally, let us suppose that this sequence converges in $L^1(\Omega \times (0,T))^{1+d}$, to $(\bar{h}, \bar{\boldsymbol{u}})$. Then $(\bar{h}, \bar{\boldsymbol{u}})$ satisfies a weak formulation of the shallow water equations.*

Proof (*Sketch of the proof*) The proof uses the following arguments.

(i) Thanks to the assumption (10), we prove that if the discrete solution $(h^{(m)}, \ \boldsymbol{u}^{(m)})$ satisfies (9a–9b) then $(\tilde{h}^{(m)}, \ \tilde{\boldsymbol{u}}^{(m)})$ satisfies the same estimates.

(ii) Then, from the equations (3a) and (3b), we deduce that, if $(h^{(m)}, \boldsymbol{u}^{(m)})$ converges in $L^1(\Omega \times (0,T))^{1+d}$ to $(\bar{h}, \bar{\boldsymbol{u}})$, then $(\tilde{h}^{(m)}, \tilde{\boldsymbol{u}}^{(m)})$ also converges in $L^1(\Omega \times [0,T))^{1+d}$ to the same limit.

(iii) Finally we recast the Heun scheme under the form:

$$h_K^{n+1} = h_K^n - \delta t \left[\tfrac{1}{2}\mathrm{div}_K(h^n \boldsymbol{u}^n) + \tfrac{1}{2}\mathrm{div}_K(\tilde{h}^{n+1}\tilde{\boldsymbol{u}}^{n+1})\right], \qquad \forall K \in \mathfrak{M},$$

$$h_{D_\sigma}^{n+1}\ u_{i,\sigma}^{n+1} = h_{D_\sigma}^n u_{i,\sigma}^n - \delta t \left[\tfrac{1}{2}\mathscr{F}_{D_\sigma}(h^n, u_i^n) + \tfrac{1}{2}\mathscr{F}_{D_\sigma}(\tilde{h}^{n+1}, \tilde{u}_i^{n+1})\right],$$
$$\forall \sigma \in \mathscr{E}_{\mathrm{int}}^{(i)},\ 1 \leq i \leq d.$$

From Points (i) and (ii), it is thus sufficient to check the consistency, in the Lax-Wendroff sense, of the numerical fluxes of the Euler scheme. This is done using the tools of [5], with some technicalities to cope with the non-linearity of the problem and the staggered discretization.

4 Numerical Tests

The scheme under consideration has been developed in the CALIF^{3}S open-source software [2] of the French Institut de Sûreté et de Radioprotection Nucléaire (IRSN) and tested against various benchmarks of the literature. We begin here by checking the accuracy of the scheme on a known regular solution consisting in a travelling vortex. This solution is obtained through the following steps: we first derive a compact-support H^2 solution consisting in a standing vortex which becomes time-dependent by adding a constant velocity motion. The velocity field of the standing vortex and the pressure are sought under the form: $\hat{\boldsymbol{u}} = f(\xi) \begin{bmatrix} -x_2 \\ x_1 \end{bmatrix}$, $\hat{p} = \wp(\xi)$, with $\xi = x_1^2 + x_2^2$. A simple derivation of these expressions yields: $\hat{\boldsymbol{u}} \cdot \nabla \hat{\boldsymbol{u}} = -f(\xi)^2 \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ and $\nabla \hat{p} = 2\, \wp'(\xi) \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$. Using the relation $p = \frac{1}{2}\, g h^2$, we thus obtain a stationary solution of the shallow water Eq. (1) if $\wp$ satisfies $8\, g\, \wp = (F + c)^2$, where F is such that $F' = f^2$, $F(0) = 0$ and c is a positive real number. For the present numerical study, we choose $f(\xi) = 10\, \xi^2 (1 - \xi)^2$ if $\xi \in (0, 1)$, $f = 0$ otherwise, which indeed yields an $H^2(\mathbb{R}^2)$ velocity field (note that as a consequence, the pressure and the water height are also regular), and $c = 1$. The problem is made unsteady by a time translation: given a constant vector field $\boldsymbol{a}$, the pressure p and the velocity $\boldsymbol{u}$ are deduced from the steady state solution $\hat{p}$ and $\hat{\boldsymbol{u}}$, we set $h(\boldsymbol{x}, t) = \hat{h}(\boldsymbol{x} - \boldsymbol{a}t)$ and $\boldsymbol{u}(\boldsymbol{x}, t) = \hat{\boldsymbol{u}}(\boldsymbol{x} - \boldsymbol{a}t) + \boldsymbol{a}$. The center of the vortex is initially located at $\boldsymbol{x}_0 = (0, 0)^t$, the translation velocity $\boldsymbol{a}$ is set to $\boldsymbol{a} = (1, 1)^t$, the computational domain is $\Omega = (-1.2,\ 2.)^2$ and the computation is run on the time interval $(0, 0.8)$.

Computations are performed with successively refined meshes with square cells, and the time step is $\delta t = \delta_{\mathfrak{M}}/8$, and corresponds to a Courant number CFL close to $1/3$, computed as $\mathrm{CFL} = \max(|\boldsymbol{u}| + \sqrt{gh}) \frac{\delta t}{\delta_{\mathfrak{M}}}$. The discrete L^1-norm of the difference between the computed and the exact solutions is given in Table 1.

The observed order of convergence over the whole sequence is 2 for the water height and 1.5 for the velocity.

Table 1 Measured numerical errors for the travelling vortex, $\mathrm{error}(h) = ||h(\cdot, 0.8) - \bar{h}(\cdot, 0.8)||_{L^1_{\mathfrak{M}}(\Omega)}$, $\mathrm{error}(u) = ||\boldsymbol{u}(\cdot, 0.8) - \bar{\boldsymbol{u}}(\cdot, 0.8)||_{L^1_{\mathfrak{M}}(\Omega)}$ and corresponding order of convergence

Mesh	Error (h)	Ord (h)	Error (u)	Ord (u)
32×32	$3.61\, 10^{-3}$	/	$2.93\, 10^{-1}$	/
64×64	$1.15\, 10^{-3}$	1.57	$1.14\, 10^{-1}$	1.28
128×128	$2.58\, 10^{-4}$	2.23	$4.06\, 10^{-2}$	1.40
256×256	$5.85\, 10^{-5}$	2.20	$1.49\, 10^{-2}$	1.36
512×512	$1.53\, 10^{-5}$	1.91	$4.66\, 10^{-3}$	1.60

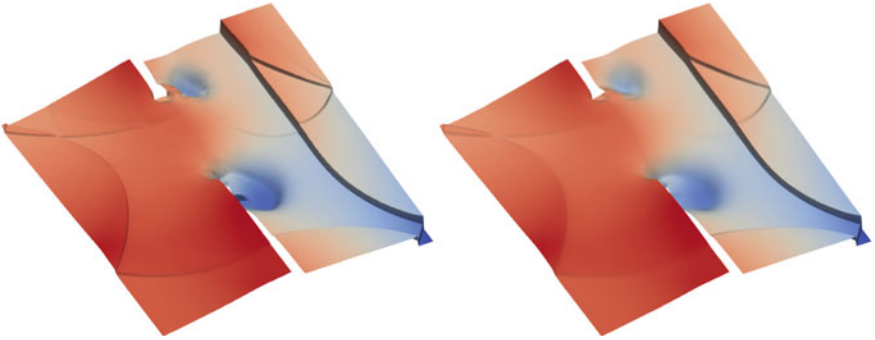


Fig. 2 Partial dam-break flow. Left: MUSCL scheme—Right: upwind scheme

We now turn to a test consisting in a partial dam-break problem with reflection phenomena, and with a non-flat topography. In this test, the computational domain is $\Omega = (0, 200) \times (0, 200) \setminus \Omega_w$ with $\Omega_w = (95, 105) \times (0, 95) \cup (95, 105) \times (170, 200)$. The fluid is supposed to be initially at rest, the initial water height is $h = 10$ for $x_1 \leq 100$ and $h = 5 - 0.04\,(x_1 - 100)$ otherwise, and the topography is $z = 0$ if $x_1 \leq 100$ and $z = 0.04\,(x_1 - 100)$ otherwise. The computation is performed with a mesh obtained from a 1000×1000 regular grid by removing the cells included in Ω_w. The time step is $\delta t = \delta_{\mathfrak{M}}/40$ (the maximal speed of sound and the maximal velocity are both close to 10). A slight stabilization (corresponding to a diffusion coefficient equal to the mesh step divided by 4, so two orders of magnitude lower than the artificial viscosity generated by the upwind scheme in high velocity zones) is added to damp oscillations appearing in the zones at rest, where no numerical diffusion is generated by our schemes. Results obtained at $t = 20$ with the first order in time and space and the present scheme are compared on Fig. 2. One can observe that the second-order scheme is clearly less diffusive. In addition, these results illustrate the capacity of the staggered scheme to deal with reflection conditions by simply imposing the normal velocity to the boundary at zero.

References

1. Arakawa, A., Lamb, V.: A potential enstrophy and energy conserving scheme for the shallow water equations. Mon. Weather Rev. **109**, 18–36 (1981)
2. CALIF^{3}S: A software components library for the computation of fluid flows. https://gforge.irsn.fr/gf/project/califs
3. Doyen, D., Gunawan, H.: An explicit staggered finite volume scheme for the shallow water equations. In: Finite volumes for complex applications, VII. Methods and theoretical aspects, Springer Proceedings in Mathematics and Statistics, vol. 77, pp. 227–235. Springer, Cham (2014)
4. Duran, A., Vila, J.P., Baraille, R.: Energy-stable staggered schemes for the shallow water equations. https://hal.archives-ouvertes.fr (2019)

5. Gallouët, T., Herbin, R., Latché: On the weak consistency of finite volumes schemes for conservation laws on general meshes. SeMA J. (2019) (online)
6. Herbin, R., Latché, J.C., Nasseri, Y., Therme, N.: A decoupled staggered scheme for the shallow water equations. In: Accepted for publication, Proceedings of the 15 International Conference Zaragoza-Pau on Mathematics and its Applications, Jaca, 10–12 Sept 2018. https://arxiv.org/abs/1906.11001 (2019)
7. Herbin, R., Latché, J.C., Nguyen, T.: Consistent segregated staggered schemes with explicit steps for the isentropic and full Euler equations. Math. Model. Numer. Anal. **52**, 893–944 (2018)

Post-processing of Fluxes for Finite Volume Methods for Elliptic Problems

Hanz Martin Cheng

Abstract We develop post-processing techniques for fluxes obtained from finite volume methods, which enables us to reconstruct flow densities in an H-div space. These post-processing techniques ensure that mass is conserved locally, which is very important, e.g., in geophysical applications.

Keywords Diffusion problem · Finite volume methods · Fluxes · Polyhedral mesh

1 Introduction

Let Ω be an open connected subset of $\mathbb{R}^d$ and consider an anisotropic diffusion problem: Find $p \in H^1(\Omega)$ such that

$$-\mathrm{div}(\Lambda \nabla p) = f \text{ on } \Omega, \tag{1}$$

with suitable boundary conditions. Here, we assume that the source term $f \in L^2(\Omega)$, and the diffusion tensor Λ is a measurable function from Ω to the set of $d \times d$ symmetric matrices and there exists $\underline{\lambda}, \overline{\lambda} > 0$ such that, for a.e. $\boldsymbol{x} \in \Omega$, the eigenvalues of $\Lambda(\boldsymbol{x})$ are in $[\underline{\lambda}, \overline{\lambda}]$.

To name a few applications, Eq. (1) is encountered in heat conduction (Fourier's law), porous media flow (Darcy's law), or electric conduction (Ohm's law). Aside from p, another main quantity of interest is the *flow density* $\mathbf{u} := -\Lambda \nabla p$. In the context of porous media flow, (1) is referred to as the pressure equation, where Λ is the permeability tensor, p is the pressure, and $\mathbf{u}$ is the Darcy velocity. The aim of this paper is to post-process the fluxes obtained from finite volume methods for (1),

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

School of Mathematics, Monash University, 3800 Clayton, VIC, Australia

R. Klöfkorn et al. (eds.), *Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples*, Springer Proceedings in Mathematics & Statistics 323,
https://doi.org/10.1007/978-3-030-43651-3_10

so that we can reconstruct a flow density $\mathbf{u} \in H(\mathrm{div}, \Omega)$. The reason behind taking $\mathbf{u} \in H(\mathrm{div}, \Omega)$ is to ensure that mass is conserved locally, which is very important, e.g., in geophysical applications.

2 Finite Volume Methods for the Diffusion Equation

For our discretisation, "mesh" is to be understood in the simplest intuitive way: a partition of Ω into polygonal (in 2D) or polyhedral (in 3D) sets. Following the notations and definitions in [6, Definition 7.2], we denote $\mathcal{T} = (\mathcal{M}, \mathcal{E})$ to be the set of cells K and faces (edges in 2D) σ of our mesh, respectively. For a cell $K \in \mathcal{M}$, we denote by $\mathcal{E}_K \subset \mathcal{E}$ the set of faces (edges) of the cell K. Finite volume schemes are formulated by taking the integral of (1) over a control volume K, and using Stokes' formula to obtain the balance of fluxes: $\sum_{\sigma \in \mathcal{E}_K} \int_\sigma (-\Lambda \nabla p) \cdot \mathbf{n}_{K,\sigma} = \int_K f$. Now, if σ is a face shared by two distinct cells K and L, then $\int_\sigma (-\Lambda \nabla p) \cdot \mathbf{n}_{K,\sigma} + \int_\sigma (-\Lambda \nabla p) \cdot \mathbf{n}_{L,\sigma} = 0$. This is known as the conservation of fluxes. Key to the definition of a finite volume scheme is the choice for the discrete approximation of the fluxes, $F_{K,\sigma} \approx \int_\sigma (-\Lambda \nabla p) \cdot \mathbf{n}_{K,\sigma}$, that satisfy a discrete version of the balance and conservation of fluxes. That is, for each $K \in \mathcal{M}$,

$$\sum_{\sigma \in \mathcal{E}_K} F_{K,\sigma} = \int_K f, \tag{2a}$$

and for each face σ shared by distinct cells K and L,

$$F_{K,\sigma} + F_{L,\sigma} = 0. \tag{2b}$$

A review of some finite volume schemes (e.g. the TPFA, MPFA, and HMM method) that yield fluxes for which the post-processing technique described in Sect. 3 is applicable can be found in [7].

3 Post-processing of the Fluxes for Reconstructing a Flow Density

In this section, we aim to reconstruct, from the approximate fluxes $(F_{K,\sigma})_{K \in \mathcal{M},\, \sigma \in \mathcal{E}_K}$ obtained in (2), a flow density $\mathbf{u}$, taking into account two important features: the preservation of the divergence in (1), to avoid creating regions with artificial wells or sinks (leading to non-physical flows), and the continuity of the normal component of $\mathbf{u}$ (in the context of porous media flow, this ensures that the velocity is well defined, and does not freeze along an interface). A natural space for functions that satisfies these properties is the $H(\mathrm{div}, \Omega)$ space, and the most common examples of functions

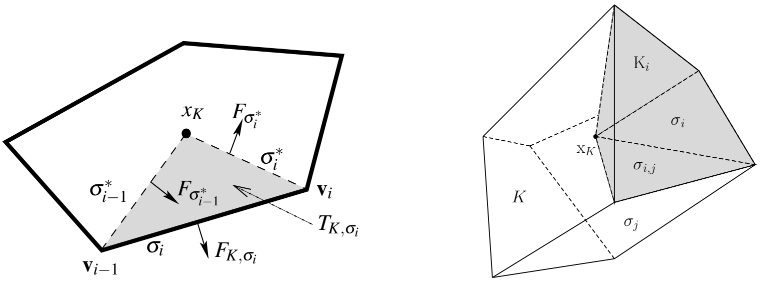


Fig. 1 Left: triangulation in 2D. Right: division of a polyhedron into sub-cells

that live in $H(\mathrm{div}, \Omega)$ are the $\mathbb{RT}_k$ finite elements on simplices [4]. In particular, we focus on $\mathbb{RT}_0$ elements on simplices. To do so, we divide each cell $K \in \mathscr{M}$ into simplices (triangles for $d = 2$, tetrahedra for $d = 3$), gathered in a set $\mathscr{S}_K$, that share $\boldsymbol{x}_K$ as apex and whose bases are faces or subsets of the faces of K. In [10], hexahedral type meshes were subdivided into 5 simplices. The $\mathbb{RT}_0$ elements are then obtained by solving a local Neumann problem. Here, we describe a generic triangulation for polytopal meshes, and how to reconstruct $\mathbb{RT}_0$ elements for these.

For the 2 dimensional case, a triangulation of each cell K is performed by choosing a point $\boldsymbol{x}_K$ in the interior of K, and forming a triangle T_{K,σ_i} with apex $\boldsymbol{x}_K$ and base σ_i for each edge $\sigma_i \in \mathscr{E}_K$ (Fig. 1, left). For simplicity of notation, we order the edges σ_i, and thus the corresponding associated triangles T_{K,σ_i} of cell K in counterclockwise order. We then denote by σ_i^* the edge shared between T_{K,σ_i} and $T_{K,\sigma_{i+1}}$, $i = 1, \ldots, n_e$, with the convention that $\sigma_{n_e+1} = \sigma_1$ (see Fig. 1, left).

Now, we present how to divide a cell into simplices in 3D. We start by picking a point $\boldsymbol{x}_K$ in the interior of K and forming, for $i = 1, 2, \ldots, n_f$, sub-cells with vertices $\boldsymbol{x}_K$, $v_{\sigma_{i,k}}$ for $k = 1, 2, \ldots, (n_v)_i$, where n_f and $(n_v)_i$ denote the number of faces of the cell K and the number of vertices in the face σ_i of cell K, respectively. We then denote the sub-cell in cell K associated to face σ_i as K_i (Fig. 1, right). This results to n_f polyhedra, as each face corresponds to one sub-cell.

A simplex $S_{i,j}$ is then constructed by joining the point $\boldsymbol{x}_K$ to a triangular base $T_{i,j}$ formed by joining the edge $e_{i,j}$ being shared by the faces σ_i and σ_j, to a point $\boldsymbol{x}_{\sigma_i}$ on the face σ_i (Fig. 2). Over a simplex S of dimension d, an $\mathbb{RT}_0$ finite element is defined to be $\mathbb{RT}_0(S) := (\mathbb{P}^0(S))^d + \boldsymbol{x}\mathbb{P}^0(S)$. This space has $d + 1$ degrees of freedom, but as can be seen in (2), finite volume schemes only provide an approximation for the fluxes $F_{K,\sigma}$ on the boundary ∂K. Hence, in order to reconstruct functions in $\mathbb{RT}_0$, we need to find approximations for the fluxes on the internal faces.

Denote by $\mathscr{E}_K^*$ the set of internal faces of K, that is, all of the faces of the simplices $S \in \mathscr{S}_K$, that do not lie on $\sigma \in \mathscr{E}_K$. Then, for every simplex $S \in \mathscr{S}_K$, denote by σ_S the face of K on which it sits, $\widetilde{\sigma}_S$ the part of σ_S it occupies, and $\mathscr{E}_S^*$ its internal faces (that is, all its faces except σ_S). For every internal face $\sigma^* \in \mathscr{E}_S^*$ that lies on a simplex S, we denote by S' the simplex which shares σ^* with S. In order to preserve the divergence in (1), we impose the balance equation (so that the divergence of

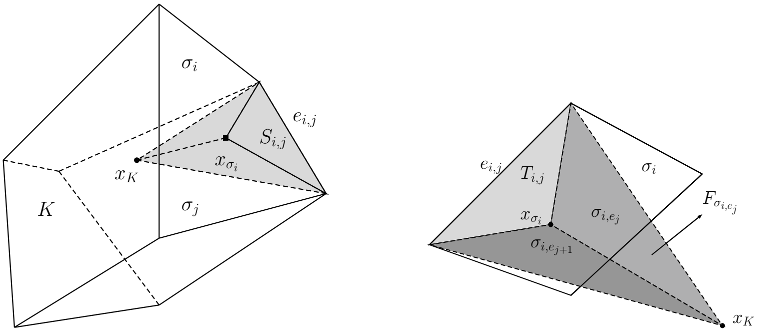


Fig. 2 Left: triangulation in 3D; Right: simplex sliced from a sub-cell

these fluxes in each simplex is equal to the divergence of the fluxes on K), and the conservativity of the fluxes:

$$\forall S \in \mathscr{S}_K\ ,\ \sum_{\sigma^* \in \mathscr{E}_S^*} F_{S,\sigma^*} + \frac{|\widetilde{\sigma}_S|}{|\sigma_S|} F_{K,\sigma_S} = \frac{|S|}{|K|} \sum_{\sigma' \in \mathscr{E}_K} F_{K,\sigma'}, \tag{3a}$$

$$\forall \sigma^* \in \mathscr{E}_S^*\ ,\ F_{S,\sigma^*} + F_{S',\sigma^*} = 0. \tag{3b}$$

The second term in the left hand side of (3a) is the contribution of the external face $\widetilde{\sigma}_S$ of S, on which we assume that the flux is the corresponding proportion of the flux $F_{K,\sigma}$ (In 2D, we have $\frac{|\widetilde{\sigma}_S|}{|\sigma_S|} = 1$). We note however that the system (3) is rank-deficient. This can be seen by taking the sum over all $S \in \mathscr{S}_K$ in (3a), which, in view of (3b), leads to the trivial relation $\sum_{\sigma \in \mathscr{E}_K} F_{K,\sigma} = \sum_{\sigma \in \mathscr{E}_K} F_{K,\sigma}$. We illustrate three methods (Sects. 3.1–3.3) to deal with the rank deficiency.

3.1 Minimal l^2 Norm (KR Method)

We may take the solution to (3) with minimal l^2 norm (both for 2D and 3D), as in [9]. This method will be referred to as the *KR method* (attributed to Y. Kuznetsov and S. Repin). However, on distorted meshes, the internal fluxes F_{K,σ^*} obtained from the least norm solution of (3) cannot construct constant velocity fields accurately.

3.2 Consistency Condition (C Method)

In 2D, the system (3) consists of n_e equations in n_e unknowns, and it can be shown that it is rank-deficient by 1. Hence, we remove one of the n_e equations, and replace it with a 'consistent' equation (hence *C method*) so that (3) is of full rank. A generic method for finding an equation so that the recovered fluxes can reconstruct $\mathbb{RT}_0$ velocities exactly is described in [5]. One such equation can be obtained by taking $\boldsymbol{x}_K$ as the barycenter of K, and setting $\sum_{i=1}^{n_e} F_{\sigma_i^*} = 0$. This particular choice is equivalent to the composite polygonal mixed finite element described in [2]. An expression for the fluxes can then be found explicitly. This means that we do not have to solve any local system, which is one of the main advantages of the C method over the KR method. The weakness of this method is that it is highly dependent on the fact that we only need one closing equation in 2D. An extension onto 3D pyramidal cells has been presented in [3]; however, extension onto meshes with more generic geometries is non-trivial.

3.3 Introducing Auxiliary Cell-Centered Unknowns (A Method)

In this section, we look for internal fluxes that are composed of a consistent flux coming from a constant velocity in the cell, and a stabilisation term, similar to a Brezzi-Pitkäranta stabilisation (a discrete inconsistent Laplacian on the submesh). This was actually inspired by the post-processing technique in [1]. After cutting the cell K into simplices, for each simplex $S \in \mathscr{S}_K$ we look for internal fluxes of the form

$$\forall \sigma^* \in \mathscr{E}_S^* \,,\;\; F_{S,\sigma^*} = \bar{F}_{S,\sigma^*} + \frac{|\sigma^*|}{h_{\sigma^*}}(Q_S - Q_{S'}), \tag{4}$$

where S' is the simplex on the other side of σ^*, h_{σ^*} is a characteristic distance between S and S' (for example, the distance between their centers of mass), and $(Q_S)_{S\in\mathscr{S}_K}$ are real numbers (if $\mathbf{u}$ is a Darcy velocity $-\Lambda\nabla p$, then these could be considered as potentials inside each simplex). Here, $\bar{F}_{S,\sigma^*} = |\sigma^*|\mathbf{u}_K \cdot \mathbf{n}_{S,\sigma^*}$ and $\mathbf{u}_K = \frac{1}{|K|}\sum_{\sigma\in\mathscr{E}_K} F_{K,\sigma}(\overline{\boldsymbol{x}}_\sigma - \boldsymbol{x}_K)$. As a remark, we note that if the fluxes $F_{K,\sigma}$ come from a constant velocity field ξ, then this definition of $\mathbf{u}_K$ allows us to recover ξ exactly.

Substituting (4) into (3), we obtain the following square system on the unknowns $(Q_S)_{S\in\mathscr{S}_K}$:

$$\forall S \in \mathscr{S}_K \,,\; \sum_{\sigma^*\in\mathscr{E}_S^*} \frac{|\sigma^*|}{h_{\sigma^*}}(Q_S - Q_{S'}) = b_S, \tag{5}$$

where b_S depends on $\mathbf{u}_K$ and the fluxes around K.

Lemma 1 *If $(Q_S)_{S\in\mathscr{S}_K}$ is a solution of (5), then*

$$\sum_{\sigma^*\in\mathscr{E}_K^*} \frac{|\sigma^*|}{h_{\sigma^*}}(Q_S - Q_{S'})^2 = \sum_{S\in\mathscr{S}_K} b_S Q_S.$$

Proof This can be established by multiplying (5) with Q_S, taking the sum over $S \in \mathscr{S}_K$, and gathering the terms by internal faces $\sigma^* \in \mathscr{E}_K^*$. ■

In particular, if $(b_S)_{S\in\mathscr{S}_K} = 0$ then all $(Q_S)_{S\in\mathscr{S}_K}$ are identical and thus $F_{S,\sigma^*} = \bar{F}_{S,\sigma^*}$. As a consequence, the A method can reconstruct constant velocity fields exactly. Also, the matrix of (5) is rank-deficient by 1, since it only has the constant vector **1** in its kernel. Velocities reconstructed from fluxes that satisfy (4)–(5) will be denoted as *A velocities* ('A' for 'auxiliary').

3.3.1 Detailed Computations in 2D

Here, we start by introducing an auxiliary unknown Q_i associated to each of the sub-cells T_{K,σ_i} as in (4). Writing $\beta_i = \frac{|\sigma_i^*|}{h_{\sigma_i^*}}$, the system of Eq. (5) can be written as $A\mathbf{Q} = \mathbf{b}$ with unknowns $\mathbf{Q} = (Q_1, \ldots, Q_{n_e})$. Here, $A = P^T DP$ where D is the diagonal matrix with $d_{i,i} = \beta_i$, $i = 1, \ldots, n_e$ and P is the matrix such that for $i = 1, \ldots, n_e$, $p_{i,i} = 1$, $p_{i,i+1} = -1$, where the entry p_{n_e,n_e+1} is equal to $p_{n_e,1}$. Setting $\bar{F}_{\sigma_0^*} = \bar{F}_{\sigma_{n_e}^*}$, $\mathbf{b}$ is composed of entries $b_i = \frac{|T_{K,\sigma_i}|}{|K|} \sum_{\sigma\in\mathscr{E}_K} F_{K,\sigma} - F_{K,\sigma_i} - \bar{F}_{\sigma_i^*} + \bar{F}_{\sigma_{i-1}^*}$. The quantities $q_i = (Q_i - Q_{i+1})$ can uniquely be determined by removing any one of the equations in (3) and replacing it with a relation involving the q_i's. In particular, we remove the equation corresponding to the n_e-th row of A, and replace it with $\sum_{i=1}^{n_e} q_i = 0$. This yields a matrix system $\hat{A}\mathbf{q} = \hat{\mathbf{b}}$, where $\mathbf{q}$ is the $n_e \times 1$ vector with ith entry q_i, $\hat{\mathbf{b}}$ is the $n_e \times 1$ column vector with the first $n_e - 1$ entries identical to $\mathbf{b}$ but with last entry 0, and $\hat{A}$ being the matrix formed by the first $n_e - 1$ rows of $P^T D$, augmented by the $1 \times n_e$ row vector of all ones **1**. We note now that $\hat{A}$ has full rank, consisting of the equations

$$\beta_j q_j = \beta_{n_e} q_{n_e} + \sum_{i=1}^{j} b_i,\ j = 1, \ldots, n_e - 1, \quad \text{and} \quad \sum_{i=1}^{n_e} q_i = 0. \tag{6}$$

The values q_i, $i = 1, \ldots, n_e$, and hence the fluxes, can then be explicitly obtained by solving (6).

3.3.2 A Method in 3D

To implement the A method in 3D, the idea is to perform a 2-step process, the first of which involves solving a local linear system of equations, followed by a second step, which gives explicit expressions for the fluxes. We start by partitioning each cell K into n_f sub-cells as in Fig. 1, right. Auxiliary cell-centered unknowns (1 for each sub-cell) are then introduced as in Sect. 3.3, but over generic polyhedrons, instead of simplices. Each interior flux F_{σ_i,σ_j} then corresponds to the face $\sigma_{i,j}$ formed by joining the interior point $\boldsymbol{x}_K$ to an edge $e_{i,j}$ of K being shared by the faces σ_i and σ_j

of cell K. An analog of (4) is then given by $F_{\sigma_i,\sigma_j} = \bar{F}_{\sigma_i,\sigma_j} + \frac{|\sigma_{i,j}|}{h_{i,j}}(Q_i - Q_j)$, with $h_{i,j}$ being a characteristic distance between the sub-cells K_i and K_j. We then write (3a) for generic cells (i.e. the average divergence of the fluxes for each sub-cell is equal to the average divergence of the fluxes on K). Denoting by $(n_e)_i$ the number of edges of a face σ_i of K, we then have, on each sub-cell K_i,

$$\sum_{j=1}^{(n_e)_i} \beta_{i,j}(Q_i - Q_j) = \frac{|K_i|}{|K|} \sum_{\sigma \in \mathscr{E}_K} F_{K,\sigma} - \sum_{j=1}^{(n_e)_i} \bar{F}_{\sigma_i,\sigma_j} - F_{K,\sigma_i}, \tag{7}$$

where $\beta_{i,j} = \frac{|\sigma_{i,j}|}{h_{i,j}}$. Owing to Lemma 1, the values $(Q_i - Q_j)$ can be uniquely determined by fixing one of the values Q_i. Hence, after setting the value of Q_1, we can solve a local system of size $(n_f - 1) \times (n_f - 1)$ to determine the values of Q_j, $j = 2, \ldots n_f$, and thus the fluxes F_{σ_i,σ_j}. At this stage, we recall that our aim is to reconstruct $\mathbb{RT}_0$ functions over simplices. Hence, we proceed by breaking each of the sub-cells K_i into simplices. On each sub-cell K_i, we pick a point $\boldsymbol{x}_{\sigma_i}$ on the face σ_i and associate, for each edge e_j $(j = 1, \ldots (n_e)_i)$ of the face σ_i an interior face σ_{i,e_j}. We then form a simplex $S_{i,j}$ with base on the sub-triangle $T_{i,j}$ in σ_i and faces $\sigma_{i,j}$, σ_{i,e_j}, $\sigma_{i,e_{j+1}}$ (Fig. 2, right). Since (7) ensures that each sub-cell K_i preserves the divergence of the entire cell K, we only need each simplex to preserve the divergence of the sub-cell K_i it resides in. Hence, we set for each edge e_j of the face σ_i,

$$F_{\sigma_i,e_j} + F_{\sigma_i,e_{j+1}} = \frac{|S_{i,j}|}{|K_i|} \sum_{\sigma \in \mathscr{E}_{K_i}} F_{K,\sigma} - F_{\sigma_i,\sigma_j} - \frac{|T_{i,j}|}{|\sigma_i|} F_{K,\sigma_i}, \tag{8}$$

where F_{σ_i,e_j} is the interior flux along the interior face σ_{i,e_j}, oriented outward of the simplex $S_{i,j}$. Each sub-cell consists of $(n_e)_i$ simplices and $(n_e)_i$ interior fluxes, which corresponds to $(n_e)_i$ equations in $(n_e)_i$ unknowns. System (8) is essentially the same as the one in 2D (i.e. the connectivity is determined through the adjacency of the triangles $T_{i,j}$ and $T_{i,j+1}$, corresponding to the edges e_j and e_{j+1} of the face σ_i, respectively). Expressions for the fluxes can then be obtained explicitly.

4 Numerical Tests in 2D

In this section, we present numerical results that illustrate the advantages of the C and A methods over the KR method. The approximate fluxes $(F_{K,\sigma})_{K \in \mathscr{M}, \sigma \in \mathscr{E}_K}$, are obtained by solving (1) with the HMM method [8]. The sub-interior fluxes are then obtained through the methods described in Sects. 3.1–3.3, and used to construct flow densities that are $\mathbb{RT}_0$ functions. We will consider tests on the domain $\Omega = (0, 1) \times (0, 1)$, for 2 types of velocity fields: $V = (0, 1)$ and $V = -\Lambda \nabla p = (\pi \sin(\pi x) \cos(\pi y), \pi \cos(\pi x) \sin(\pi y))$. The purpose of testing $V = (0, 1)$ is to illustrate that the KR method is not able to reconstruct constant velocities exactly

Table 1 Relative errors in the velocity reconstruction

Mesh	KR	C	A
Cartesian	1.18e-14	1.18e-14	1.18e-14
Hexahedral	3.64e-02	2.91e-13	2.90e-13
Kershaw	3.05e-01	5.14e-14	5.10e-14

(a) $V = (0, 1)$

Cartesian	KR	C	A
$h = 0.1768$	1.02e-01	1.02e-01	1.02e-01
$h = 0.0884$	5.12e-02	5.12e-02	5.12e-02
$h = 0.0442$	2.56e-02	2.56e-02	2.56e-02

(b) $V = (\pi \sin(\pi x)\cos(\pi y), \pi \cos(\pi x)\sin(\pi y))$

Table 2 Relative errors in the velocity reconstruction

Hexahedral	KR	C	A
$h = 0.2414$	1.00e-01	8.07e-02	8.01e-02
$h = 0.1297$	5.79e-02	3.96e-02	3.92e-02
$h = 0.0657$	3.52e-02	1.94e-02	1.92e-02

(a) $V = (\pi \sin(\pi x)\cos(\pi y), \pi \cos(\pi x)\sin(\pi y))$

Kershaw	KR	C	A
$h = 0.5154$	6.04e-01	4.72e-01	4.14e-01
$h = 0.2881$	2.99e-01	2.40e-01	1.95e-01
$h = 0.1517$	1.58e-01	1.32e-01	1.07e-01

(b) $V = (\pi \sin(\pi x)\cos(\pi y), \pi \cos(\pi x)\sin(\pi y))$

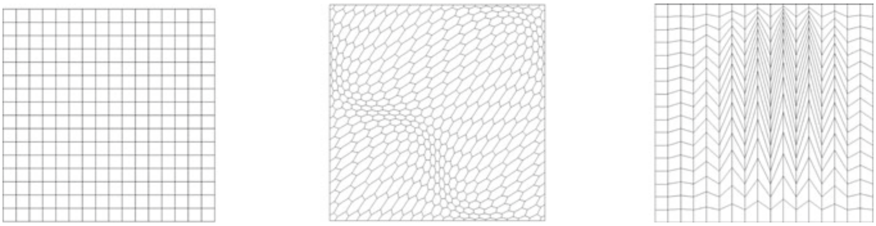


Fig. 3 Mesh types: Cartesian (left); Hexahedral (middle); Kershaw (right)

on distorted cells (see Table 1a). On the other hand, as expected, using the A and C methods enable us to recover V up to machine precision, regardless of the mesh. For the second test case, we aim to illustrate the first order accuracy of the velocity reconstructions on a variety of meshes. These are presented in Tables 1b and 2a–b. The coarsest and finest mesh consists of 8×8 and 32×32 cells, respectively. Here, the parameter h denotes the maximum diameter of the cells in the mesh.

Looking at Table 1b, we see that the KR, C, and A methods are equivalent for Cartesian meshes, and all of them exhibit first order convergence. Now, we look at the distorted meshes in Table 2a and b. Here, it can be noted that although all three methods exhibit first order convergence, the C and A methods yield much better reconstructions than the KR method. We also notice that on the very distorted Kershaw mesh, the A method performs slightly better than the C method (Fig. 3).

5 Summary

In this work, we presented post-processing techniques for fluxes obtained from finite volume methods. Tests in 2D show that the new methods (C and A) are better than the KR method. Details on how to implement the A method in 3D are also presented. Future work will focus on the possibility of obtaining higher order moments along the interior faces, which will be useful for reconstructing $\mathbb{RT}_k$ elements for $k \geq 1$.

References

1. Beirão da Veiga, L., Manzini, G., Putti, M.: Post processing of solution and flux for the nodal mimetic finite difference method. Numer. Methods Part. Differ. Equ. **31**(1), 336–363 (2015)
2. Birgle, N., Jaffré, J., Roberts, J.E.: A 2-D Composite Polygonal Mixed Finite Element. hal-01251652 (2015)
3. Birgle, N.: Underground flow, numerical methods and high performance computing. Ph.D. thesis, Université Pierre et Marie Curie - Paris VI (2016)
4. Brezzi, F., Fortin, M.: Mixed and Hybrid Finite Element Methods. Springer Series in Computational Mathematics, vol. 15. Springer, New York (1991)
5. Cheng, H.M., Droniou, J.: An HMM-ELLAM scheme on generic polygonal meshes for miscible incompressible flows in porous media. J. Pet. Sci. Eng. **172**, 707–723 (2019)
6. Droniou, J., Eymard, R., Gallouët, T., Guichard, C., Herbin, R.: The gradient discretisation method. Mathematics & Applications, vol. 82. Springer, Berlin (2018)
7. Droniou, J.: Finite volume schemes for diffusion equations: introduction to and review of modern methods. Math. Models Methods Appl. Sci. **24**(8), 1575–1619 (2014)
8. Droniou, J., Eymard, R., Gallouët, T., Herbin, R.: A unified approach to mimetic finite difference, hybrid finite volume and mixed finite volume methods. Math. Models Methods Appl. Sci. **20**(2), 265–295 (2010)
9. Kuznetsov, Y., Repin, S.: New mixed finite element method on polygonal and polyhedral meshes. Russ. J. Numer. Anal. Math. Model. **18**(3) (2003)
10. Sboui, A., Jaffré, J., Roberts, J.E.: A composite mixed finite element for hexahedral grids. SIAM J. Sci. Comput. **31**, 2623–2645 (2009)

Exponential Decay to Equilibrium of Nonlinear DDFV Schemes for Convection-Diffusion Equations

Claire Chainais-Hillairet and Stella Krell

Abstract We introduce a nonlinear DDFV scheme for an anisotropic linear convection-diffusion equation with mixed boundary conditions and we establish the exponential decay of the scheme towards its steady-state.

Keywords Discrete duality finite volume scheme · Discrete entropy/dissipation relation · Long time behaviour

MSC (2010) 65M08 · 35B40

1 Motivation

We are interested in the numerical discretization of linear anisotropic convection-diffusion equations on almost general meshes. Let Ω be a polygonal connected open bounded subset of $\mathbb{R}^2$ and let $T > 0$. The boundary $\Gamma = \partial\Omega$ is divided into two parts $\Gamma = \Gamma^D \cup \Gamma^N$ with $\mathrm{m}(\Gamma^D) > 0$. The problem writes:

$$\partial_t u + \mathrm{div}\mathbf{J} = 0,\ \mathbf{J} = -\Lambda(\nabla u + u\nabla V)\ \text{ in } \Omega \times (0, T), \tag{1a}$$

$$\mathbf{J} \cdot \mathbf{n} = 0\ \text{ on } \Gamma^N \times (0, T),\ \text{ and } u = u^D\ \text{ on } \Gamma^D \times (0, T), \tag{1b}$$

$$u(\cdot, 0) = u_0\ \text{ in } \Omega. \tag{1c}$$

We assume that the initial condition u_0 belong to $L^\infty(\Omega)$ and is positive, that the exterior potential V belongs to $C^1(\bar{\Omega}, \mathbb{R})$. The anisotropy tensor is supposed to be bounded, symmetric and uniformly elliptic: there exists $\lambda^M \geq \lambda_m > 0$ such that

C. Chainais-Hillairet
Univ. Lille, CNRS, UMR 8524-Laboratoire Paul Painlevé, 59000 Lille, France
e-mail: claire.chainais@univ-lille.fr

S. Krell (✉)
Université Côte d'Azur, CNRS, Inria, LJAD, Nice, France
e-mail: stella.krell@univ-cotedazur.fr

R. Klöfkorn et al. (eds.), *Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples*, Springer Proceedings in Mathematics & Statistics 323,
https://doi.org/10.1007/978-3-030-43651-3_11

$$\lambda_m |\mathbf{v}|^2 \leq \Lambda(\mathbf{x})\mathbf{v}\cdot\mathbf{v} \leq \lambda^M |\mathbf{v}|^2 \quad \text{for all } \mathbf{v} \in \mathbb{R}^2 \text{ and almost all } \mathbf{x} \in \Omega. \tag{2}$$

We finally assume that the boundary data u^D corresponds to a thermal Gibbs equilibrium, which means the existence of $\rho > 0$ such that $u^D = \rho e^{-V}$ on Γ^D. This implies that $u^\infty = \rho e^{-V}$ is a steady-state of (1). Moreover, the exponential decay of the solution u to (1) towards u^∞ is well-known, see for instance [2] (for even more general results) and the references therein.

When designing numerical schemes for (1), it is crucial to ensure that the scheme has a similar large time behavior than the continuous model. This property is ensured by classical TPFA schemes with linear B-fluxes on admissible meshes when $\Lambda = I$, as shown in [4]. Unfortunately, these schemes cannot be used in the anisotropic case and/or on general meshes. In [3], a nonlinear DDFV scheme has been introduced for (1) in the case of Neumann boundary conditions. The convergence of the scheme has been proved and numerical experiments show the exponential decay of the scheme towards equilibrium. This last property has been recently established in [5]. The aim of this paper is to introduce the nonlinear DDFV scheme for (1) and to prove its exponential decay towards equilibrium. The result is obtained under an assumption of uniform boundedness of the discrete solution away from 0 and ∞; it is is stated in Theorem 1.

2 Presentation of the Numerical Scheme

2.1 Meshes and Notations

In order to define a DDFV scheme, we need to introduce three different meshes—the primal mesh, the dual mesh and the diamond mesh—and some associated notations, for more details and also illustrations see [1, 3].

The primal mesh denoted $\overline{\mathfrak{M}}$ is composed of the interior primal mesh $\mathfrak{M}$ (a partition of Ω with polygonal control volumes) and the set $\partial\mathfrak{M}$ of boundary edges seen as degenerate control volumes. For all $K \in \overline{\mathfrak{M}}$, we define x_K the center of K.

For any vertex x_{K^*} of the primal mesh satisfying $x_{K^*} \in \Omega$, we define a polygonal control volume K^* by connecting all the centers of the primal cells sharing x_{K^*} as vertex. The set of such control volumes is the interior dual mesh denoted $\mathfrak{M}^*$. For any vertex $x_{K^*} \in \partial\Omega$, we define a polygonal control volume K^* by connecting the centers x_K of the interior primal cells and the midpoints of the boundary edges sharing x_{K^*} as vertex. The set of such control volumes is the boundary dual mesh, denoted $\partial\mathfrak{M}^*$. Finally, the dual mesh is $\mathfrak{M}^* \cup \partial\mathfrak{M}^*$, denoted by $\overline{\mathfrak{M}^*}$.

For all neighboring primal cells K and L, we assume that $\partial K \cap \partial L$ is a segment, corresponding to an edge of the mesh $\mathfrak{M}$, denoted by $\sigma = K|L$. Let $\mathcal{E}$ be the set of such edges. We similarly define the set $\mathcal{E}^*$ of the edges of the dual mesh. For each couple $(\sigma, \sigma^*) \in \mathcal{E} \times \mathcal{E}^*$ such that $\sigma = K|L$ and $\sigma^* = K^*|L^*$ cross, we define the quadrilateral diamond $\mathcal{D}_{\sigma,\sigma^*}$ whose diagonals are σ and σ^* (if $\sigma \subset \partial\Omega$, it degenerates into a triangle). The set of the diamonds defines the diamond mesh $\mathfrak{D}$, which is a partition of Ω. Finally, the DDFV mesh is made of $\mathcal{T} = (\overline{\mathfrak{M}}, \overline{\mathfrak{M}^*})$ and $\mathfrak{D}$.

For each $K \in \overline{\mathfrak{M}}$ and $K^* \in \overline{\mathfrak{M}^*}$, we define m_K the measure of K, m_{K^*} the measure of K^*. For a diamond $\mathcal{D} = \mathcal{D}_{\sigma,\sigma^*}$, whose vertices are $(x_K, x_{K^*}, x_L, x_{L^*})$, we define: $x_{\mathcal{D}}$ its center ($\{x_{\mathcal{D}}\} = \sigma \cap \sigma^*$), m_σ and m_{σ^*} the lengths of the edges, $\mathrm{m}_{\mathcal{D}}$ its measure, $d_{\mathcal{D}}$ its diameter, $\alpha_{\mathcal{D}}$ the angle between (x_K, x_L) and (x_{K^*}, x_{L^*}). We have $\mathrm{m}_{\mathcal{D}} = \frac{1}{2}\mathrm{m}_\sigma \mathrm{m}_{\sigma^*} \sin(\alpha_{\mathcal{D}})$. We will also use two direct basis $(\boldsymbol{\tau}_{K^*\!,L^*}, \mathbf{n}_{\sigma K})$ and $(\mathbf{n}_{\sigma^* K^*}, \boldsymbol{\tau}_{K,L})$, where $\mathbf{n}_{\sigma K}$ is the unit normal to σ outward K, $\mathbf{n}_{\sigma^* K^*}$ is the unit normal to σ^* outward K^*, $\boldsymbol{\tau}_{K^*\!,L^*}$ is the unit tangent vector to σ, oriented from K^* to L^*, $\boldsymbol{\tau}_{K,L}$ is the unit tangent vector to σ^*, oriented from K to L.

We define two local regularity factors $\theta_{\mathcal{D}}, \tilde{\theta}_{\mathcal{D}}$ of the diamond $\mathcal{D}$ by

$$\theta_{\mathcal{D}} = \frac{1}{2\sin(\alpha_{\mathcal{D}})}\left(\frac{\mathrm{m}_\sigma}{\mathrm{m}_{\sigma^*}} + \frac{\mathrm{m}_{\sigma^*}}{\mathrm{m}_\sigma}\right), \; \tilde{\theta}_{\mathcal{D}} = \max\left(\max_{\substack{K\in\mathfrak{M},\\ \mathrm{m}_{\mathcal{D}\cap K}>0}} \frac{\mathrm{m}_{\mathcal{D}}}{\mathrm{m}_{\mathcal{D}\cap K}} ; \max_{\substack{K^*\in\mathfrak{M}^*,\\ \mathrm{m}_{\mathcal{D}\cap K^*}>0}} \frac{\mathrm{m}_{\mathcal{D}}}{\mathrm{m}_{\mathcal{D}\cap K^*}}\right)$$

and we assume the following regularity of the mesh:

$$\exists \Theta \geq 1 \text{ such that } 1 \leq \theta_{\mathcal{D}}, \tilde{\theta}_{\mathcal{D}} \leq \Theta, \qquad \forall \mathcal{D} \in \mathfrak{D}.$$

Finally, we define the approximation $\Lambda^{\mathcal{D}}$ of the anisotropy tensor Λ on each diamond $\mathcal{D} \in \mathfrak{D}$ as the mean value of Λ over $\mathcal{D}$.

2.2 Discrete Unknowns and Discrete Operators

We need several types of degrees of freedom to represent scalar and vector fields in the discrete setting. $\mathbb{R}^{\mathcal{T}}$ is the linear space of scalar fields which are associated to each primal and dual cell and $(\mathbb{R}^2)^{\mathfrak{D}}$ the linear space of vector fields constant on the diamonds:

$$\begin{aligned} u_{\mathcal{T}} \in \mathbb{R}^{\mathcal{T}} &\iff u_{\mathcal{T}} = \left((u_K)_{K\in\overline{\mathfrak{M}}}, (u_{K^*})_{K^*\in\overline{\mathfrak{M}^*}}\right) \\ \boldsymbol{\xi}_{\mathfrak{D}} \in (\mathbb{R}^2)^{\mathfrak{D}} &\iff \boldsymbol{\xi}_{\mathfrak{D}} = \left(\boldsymbol{\xi}_{\mathcal{D}}\right)_{\mathcal{D}\in\mathfrak{D}} \end{aligned}$$

We also define a positive semi-definite bilinear form on $\mathbb{R}^{\mathcal{T}}$ and a scalar product on $(\mathbb{R}^2)^{\mathfrak{D}}$ by

$$\begin{aligned} [\![v_{\mathcal{T}}, u_{\mathcal{T}}]\!]_{\mathcal{T}} &= \frac{1}{2}\left(\sum_{K\in\mathfrak{M}} \mathrm{m}_K u_K v_K + \sum_{K^*\in\overline{\mathfrak{M}^*}} \mathrm{m}_{K^*} u_{K^*} v_{K^*}\right), \quad \forall u_{\mathcal{T}}, v_{\mathcal{T}} \in \mathbb{R}^{\mathcal{T}}, \\ \left(\boldsymbol{\xi}_{\mathfrak{D}}, \boldsymbol{\varphi}_{\mathfrak{D}}\right)_{\Lambda,\mathfrak{D}} &= \sum_{\mathcal{D}\in\mathfrak{D}} \mathrm{m}_{\mathcal{D}}\, \boldsymbol{\xi}_{\mathcal{D}} \cdot \Lambda^{\mathcal{D}} \boldsymbol{\varphi}_{\mathcal{D}}, \quad \forall \boldsymbol{\xi}_{\mathfrak{D}}, \boldsymbol{\varphi}_{\mathfrak{D}} \in (\mathbb{R}^2)^{\mathfrak{D}}. \end{aligned}$$

The associated norms are respectively denoted by $\|\cdot\|_{2,\mathscr{T}}$ and $\|\cdot\|_{\Lambda,\mathfrak{D}}$.

The DDFV method is based on the definitions of a discrete gradient $\nabla^{\mathfrak{D}}$, of a discrete divergence $\mathrm{div}^{\mathscr{T}}$ and a duality formula (see [1]). Here we do not recall the definition of the discrete operators as we will use a compact form of the scheme, as in [3]. For $u_{\mathscr{T}} \in \mathbb{R}^{\mathscr{T}}$, we just define $\delta^{\mathfrak{D}} u_{\mathscr{T}} = (\delta^{\mathscr{D}} u_{\mathscr{T}})_{\mathscr{D}\in\mathfrak{D}}$ by $\delta^{\mathscr{D}} u_{\mathscr{T}} = \begin{pmatrix} u_K - u_L \\ u_{K^*} - u_{L^*} \end{pmatrix}$ for all $\mathscr{D} \in \mathfrak{D}$. Then, the usual definition of the discrete gradient ensures that:

$$(\nabla^{\mathfrak{D}} u_{\mathscr{T}}, \nabla^{\mathfrak{D}} v_{\mathscr{T}})_{\Lambda,\mathfrak{D}} = \sum_{\mathscr{D}\in\mathfrak{D}} \delta^{\mathscr{D}} u_{\mathscr{T}} \cdot \mathbb{A}^{\mathscr{D}} \delta^{\mathscr{D}} v_{\mathscr{T}}, \quad \forall u_{\mathscr{T}}, v_{\mathscr{T}} \in \mathbb{R}^{\mathscr{T}},$$

where for all $\mathscr{D} \in \mathfrak{D}$

$$\mathbb{A}^{\mathscr{D}} = \frac{1}{4\mathrm{m}_{\mathscr{D}}} \begin{pmatrix} \mathrm{m}_\sigma^2 \mathbf{n}_{\sigma K} \cdot \Lambda^{\mathscr{D}} \mathbf{n}_{\sigma K} & \mathrm{m}_\sigma \mathrm{m}_{\sigma^*} \mathbf{n}_{\sigma K} \cdot \Lambda^{\mathscr{D}} \mathbf{n}_{\sigma^* K^*} \\ \mathrm{m}_\sigma \mathrm{m}_{\sigma^*} \mathbf{n}_{\sigma K} \cdot \Lambda^{\mathscr{D}} \mathbf{n}_{\sigma^* K^*} & \mathrm{m}_{\sigma^*}^2 \mathbf{n}_{\sigma^* K^*} \cdot \Lambda^{\mathscr{D}} \mathbf{n}_{\sigma^* K^*} \end{pmatrix} = \begin{pmatrix} A_{\sigma,\sigma}^{\mathscr{D}} & A_{\sigma,\sigma^*}^{\mathscr{D}} \\ A_{\sigma^*,\sigma}^{\mathscr{D}} & A_{\sigma^*,\sigma^*}^{\mathscr{D}} \end{pmatrix}.$$

Finally, we introduce a reconstruction operator on diamonds $r^{\mathfrak{D}}$. It is a mapping from $\mathbb{R}^{\mathscr{T}}$ to $\mathbb{R}^{\mathfrak{D}}$ defined for all $u_{\mathscr{T}} \in \mathbb{R}^{\mathscr{T}}$ by $r^{\mathfrak{D}} u_{\mathscr{T}} = \left(r^{\mathscr{D}} u_{\mathscr{T}}\right)_{\mathscr{D}\in\mathfrak{D}}$, where for $\mathscr{D} \in \mathfrak{D}$, whose vertices are $x_K, x_L, x_{K^*}, x_{L^*}$, $r^{\mathscr{D}} u_{\mathscr{T}} = \frac{1}{4}(u_K + u_L + u_{K^*} + u_{L^*})$.

2.3 The Scheme

A nonlinear DDFV scheme for the convection-diffusion equation (1a) with Neumann boundary conditions has already been introduced and analyzed in [3]. In this paper, we want to take into account Dirichlet boundary conditions on the part Γ^D of the boundary. We assume that the primal mesh is compatible with the partition of $\partial\Omega$. Let us introduce the set of Dirichlet boundary primal and dual cells:

$$\partial\mathfrak{M}_D = \{K \in \partial\mathfrak{M} : K \subset \Gamma_D\}, \qquad \partial\mathfrak{M}_D^* = \{K^* \in \partial\mathfrak{M}^* : x_{K^*} \in \overline{\Gamma}_D\},$$

Then, for a given $v \in C(\Gamma^D)$, we define

$$\mathrm{E}_v^{\Gamma_D} = \{u_{\mathscr{T}} \in \mathbb{R}^{\mathscr{T}}, \text{ s. t. } \forall K \in \partial\mathfrak{M}_D,\ u_K = v(x_K) \text{ and } \forall K^* \in \partial\mathfrak{M}_D^*,\ u_{K^*} = v(x_{K^*})\}.$$

Let Δt be a time step. We first discretize the initial condition by taking the mean values of u_0 on the primal and dual cells and the exterior potential V by taking its nodal values on the primal and dual cells. It defines $u_{\mathscr{T}}^0$ and $V_{\mathscr{T}}$. Then, for all $n \geq 0$, we look for $u_{\mathscr{T}}^{n+1} \in \mathrm{E}_{u^D}^{\Gamma_D}$ solution to:

$$\left[\!\left[\frac{u_{\mathscr{T}}^{n+1} - u_{\mathscr{T}}^n}{\Delta t}, \psi_{\mathscr{T}} \right]\!\right]_{\mathscr{T}} + T_{\mathfrak{D}}(u_{\mathscr{T}}^{n+1}; g_{\mathscr{T}}^{n+1}, \psi_{\mathscr{T}}) = 0, \quad \forall \psi_{\mathscr{T}} \in \mathrm{E}_0^{\Gamma_D}, \tag{3a}$$

$$T_{\mathfrak{D}}(u_{\mathcal{T}}^{n+1}; g_{\mathcal{T}}^{n+1}, \psi_{\mathcal{T}}) = \sum_{\mathcal{D}\in\mathfrak{D}} r^{\mathcal{D}} u_{\mathcal{T}}^{n+1}\, \delta^{\mathcal{D}} g_{\mathcal{T}}^{n+1} \cdot \mathbb{A}^{\mathcal{D}} \delta^{\mathcal{D}} \psi_{\mathcal{T}}, \tag{3b}$$

$$g_{\mathcal{T}}^{n+1} = \log(u_{\mathcal{T}}^{n+1}) + V_{\mathcal{T}}. \tag{3c}$$

The scheme is written here under a compact form. But it can also be expanded on primal and dual meshes after the introduction of conservative numerical fluxes.

3 Main Results

Steady-state. As the boundary conditions are at thermal equilibrium, $u^D = \rho e^{-V}$ on Γ^D with $\rho > 0$. Then, $u_{\mathcal{T}}^{\infty} = \rho e^{-V_{\mathcal{T}}}$ belongs to $\mathbb{E}_{u^D}^{\Gamma_D}$ and verifies $\delta^{\mathcal{D}}(\log u_{\mathcal{T}}^{\infty} + V_{\mathcal{T}}) = 0$ for all $\mathcal{D} \in \mathfrak{D}$, so that it is a steady-state to the scheme (3). Let us remark that, due to the definition of the steady-state, it is clearly bounded: there exists $m^{\infty} > 0$ and $M^{\infty} > 0$, such that $m^{\infty} \leq u_{\mathcal{T}}^{\infty} \leq M^{\infty}$.

Entropy-dissipation estimate. Let $\Phi_1 : x \mapsto x \log x - x + 1$ the Gibbs entropy. We define the discrete relative entropy $(\mathbb{E}_{1,\mathcal{T}}^{n})_{n\geq 0}$ and its associated discrete dissipation $(\mathbb{I}_{1,\mathcal{T}}^{n+1})_{n\geq 0}$ by:

$$\mathbb{E}_{1,\mathcal{T}}^{n} = \left[\!\!\left[u_{\mathcal{T}}^{\infty} \Phi_1\left(\frac{u_{\mathcal{T}}^{n}}{u_{\mathcal{T}}^{\infty}}\right), 1_{\mathcal{T}} \right]\!\!\right], \quad \forall n \geq 0$$

$$\mathbb{I}_{1,\mathcal{T}}^{n+1} = T_{\mathfrak{D}}(u_{\mathcal{T}}^{n+1}; g_{\mathcal{T}}^{n+1}, g_{\mathcal{T}}^{n+1}), \quad \forall n \geq 0$$

The definition of the steady-state implies that $\delta^{\mathfrak{D}} g_{\mathcal{T}}^{n+1} = \delta^{\mathfrak{D}} \log(u_{\mathcal{T}}^{n+1}/u_{\mathcal{T}}^{\infty})$, so that

$$\mathbb{I}_{1,\mathcal{T}}^{n+1} = \sum_{\mathcal{D}\in\mathfrak{D}} r^{\mathcal{D}}(u_{\mathcal{T}}^{n+1})\, \delta^{\mathcal{D}} \log\left(\frac{u_{\mathcal{T}}^{n+1}}{u_{\mathcal{T}}^{\infty}}\right) \cdot \mathbb{A}^{\mathcal{D}} \delta^{\mathcal{D}} \log\left(\frac{u_{\mathcal{T}}^{n+1}}{u_{\mathcal{T}}^{\infty}}\right), \quad \forall n \geq 0. \tag{4}$$

Proposition 1 *Let assume that the scheme (3) has a solution $u_{\mathcal{T}}^{n+1} \in \mathbb{E}_{u^D}^{\Gamma_D} \cap (0, +\infty)^{\mathcal{T}}$ for all $n \geq 0$. Then, the following entropy-dissipation estimate holds:*

$$\frac{\mathbb{E}_{1,\mathcal{T}}^{n+1} - \mathbb{E}_{1,\mathcal{T}}^{n}}{\Delta t} + \mathbb{I}_{1,\mathcal{T}}^{n+1} \leq 0, \textit{ for all } n \geq 0. \tag{5}$$

Proof Due to the convexity of Φ_1 and the fact that $\Phi_1'(x) = \log x$, we have

$$\frac{\mathbb{E}_{1,\mathcal{T}}^{n+1} - \mathbb{E}_{1,\mathcal{T}}^{n}}{\Delta t} \leq \left[\!\!\left[\frac{u_{\mathcal{T}}^{n+1} - u_{\mathcal{T}}^{n}}{\Delta t} \log\left(\frac{u_{\mathcal{T}}^{n+1}}{u_{\mathcal{T}}^{\infty}}\right), 1_{\mathcal{T}} \right]\!\!\right].$$

Then, we obtain (5) by taking $\psi_{\mathcal{T}} = \log\left(u_{\mathcal{T}}^{n+1}/u_{\mathcal{T}}^{\infty}\right) \in \mathbb{E}_0^{\Gamma_D}$ in the scheme (3). ■

As a consequence of the entropy-dissipation estimate, we can obtain the existence of a solution to the scheme (3). The result is a consequence of the control of the dissipation implied by (5) and of a topological degree argument. We refer to [3] for the idea of the proof. For the study of the exponential decay, we will further assume that the solution to the scheme satisfies uniform bounds.

Exponential decay.

Theorem 1 *Assume that the solution to the scheme (3) is uniformly bounded:*

$$\exists m_* \in (0, m^\infty] \text{ and } M^* \in [M^\infty, +\infty) \text{ such that } m_* \le u^n_{\mathscr{T}} \le M^* \quad \forall n \ge 0. \tag{6}$$

Then, there exists ν depending only Ω, Θ, m_*, M^* and Λ, such that, for any $k > 0$, if $\Delta t \le k$,

$$\mathbb{E}^n_{1,\mathscr{T}} \le e^{-\tilde{\nu} t^n} \mathbb{E}^0_{1,\mathscr{T}}, \ \forall n \ge 0, \ \textit{with } \tilde{\nu} = \frac{1}{k}\log(1+\nu k). \tag{7}$$

Proof Based on the entropy-dissipation estimate (5), the proof consists in establishing the existence of some $\nu > 0$ such that

$$\mathbb{I}^{n+1}_{1,\mathscr{T}} \ge \nu \mathbb{E}^{n+1}_{1,\mathscr{T}}, \quad \forall n \ge 0. \tag{8}$$

The main difference for the proof of (8) when dealing with Dirichlet boundary conditions instead of Neumann boundary conditions (see [5]) is that we can not use a discrete log-Sobolev inequality. The inequality (8) is based on some reformulations of the discrete entropy and entropy dissipation and on a discrete Poincaré inequality.

Using the definition of Φ_1, $\mathbb{E}^{n+1}_{1,\mathscr{T}}$ rewrites: $\mathbb{E}^{n+1}_{1,\mathscr{T}} = [\![u^{n+1}_{\mathscr{T}} \log(u^{n+1}_{\mathscr{T}}/u^\infty_{\mathscr{T}}) - u^{n+1}_{\mathscr{T}} + u^\infty_{\mathscr{T}}, 1_{\mathscr{T}}]\!]$. As $x\log(x/y) - x + y \le (x-y)^2/(2\min(x,y))$ for all $x, y > 0$, we obtain:

$$\mathbb{E}^{n+1}_{1,\mathscr{T}} \le \frac{1}{2m_*} \| u^{n+1}_{\mathscr{T}} - u^\infty_{\mathscr{T}} \|^2_{2,\mathscr{T}}, \quad \forall n \ge 0. \tag{9}$$

For all $\mathscr{D} \in \mathfrak{D}$, we introduce the diagonal matrix $\mathbb{B}^{\mathscr{D}}$, whose diagonal coefficients are $B^{\mathscr{D}}_{\sigma,\sigma} = |A^{\mathscr{D}}_{\sigma,\sigma}| + |A^{\mathscr{D}}_{\sigma,\sigma^*}|$ and $B^{\mathscr{D}}_{\sigma^*,\sigma^*} = |A^{\mathscr{D}}_{\sigma^*,\sigma^*}| + |A^{\mathscr{D}}_{\sigma,\sigma^*}|$. As shown in [3], there exists a constant $C(\Theta, \Lambda)$ depending only on Θ, λ_m and λ^M such that

$$\mathbf{w} \cdot \mathbb{A}^D \mathbf{w} \le \mathbf{w} \cdot \mathbb{B}^D \mathbf{w} \le C(\Theta, \Lambda) \mathbf{w} \cdot \mathbb{A}^D \mathbf{w}, \quad \forall \mathbf{w} \in \mathbb{R}^2, \ \forall \mathscr{D} \in \mathfrak{D}. \tag{10}$$

But, as $\mathbb{B}^{\mathscr{D}}$ is a diagonal matrix, for all $\mathscr{D} \in \mathfrak{D}$ we have:

$$\delta^{\mathscr{D}} \log \frac{u^{n+1}_{\mathscr{T}}}{u^\infty_{\mathscr{T}}} \cdot \mathbb{B}^{\mathscr{D}} \delta^{\mathscr{D}} \log \frac{u^{n+1}_{\mathscr{T}}}{u^\infty_{\mathscr{T}}} = B^{\mathscr{D}}_{\sigma,\sigma} \left(\log \frac{u^{n+1}_K}{u^\infty_K} - \log \frac{u^{n+1}_L}{u^\infty_L} \right)^2 + B^{\mathscr{D}}_{\sigma^*,\sigma^*} \left(\log \frac{u^{n+1}_{K^*}}{u^\infty_{K^*}} - \log \frac{u^{n+1}_{L^*}}{u^\infty_{L^*}} \right)^2.$$

As $(\log x - \log y)^2 \geq (x-y)^2/\max^2(x,y)$ for all $x, y > 0$, we deduce from (6) that

$$\delta^{\mathscr{D}} \log \frac{u_{\mathscr{T}}^{n+1}}{u_{\mathscr{T}}^{\infty}} \cdot \mathbb{B}^{\mathscr{D}} \delta^{\mathscr{D}} \log \frac{u_{\mathscr{T}}^{n+1}}{u_{\mathscr{T}}^{\infty}} \geq \left(\frac{m_*}{M^*}\right)^2 \delta^{\mathscr{D}} \frac{u_{\mathscr{T}}^{n+1}}{u_{\mathscr{T}}^{\infty}} \cdot \mathbb{B}^{\mathscr{D}} \delta^{\mathscr{D}} \frac{u_{\mathscr{T}}^{n+1}}{u_{\mathscr{T}}^{\infty}}.$$

From (4), (6), (10) (applied twice) and (2), we deduce that

$$\mathbb{I}_{1,\mathscr{T}}^{n+1} \geq C(\Theta, \Lambda) m_* \left(\frac{m_*}{M^*}\right)^2 \lambda_m \left\| \nabla^{\mathfrak{D}} \frac{u_{\mathscr{T}}^{n+1}}{u_{\mathscr{T}}^{\infty}} \right\|_{I,\mathfrak{D}}^2 . \tag{11}$$

Let us now apply the discrete Poincaré inequality to $u_{\mathscr{T}}^{n+1}/u_{\mathscr{T}}^{\infty} - 1 \in \mathrm{E}_0^{\Gamma_D}$ (see [1]). Combined with (6), this yields

$$\|u_{\mathscr{T}}^{n+1} - u_{\mathscr{T}}^{\infty}\|_{2,\mathscr{T}}^2 \leq C_P(\Omega) M^{*2} \left\| \nabla^{\mathfrak{D}} \frac{u_{\mathscr{T}}^{n+1}}{u_{\mathscr{T}}^{\infty}} \right\|_{I,\mathfrak{D}}^2 . \tag{12}$$

From (9), (11), (12), we finally deduce (8), with

$$\nu = C(\Theta, \Lambda, \Omega) \left(\frac{m_*}{M^*}\right)^4 .$$

This concludes the proof of Theorem 1.

4 Numerical Experiments

We consider a test case where $\Omega = (0,1)^2$, $V(x_1, x_2) = -x_1$, $\Gamma_D = \{x_1 = 0\} \cup \{x_1 = 1\}$ and the exact solution u_{ex} is defined by

$$u_{\mathrm{ex}}((x_1, x_2), t) = e^{-\alpha t + \frac{x_1}{2}} \sin(\pi x_1) + e^{x_1}$$

with $\alpha = \pi^2 + \frac{1}{4}$. We choose $u_0 = u_{\mathrm{ex}}(\cdot, 0)$.

In order to illustrate the convergence and the robustness of our method, we test its convergence on two sequences of meshes. The first sequence of primal meshes is made of successively refined Kershaw meshes. The second sequence of primal meshes is the so-called quadrangle meshes `mesh_quad_i` of the FVCA8 benchmark on incompressible flows. In the refinement procedure, the time step is divided by 4 when the mesh size is divided by 2. The nonlinear system (3) is solved thanks to Newton's method. In order to avoid the singularity of the log near 0, the sequence $(u_{\mathscr{T}}^{n+1,i})_{i\geq 0}$ to compute $u_{\mathscr{T}}^{n+1}$ from the previous state $(u_{\mathscr{T}}^{n})_{i\geq 0}$ is initialized by $u_{\mathscr{T}}^{n+1,0} = \max(u_{\mathscr{T}}^{n}, 10^{-12})$. As a stopping criterion, we require the ℓ^1-norm of the residual to be smaller than 10^{-10}. In Table 1, the quantities erru and errgu respectively

Table 1 Numerical results on the Quadrangle mesh family, final time T = 0.1

M	dt	errgu	ordgu	erru	ordu	$N_{\max}$	N_{mean}	Min u^n	Max u^n
1	1.613E-03	3.447E-02	–	5.208E-03	–	2	2	1.0	3.148
2	4.032E-04	1.578E-02	1.12	1.389E-03	1.90	2	2	1.0	3.161
3	1.008E-04	8.629E-03	0.92	4.467E-04	1.72	2	2	1.0	3.161
4	2.520E-05	3.934E-03	1.19	1.157E-04	2.04	2	1	1.0	3.162
5	6.300E-06	9.668E-04	1.66	2.402E-05	1.86	1	1	1.0	3.162

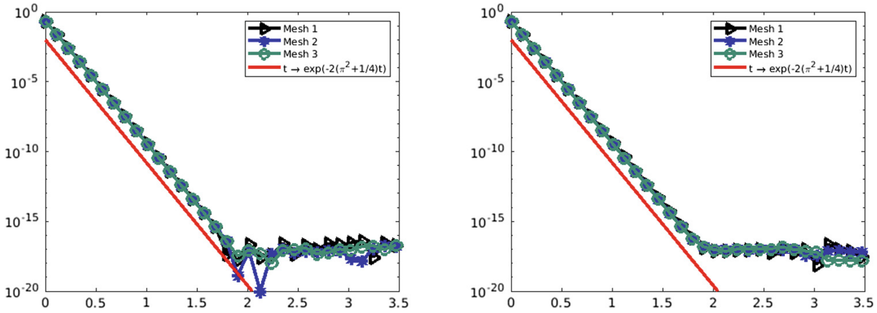


Fig. 1 Discrete relative entropy $\mathbb{E}^n_{1,\mathscr{T}}$ as a function of $n\Delta t$. Left: computed on the first three Quadrangle meshes. Right: computed on the first three Kershaw meshes

denote the $L^\infty((0,T); L^2(\Omega))$ error on the solution and the $L^2(\Omega \times (0,T))^2$ error on the gradient, whereas ordu and ordgu are the corresponding convergence orders. It appears that the method is second order accurate w.r.t. space.

The maximal (resp. mean) number of Newton iterations by time step is denoted by $N_{\max}$ (resp. N_{mean}). We observe that the needed number of Newton iterations starts from a reasonably small value and falls down to 1 after a small number of time steps. Therefore, our method does not imply an important extra computational cost when compared to linear methods. Eventually, we observe numerically that the numerical solution remains bounded in time along the simulation (the bounds are reached at the initial time), which validates the hypothesis (6) of Theorem 1.

In order to give an evidence of the good large-time behavior of our scheme, we plot in Fig. 1 the evolution of $\mathbb{E}^n_{1,\mathscr{T}}$ computed on the Kershaw and Quadrangle meshes. We observe the exponential decay of the relative energy.

References

1. Andreianov, B., Boyer, F., Hubert, F.: Discrete duality finite volume schemes for Leray-Lions-type elliptic problems on general 2D meshes. Numer. Methods Part. Differ. Equ. **23**(1), 145–195 (2007)
2. Bodineau, T., Lebowitz, J., Mouhot, C., Villani, C.: Lyapunov functionals for boundary-driven nonlinear drift-diffusion equations. Nonlinearity **27**, 2111–2132 (2014)
3. Cancès, C., Chainais-Hillairet, C., Krell, S.: Numerical analysis of a nonlinear free-energy diminishing discrete duality finite volume scheme for convection diffusion equations. Comput. Methods Appl. Math. **18**, 407–432 (2018)
4. Chainais-Hillairet, C., Herda, M.: Large-time behavior of a family of finite volume schemes for boundary-driven convection-diffusion equations. IMAJNA (2019)
5. Cancès, C., Chainais-Hillairet, C., Herda, M., Krell, S.: Large time behavior of nonlinear finite volume schemes for convection-diffusion equations. https://hal.archives-ouvertes.fr/hal-02360155 (2019)

L^∞ Bounds for Numerical Solutions of Noncoercive Convection-Diffusion Equations

Claire Chainais-Hillairet and Maxime Herda

Abstract In this work, we apply an iterative energy method *à la* de Giorgi in order to establish L^∞ bounds for numerical solutions of noncoercive convection-diffusion equations with mixed Dirichlet-Neumann boundary conditions.

Keywords Finite volume schemes · Uniform bounds · Noncoercive elliptic equations

MSC (2010) 65M08 · 35B40

1 Introduction

The continuous problem. Let Ω be an open, bounded and connected polygonal or polyhedral domain of $\mathbb{R}^d$ with $d = 2$ or 3. We denote by $\mathrm{m}(\cdot)$ both the Lebesgue and $d-1$ dimensional Hausdorff measure. Without loss of generality, we assume that $\mathrm{m}(\Omega) = 1$. We assume that $\partial\Omega = \Gamma^D \cup \Gamma^N$ with $\Gamma^D \cap \Gamma^N = \emptyset$ and $\mathrm{m}(\Gamma^D) > 0$ and we denote by $\mathbf{n}$ the exterior normal to $\partial\Omega$. Let $\mathbf{U} \in C(\bar{\Omega})^2$ be a velocity field, $b \in L^\infty(\Omega)$ assumed to be nonnegative, $f \in L^p(\Omega)$, with $p > d/2$ a source term and $v^D \in L^\infty(\Gamma^D)$ a boundary condition. We consider the following convection-diffusion equation with mixed boundary conditions:

C. Chainais-Hillairet
Univ. Lille, CNRS,UMR 8524, Inria - Laboratoire Paul Painlevé,
59000 Lille, France
e-mail: claire.chainais@univ-lille.fr

M. Herda (✉)
Inria, Univ. Lille, CNRS, UMR 8524 - Laboratoire Paul Painlevé,
59000 Lille, France
e-mail: maxime.herda@inria.fr

R. Klöfkorn et al. (eds.), *Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples*, Springer Proceedings in Mathematics & Statistics 323,
https://doi.org/10.1007/978-3-030-43651-3_12

$$\mathrm{div}(-\nabla v + \mathbf{U}v) + bv = f \qquad \text{in } \Omega, \tag{1a}$$

$$(-\nabla v + \mathbf{U}v) \cdot \mathbf{n} = 0 \qquad \text{on } \Gamma^N, \tag{1b}$$

$$v = v^D \qquad \text{on } \Gamma^D. \tag{1c}$$

This noncoercive elliptic linear problem has been widely studied by Droniou and coauthors, even with less regularity on the data, see for instance [2–5]. Nevertheless, up to our knowledge, the derivation of explicit L^∞ bounds on numerical solutions has not been done in the literature.

The numerical scheme. The mesh of the domain Ω is denoted by $\mathcal{M} = (\mathcal{T}, \mathcal{E}, \mathcal{P})$ and classically given by: $\mathcal{T}$, a set of open polygonal or polyhedral control volumes; $\mathcal{E}$, a set of edges or faces; $\mathcal{P} = (x_K)_{K\in\mathcal{T}}$, a set of points satisfying $x_K \in K$ for all $K \in \mathcal{T}$. In the following, we also use the denomination "edge" for a face in dimension 3. As we deal with a Two-Point Flux Approximation (TPFA) of convection-diffusion equations, we assume that the mesh is admissible in the sense of [6] (Definition 9.1).

We distinguish in $\mathcal{E}$ the interior edges, $\sigma = K|L$, from the exterior edges: $\mathcal{E} = \mathcal{E}_{int} \cup \mathcal{E}_{ext}$. Among the exterior edges, we distinguish the edges included in Γ^D from the edges included in Γ^N: $\mathcal{E}_{ext} = \mathcal{E}^D \cup \mathcal{E}^N$. For a given control volume $K \in \mathcal{T}$, we define $\mathcal{E}_K$ the set of its edges, which is also split into $\mathcal{E}_K = \mathcal{E}_{K,int} \cup \mathcal{E}_K^D \cup \mathcal{E}_K^N$.

Let $\mathrm{d}(\cdot,\cdot)$ denote the Euclidean distance. For all edges $\sigma \in \mathcal{E}$, we set $\mathrm{d}_\sigma = \mathrm{d}(x_K, x_L)$ if $\sigma = K|L \in \mathcal{E}_{int}$ and $\mathrm{d}_\sigma = \mathrm{d}(x_K, \sigma)$ if $\sigma \in \mathcal{E}_{ext}$ with $\sigma \in \mathcal{E}_K$ and the transmissibility coefficient is defined by $\tau_\sigma = \mathrm{m}(\sigma)/\mathrm{d}_\sigma$, for all $\sigma \in \mathcal{E}$. We also denote by $\mathbf{n}_{K,\sigma}$ the normal to $\sigma \in \mathcal{E}_K$ outward K. We assume that the mesh satisfies the regularity constraint:

$$\exists \xi > 0 \text{ such that } \mathrm{d}(x_K, \sigma) \geq \xi\, \mathrm{d}_\sigma, \quad \forall K \in \mathcal{T}, \forall \sigma \in \mathcal{E}_K. \tag{2}$$

As a consequence, we obtain that

$$\sum_{\sigma\in\mathcal{E}_K} \mathrm{m}(\sigma)\mathrm{d}_\sigma \leq \frac{d}{\xi}\mathrm{m}(K) \quad \forall K \in \mathcal{T}. \tag{3}$$

The size of the mesh is defined by $h = \max\{\mathrm{diam}\,(K) \,:\, K \in \mathcal{T}\}$.

Let us define

$$f_K = \frac{1}{\mathrm{m}(K)}\int_K f, \quad b_K = \frac{1}{\mathrm{m}(K)}\int_K b \quad \forall K \in \mathcal{T},$$
$$U_{K,\sigma} = \frac{1}{\mathrm{m}(\sigma)}\int_\sigma \mathbf{U}\cdot\mathbf{n}_{K,\sigma}, \quad \forall K \in \mathcal{T}, \ \forall \sigma \in \mathcal{E}_K,$$
$$v_\sigma^D = \frac{1}{\mathrm{m}(\sigma)}\int_\sigma v^D, \quad \forall \sigma \in \mathcal{E}^D.$$

Given a Lipschitz-continuous function on $\mathbb{R}$ which satisfies

$$B(0) = 1, \quad B(s) > 0 \quad \text{and} \quad B(s) - B(-s) = -s \quad \forall s \in \mathbb{R}, \tag{4}$$

we consider the B-scheme defined by

$$\sum_{\sigma \in \mathscr{E}_K} \mathscr{F}_{K,\sigma} + \mathrm{m}(K) b_K v_K = \mathrm{m}(K) f_K, \quad \forall K \in \mathscr{T}, \tag{5}$$

where the numerical fluxes are defined by

$$\mathscr{F}_{K,\sigma} = \begin{cases} 0, \quad \forall K \in \mathscr{T}, \forall \sigma \in \mathscr{E}_K^N, \\ \tau_\sigma \Big(B(-U_{K,\sigma} \mathrm{d}_\sigma) v_K - B(U_{K,\sigma} \mathrm{d}_\sigma) v_{K,\sigma} \Big), \quad \forall K \in \mathscr{T}, \forall \sigma \in \mathscr{E}_K \setminus \mathscr{E}_K^N, \end{cases} \tag{6}$$

with the convention $v_{K,\sigma} = v_L$ if $\sigma = K|L$ and $v_{K,\sigma} = v_\sigma^D$ if $\sigma \in \mathscr{E}_K^D$. Let us recall that the upwind scheme corresponds to the case $B(s) = 1 + s^-$ (s^- is the negative part of s, while s^+ is its positive part) and the Scharfetter-Gummel scheme to the case $B(s) = s/(e^s - 1)$. They both satisfy (4). The centered scheme which corresponds to $B(s) = 1 - s/2$ does not satisfy the positivity assumption. It can however be used if $|U_{K,\sigma}|\mathrm{d}_\sigma \leq 2$ for all $K \in \mathscr{T}$ and $\sigma \in \mathscr{E}_K$. Thanks to the hypotheses (4), we notice that the numerical fluxes through the interior and Dirichlet boundary edges rewrite

$$\mathscr{F}_{K,\sigma} = \tau_\sigma B(|U_{K,\sigma}|\mathrm{d}_\sigma)(v_K - v_{K,\sigma}) + \mathrm{m}(\sigma) \left(U_{K,\sigma}^+ v_K - U_{K,\sigma}^- v_{K,\sigma} \right). \tag{7}$$

Main result. The scheme (5), (6) defines a linear system of equations $\mathbb{M}\mathbf{v} = \mathbf{S}$ whose unknown is $\mathbf{v} = (v_K)_{K \in \mathscr{T}}$; Since $\mathbb{M}$ is an M-matrix, one has existence and uniqueness of a solution to the scheme (see [1] for details). Moreover, if v^D and f are nonnegative functions, then $\mathbf{S}$ has nonnegative values and therefore $v_K \geq 0$ for all $K \in \mathscr{T}$. Our purpose is now to establish L^∞ bounds on $\mathbf{v}$ as stated in Theorem 1.

Theorem 1 *Assume that* $\mathbf{U} \in C(\bar{\Omega})^2$, $b \in L^\infty(\Omega)$ *with* $b \geq 0$ a.e., $f \in L^p(\Omega)$, *with* $p > d/2$, *and* $v^D \in L^\infty(\Gamma^D)$. *There exists non-negative constants* $\overline{M}$ (resp. $\underline{M}$) *depending only on* Ω, d, p, ξ, *the function* B, $\|\mathbf{U}\|_{L^\infty}$, $\|f^+\|_{L^p}$ *and* $\|(v^D)^+\|_{L^\infty}$ (resp. $\|f^-\|_{L^p}$ *and* $\|(v^D)^-\|_{L^\infty}$) *such that the solution* $\mathbf{v}$ *to the scheme* (5), (6) *satisfies*

$$-\underline{M} \leq v_K \leq \overline{M}, \quad \forall K \in \mathscr{T}.$$

The rest of this paper is dedicated to the proof of Theorem 1. It relies on a De Giorgi iteration method (see [7] and references therein). In Sect. 2, we start by studying a particular case where the data is normalized. Then, we give the proof of the theorem in Sect. 3. Let us mention that from the bounds of Theorem 1, it is possible to establish global-in-time L^∞ bounds for the corresponding evolution equation by using an entropy method (see [1, Theorem 2.7]). Moreover, as the problem is linear, these bounds are also sufficient to get convergence of the scheme in a weak sense by a compactness argument. Finally, let us mention that an interesting but difficult perspective would be the adaptation of the arguments in this note to numerical schemes on more general non-admissible meshes.

2 Study of a Particular Case

In this section, we consider the particular case where the source f is non-negative and the boundary data v^D is non-negative and bounded by 1.

Let us start with some notations. Given $m \geq 1$, we denote the m-th truncation threshold by

$$C_m = 2(1 - 2^{-m}) \,, \tag{8}$$

Then, we introduce the m-th energy

$$E_m(\mathbf{v}) = \sum_{\sigma \in \mathscr{E}_{int} \cup \mathscr{E}^D} \tau_\sigma \left[\log(1 + (v_{K,\sigma} - C_m)^+) - \log(1 + (v_K - C_m)^+)\right]^2 . \tag{9}$$

When there is no ambiguity we write $E_m = E_m(\mathbf{v})$. The first proposition is a fundamental estimate of the energy.

Proposition 1 *Assume that $f_K \geq 0$ for all $K \in \mathscr{T}$ and $v_\sigma^D \in [0, 1]$ for all $\sigma \in \mathscr{E}^D$. Then the solution $\mathbf{v}$ to (5), (6) satisfies $v_K \geq 0$ for all $K \in \mathscr{T}$ and one has for all $m \geq 1$ that*

$$E_m \;\leq\; \frac{4d}{\xi\, \beta_{\mathbf{U}}^2} \left(\|\mathbf{U}\|_{L^\infty}^2 + \|f\|_{L^p}\right) \Big(\sum_{\substack{K \in \mathscr{T} \\ v_K > C_m}} \mathrm{m}(K) \Big)^{1-\frac{1}{p}} . \tag{10}$$

where $\beta_{\mathbf{U}} := \inf_{x \in [-\|\mathbf{U}\|_{L^\infty}, \|\mathbf{U}\|_{L^\infty}]} B(\mathrm{diam}(\Omega)\, x)$ (because of (4), $\beta_{\mathbf{U}} \in (0, 1]$).

Proof Non-negativity of the solution follows from the M-matrix property of the scheme. In order to shorten some expressions hereafter, let us introduce $w_K^m = v_K - C_m$ for all $K \in \mathscr{T}$ and $w_\sigma^{m,D} = v_\sigma^D - C_m$ for all $\sigma \in \mathscr{E}^D$. Let us note that we identify $\mathbf{w}^m = (w_K^m)_{K \in \mathscr{T}}$ and the associate piecewise constant function. Therefore, we can write

$$\mathrm{m}(\{\mathbf{w}^m > 0\}) \;=\; \sum_{w_K^m > 0} \mathrm{m}(K) \;=\; \sum_{v_K > C_m} \mathrm{m}(K) \,.$$

First, observe that E_m is the discrete counterpart of

$$\int_\Omega \left|\nabla \log(1 + w^m)\right|^2 \mathbf{1}_{\{w^m > 0\}} = \int_\Omega \nabla w^m \cdot \frac{\nabla w^m}{(1 + w^m)^2} \mathbf{1}_{\{w^m > 0\}}, \;\; \text{with } w^m = v - C_m \,,$$

where $\mathbf{1}_A$ is the indicator function of A. Let us define $\varphi : s \mapsto s/(1 + s)\mathbf{1}_{\{s \geq 0\}}$, which satisfies $\varphi'(s) = 1/(1 + s)^2 \mathbf{1}_{\{s \geq 0\}}$ and let us introduce F_m another discrete counterpart of the preceding quantity

$$F_m = \sum_{\sigma\in\mathcal{E}_{int}\cup\mathcal{E}^D} \tau_\sigma \left((w^m_{K,\sigma})^+ - (w^m_K)^+\right)\left(\varphi(w^m_{K,\sigma}) - \varphi(w^m_K)\right).$$

It is clear that $E_m \le F_m$ for all $m \ge 1$, as for all $x, y \in \mathbb{R}$ we have

$$\left(\log(1+x^+) - \log(1+y^+)\right)^2 \le (x^+ - y^+)\left(\varphi(x) - \varphi(y)\right).$$

Let us now multiply the scheme (5) by $\varphi(w^m_K)$ and sum over $K \in \mathcal{T}$. Due to the non-negativity of b and $\mathbf{v}$, we obtain, after a discrete integration by parts,

$$\sum_{\sigma\in\mathcal{E}_{int}\cup\mathcal{E}^D} \mathscr{F}_{K,\sigma}(\varphi(w^m_K) - \varphi(w^m_{K,\sigma})) \le \sum_{K\in\mathcal{T}} \mathrm{m}(K) f_K \varphi(w^m_K).$$

Using that φ is bounded by 1 and vanishes on $\mathbb{R}_-$, we deduce that

$$\sum_{\sigma\in\mathcal{E}_{int}\cup\mathcal{E}^D} \mathscr{F}_{K,\sigma}(\varphi(w^m_K) - \varphi(w^m_{K,\sigma})) \le \|f\|_{L^p}\, \mathrm{m}(\{\mathbf{w}^m > 0\})^{1-\frac{1}{p}}. \tag{11}$$

We focus now on the left-hand-side of (11). Due to (7) and the definition of w^m_K, we can rewrite $\mathscr{F}_{K,\sigma}$ as

$$\mathscr{F}_{K,\sigma} = \tau_\sigma B(|U_{K,\sigma}|\mathrm{d}_\sigma)(w^m_K - w^m_{K,\sigma}) + \mathrm{m}(\sigma)\left(U^+_{K,\sigma}(w^m_K + C_m) - U^-_{K,\sigma}(w^m_{K,\sigma} + C_m)\right).$$

Observe that since φ is a non-decreasing function, one has

$$(x-y)\left(\varphi(x) - \varphi(y)\right) \ge (x^+ - y^+)(\varphi(x) - \varphi(y)), \quad \forall x, y \in \mathbb{R}.$$

Therefore, using the definition of $\beta_{\mathbf{U}}$ we obtain that

$$\sum_{\sigma\in\mathcal{E}_{int}\cup\mathcal{E}^D} \mathscr{F}_{K,\sigma}(\varphi(w^m_K) - \varphi(w^m_{K,\sigma})) \ge \beta_{\mathbf{U}} F_m - G_m, \tag{12}$$

with

$$G_m = -\sum_{\sigma\in\mathcal{E}_{int}\cup\mathcal{E}^D} \mathrm{m}(\sigma)\left(U^+_{K,\sigma}(w^m_K + C_m) - U^-_{K,\sigma}(w^m_{K,\sigma} + C_m)\right)(\varphi(w^m_K) - \varphi(w^m_{K,\sigma})).$$

For an interior edge, w^m_K and $w^m_{K,\sigma}$ play a symmetric role in the preceding sum. As $w^{m,D}_\sigma \le 0$ for all $\sigma \in \mathcal{E}^D$ and φ vanishes on $\mathbb{R}_-$, we can always assume that $w^m_K \ge w^m_{K,\sigma}$ and an edge has a contribution in the sum if at least $w^m_K > 0$. Then, under these assumptions one has

$$\begin{aligned} -\,\mathrm{m}(\sigma)\left(U^+_{K,\sigma}(w^m_K + C_m) - U^-_{K,\sigma}(w^m_{K,\sigma} + C_m)\right)(\varphi(w^m_K) - \varphi(w^m_{K,\sigma})) \\ \le \|\mathbf{U}\|_{L^\infty} \mathrm{m}(\sigma)(w^m_{K,\sigma} + C_m)(\varphi(w^m_K) - \varphi(w^m_{K,\sigma})). \end{aligned}$$

But, $w^m_{K,\sigma} + C_m \leq 2(1 + (w^m_{K,\sigma})^+)$ and applying the definition of φ, we get

$$(w^m_{K,\sigma} + C_m)(\varphi(w^m_K) - \varphi(w^m_{K,\sigma})) \ \leq \ 2\frac{(w^m_K)^+ - (w^m_{K,\sigma})^+}{1 + (w^m_K)^+}\,.$$

Therefore, using that $w^m_K \geq w^m_{K,\sigma}$, one has

$$G_m \leq 2\|\mathbf{U}\|_{L^\infty} \sum_{\sigma \in \mathscr{E}_{int} \cup \mathscr{E}^D} \mathrm{m}(\sigma) \frac{|(w^m_K)^+ - (w^m_{K,\sigma})^+|}{\sqrt{1 + (w^m_K)^+}\sqrt{1 + (w^m_{K,\sigma})^+}}.$$

We apply now Cauchy-Schwarz inequality in order to get

$$G_m \leq 2\|\mathbf{U}\|_{L^\infty} (F_m)^{1/2} \left(\sum_{\sigma \in \mathscr{E}^{sp}} \mathrm{m}(\sigma)\mathrm{d}_\sigma \right)^{1/2}, \tag{13}$$

where $\mathscr{E}^{sp}$ is the set of interior and Dirichlet boundary edges on which $(w^m_K)^+ - (w^m_{K,\sigma})^+ \neq 0$. It appears that, due to (3),

$$\sum_{\sigma \in \mathscr{E}^{sp}} \mathrm{m}(\sigma)\mathrm{d}_\sigma \leq \sum_{K \in \mathscr{T}; w^m_K > 0} \left(\sum_{\sigma \in \mathscr{E}_{K,int} \cup \mathscr{E}^D_K} \mathrm{m}(\sigma)\mathrm{d}_\sigma \right) \leq \frac{d}{\xi}\mathrm{m}(\{\mathbf{w}^m > 0\}). \tag{14}$$

We deduce from (11)–(14) and Young's inequality that

$$\begin{aligned} \beta_{\mathbf{U}} F_m &\leq 2\|\mathbf{U}\|_{L^\infty}(F_m)^{1/2}(\frac{d}{\xi}\mathrm{m}(\{\mathbf{w}^m > 0\}))^{1/2} + \|f\|_{L^p}\mathrm{m}(\{\mathbf{w}^m > 0\})^{1-\frac{1}{p}}, \\ &\leq \frac{\beta_{\mathbf{U}} F_m}{2} + \frac{2d}{\beta_{\mathbf{U}}\xi}\big(\|\mathbf{U}\|^2_{L^\infty} + \|f\|_{L^p}\big)\mathrm{m}(\{\mathbf{w}^m > 0\})^{1-\frac{1}{p}} \end{aligned}$$

which yields (10) using that $E_m \leq F_m$, $\beta_{\mathbf{U}} \leq 1$ and $\mathrm{m}(\Omega) = 1$.

Before stating the main result of the section, we need a technical lemma.

Lemma 1 *Let $(u_n)_{n \in \mathbb{N}}$ be a sequence of non-negative real numbers and let $K, \rho > 0$ and $\alpha > 1$. Then if $u_{n+1} \ \leq \ K\,\rho^n\,u_n^\alpha$ for all $n \in \mathbb{N}$ then one has*

$$0 \leq u_n \ \leq \ \left(u_0\,\rho^{\frac{1}{(\alpha-1)^2}}\,K^{\frac{1}{\alpha-1}}\right)^{\alpha^n}\ \rho^{-\frac{n(\alpha-1)+1}{(\alpha-1)^2}}\,K^{-\frac{1}{\alpha-1}}$$

for all $n \in \mathbb{N}$. In particular, if $u_0 < \rho^{-\frac{1}{(\alpha-1)^2}}\,K^{-\frac{1}{\alpha-1}}$, then $\lim u_n = 0$.

Proof Just observe that the sequence $v_n = u_n\,\rho^{\frac{n(\alpha-1)+1}{(\alpha-1)^2}}\,K^{\frac{1}{\alpha-1}}$ satisfies $0 \leq v_{n+1} \leq v_n^\alpha$ for all $n \geq 0$ which directly yields the result. Observe that the bound is optimal.

Proposition 2 *Assume that $f_K \geq 0$ for all $K \in \mathcal{T}$ and $v_\sigma^D \in [0,1]$ for all $\sigma \in \mathcal{E}^D$. Then, there exist $\gamma > 0$ depending on d and p, and $\eta > 0$ depending additionally on Ω and ξ such that one has*

$$E_1 \leq \eta \left(\frac{\beta_{\mathbf{U}}^2}{\|\mathbf{U}\|_{L^\infty}^2 + \|f\|_{L^p}} \right)^\gamma \quad \Rightarrow \quad (v_K \leq 2, \ \forall K \in \mathcal{T}) \,. \tag{15}$$

Proof The proof consists in establishing an induction property on E_m which guarantees that if E_1 is small enough then $\lim E_m = 0$. Then, as $\lim C_m = 2$ and thanks to the discrete Poincaré inequality, we deduce that

$$\sum_{K \in \mathcal{T}} \mathrm{m}(K) \left(\log(1 + (v_K - 2)^+) \right)^2 = 0,$$

which implies $v_K \leq 2$ for all $K \in \mathcal{T}$.

First observe that as $C_m = C_{m-1} + 2^{-m+1}$, for any $q > 0$ we have:

$$\mathbf{1}_{\{\mathbf{w}^m > 0\}} \leq \frac{\left(\log(1 + (\mathbf{w}^{m-1})^+)\right)^q}{(\log(1 + 2^{-m+1}))^q} \mathbf{1}_{\{\mathbf{w}^{m-1} > 0\}}, \tag{16}$$

and thus

$$\mathrm{m}(\{\mathbf{w}^m > 0\}) \leq \frac{1}{(\log(1 + 2^{-m+1}))^q} \sum_{K \in \mathcal{T}} \mathrm{m}(K) \left(\log(1 + (w_K^{m-1})^+) \right)^q .$$

Since $p > d/2$ there exists $q \in (2p/(p-1), \, 2d/(d-2))$. For such q, the discrete Poincaré-Sobolev inequality implies

$$\sum_{K \in \mathcal{T}} \mathrm{m}(K) \left(\log(1 + (w_K^{m-1})^+) \right)^q \ \leq \ C_{\Omega,d,q,\xi} \, E_{m-1}^{\frac{q}{2}} .$$

Then, noticing that for $x \in [0,1]$, $(\log(1+x))^q \geq (\log 2)^q x^q$, we obtain

$$\mathrm{m}(\{\mathbf{w}^m > 0\})^{1-\frac{1}{p}} \leq C_{\Omega,d,q,\xi} \, 2^{\frac{(m-1)q(p-1)}{p}} E_{m-1}^{\frac{q(p-1)}{2p}} . \tag{17}$$

We deduce from (10) and (17) that

$$E_m \ \leq \ C_{\Omega,d,q,\xi} \, \frac{\|\mathbf{U}\|_{L^\infty}^2 + \|f\|_{L^p}}{\beta_{\mathbf{U}}^2} \, 2^{\frac{(m-1)q(p-1)}{p}} E_{m-1}^{\frac{q(p-1)}{2p}} .$$

Thus the sequence $(E_m)_{m \geq 0}$ satisfies the hypothesis of Lemma 1 with $\alpha = q(p-1)/(2p) > 1$ and K proportional to $(\|\mathbf{U}\|_{L^\infty}^2 + \|f\|_{L^p})/\beta_{\mathbf{U}}^2$. We deduce the upper bound for E_1 (with $\gamma = 1/(\alpha - 1)$) under which $\lim E_m = 0$.

3 Proof of Theorem 1

First observe that if one replaces the data f and v^D by either f^+ and $(v^D)^+$, or f^- and $(v^D)^-$, in the scheme (5), (6), then the corresponding solutions, say respectively $\mathbf{P} = (P_K)_{K\in\mathscr{T}}$ and $\mathbf{N} = (N_K)_{K\in\mathscr{T}}$, are non-negative and such that $\mathbf{v} = \mathbf{P} - \mathbf{N}$ is the solution to (5), (6) in the original framework.

From there let us show that there is $\overline{M} > V_+^D := \max(\|(v^D)^+\|_{L^\infty}, 1)$ such that for all $K \in \mathscr{T}$ one has $0 \leq P_K \leq \overline{M}$. The bound for $\mathbf{N}$, which is denoted by $\underline{M}$, can be obtained in the same way.

Let $M > V_+^D$. First observe that $\mathbf{P}^M := \mathbf{P}/M$ satisfies the scheme (5), (6) where the source term and boundary data have been replaced by f^+/M and $(v^D)^+/M$ respectively. Moreover, one can apply Proposition 1, which yields

$$E_1(\mathbf{P}^M) \leq \frac{4d}{\xi\beta_{\mathbf{U}}^2}\left(\|\mathbf{U}\|_{L^\infty}^2 + \frac{\|f^+\|_{L^p}}{M}\right)\mathrm{m}(\{\mathbf{P}^M > 1\})^{1-\frac{1}{p}}\,. \tag{18}$$

Now observe that $\mathbf{P} = M\,\mathbf{P}^M = V_+^D\,\mathbf{P}^{V_+^D}$. Therefore, for $r = 2p/(p-1)$ one has

$$\begin{aligned}
E_1(\mathbf{P}^M) &\leq \frac{4d}{\xi\beta_{\mathbf{U}}^2}\Big(\|\mathbf{U}\|_{L^\infty}^2\,\mathrm{m}(\{\mathbf{P}^{V_+^D} > M/V_+^D\})^{1-\frac{1}{p}} + \frac{\|f^+\|_{L^p}}{M}\Big)\\
&\leq \frac{4d}{\xi\beta_{\mathbf{U}}^2}\Big[\|\mathbf{U}\|_{L^\infty}^2\Big(\sum_{K\in\mathscr{T}}\mathrm{m}(K)\,\frac{\log(1+(P_K^{V_+^D}-1)^+)^r}{\log(1+(M/V_+^D-1)^+)^r}\Big)^{1-\frac{1}{p}} + \frac{\|f^+\|_{L^p}}{M}\Big]\\
&\leq \frac{C_{\Omega,d,p,\xi}}{\beta_{\mathbf{U}}^2}\Big(\|\mathbf{U}\|_{L^\infty}^2\,\frac{E_1(\mathbf{P}^{V_+^D})}{\log(M/V_+^D)^2} + \frac{\|f^+\|_{L^p}}{M}\Big)\,,
\end{aligned}$$

where we used an argument similar to (16) in the second inequality and a discrete Poincaré Sobolev inequality in the third one. Then, (18) with $M = V_+^D$ yields

$$E_1(\mathbf{P}^M) \leq \frac{C_{\Omega,d,p,\xi}}{\beta_{\mathbf{U}}^4}\Big(\frac{\|\mathbf{U}\|_{L^\infty}^2}{\log(M/V_+^D)^2}\left(\|\mathbf{U}\|_{L^\infty}^2 + \frac{\|f^+\|_{L^p}}{V_+^D}\right) + \frac{\|f^+\|_{L^p}}{M}\Big)\,.$$

It remains to choose M such that the right-hand side is bounded by

$$\eta\left(\frac{\beta_{\mathbf{U}}^2}{\|\mathbf{U}\|_{L^\infty}^2 + M^{-1}\|f^+\|_{L^p}}\right)^\gamma\,.$$

It is satisfied for M large enough, which permits to define $\overline{M}$. Observe that if $v_+^D = 0$ ($V_+^D = 1$) and $\mathbf{U} = 0$, one can take $\overline{M} = \widetilde{C}_{\Omega,\xi,d,p}\|f^+\|_{L^p}$ as expected.

Acknowledgements The authors thank the Labex CEMPI (ANR-11-LABX-0007-01) and the ANR MOHYCON (ANR-17-CE40-0027-01) for their support. They also want to thank Alexis F. Vasseur for fruitful exchanges on the subject.

References

1. Chainais-Hillairet, C., Herda, M.: Large-time behaviour of a family of finite volume schemes for boundary-driven convection-diffusion equations. IMA J. Numer. Anal. (2019). https://doi.org/10.1093/imanum/drz037
2. Droniou, J.: Non-coercive linear elliptic problems. Potential Anal. **17**(2), 181–203 (2002)
3. Droniou, J.: Error estimates for the convergence of a finite volume discretization of convection-diffusion equations. J. Numer. Math. **11**(1), 1–32 (2003). https://doi.org/10.1163/156939503322004873
4. Droniou, J., Gallouët, T.: Finite volume methods for convection-diffusion equations with right-hand side in H^{-1}. M2AN Math. Model. Numer. Anal. **36**(4), 705–724 (2002)
5. Droniou, J., Gallouët, T., Herbin, R.: A finite volume scheme for a noncoercive elliptic equation with measure data. SIAM J. Numer. Anal. **41**(6), 1997–2031 (2003)
6. Eymard, R., Gallouët, T., Herbin, R.: Finite volume methods. In: Handbook of Numerical Analysis, vol. VII, pp. 713–1020. North-Holland, Amsterdam (2000)
7. Vasseur, A.F.: The De Giorgi method for elliptic and parabolic equations and some applications. In: Lectures on the Analysis of Nonlinear Partial Differential Equations, vol. 4 (2016)

On Four Numerical Schemes for a Unipolar Degenerate Drift-Diffusion Model

Clément Cancès, Claire Chainais Hillairet, Jürgen Fuhrmann, and Benoît Gaudeul

Abstract We consider a unipolar degenerate drift-diffusion system where the relation between the concentration of the charged species c and the chemical potential h is $h(c) = \log \frac{c}{1-c}$. For four different finite volume schemes based on four different formulations of the fluxes of the problem, we discuss stability and existence results. For two of them, we report a convergence proof. Numerical experiments illustrate the behaviour of the different schemes.

Keywords Finite volume methods · Drift-diffusion problems · Energy methods

MSC (2010) 65M08 · 65N08 · 65Z05

1 Introduction

The transport of a charged species with density c in the presence of a fixed or moving countercharge and a self-consistent electric field, deriving from an electrostatic potential Φ, can be described by the non-dimensionalized system of equations

C. Cancès · C. C. Hillairet · B. Gaudeul
Inria - Laboratoire Paul Painlevé, University of Lille, CNRS,
UMR 8524, 59000 Lille, France
e-mail: clement.cances@inria.fr

C. C. Hillairet
e-mail: claire.chainais@univ-lille.fr

B. Gaudeul
e-mail: benoit.gaudeul@univ-lille.fr

J. Fuhrmann (✉)
Weierstrass Institute (WIAS), Mohrenstr. 39, 10117 Berlin, Germany
e-mail: juergen.fuhrmann@wias-berlin.de

R. Klöfkorn et al. (eds.), *Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples*, Springer Proceedings in Mathematics & Statistics 323,
https://doi.org/10.1007/978-3-030-43651-3_13

$$\partial_t c + \operatorname{div}(\mathbf{J}) = 0 \quad \text{in } (0,T)\times\Omega, \tag{1}$$

$$\mathbf{J} = -c\nabla\left(h(c)+\Phi\right) \quad \text{in } (0,T)\times\Omega, \tag{2}$$

where $h(c) = \log\left(\frac{c}{1-c}\right)$ is the chemical potential. The electrostatic potential Φ is related to space charge density via the Poisson equation

$$-\Delta\Phi = c + c^{\text{dop}} \quad \text{in } (0,T)\times\Omega. \tag{3}$$

In (3), c^{dop} describes the doping profile of the media. Such models occur in applications ranging from organic semiconductors [5], high-temperature fuel cells [13] or simplified models of ionic liquids [8]. Because of the singularity of h near 1, the density c remains in the interval $(0,1)$. We consider the evolution in a connected bounded open domain Ω of $\mathbb{R}^d$ $(d \le 3)$ with polyhedral and Lipschitz continuous boundary $\partial\Omega$ during a finite but arbitrary time $T > 0$. The doping profile c^{dop} is assumed to be constant w.r.t. time and to be bounded, i.e., $c^{\text{dop}} \in L^\infty(\Omega)$. The system is supplemented with the prescription of the initial concentration

$$c_{|_{t=0}} = c^0 \in L^\infty(\Omega) \quad \text{with} \quad 0 \le c^0 \le 1 \quad \text{and} \quad 0 < \overline{c} = \fint_\Omega c^0 \mathrm{d}\mathbf{x} < 1, \tag{4}$$

of no-flux boundary conditions for the concentration

$$\mathbf{J}\cdot\mathbf{n} = 0 \quad \text{on } (0,T)\times\partial\Omega. \tag{5}$$

For the electrostatic potential, inhomogeneous Dirichlet boundary conditions are imposed on a subset Γ_D of positive measure of $\partial\Omega$, whereas homogeneous Neumann boundary conditions are imposed on $\Gamma_N = \partial\Omega\setminus\Gamma_D$:

$$\Phi = \Phi^D \quad \text{on } (0,T)\times\Gamma_D, \qquad \nabla\Phi\cdot\mathbf{n} = 0 \quad \text{on } (0,T)\times\Gamma_N. \tag{6}$$

We assume that Φ^D is defined in the whole domain, with $\Phi^D \in H^1(\Omega)\cap L^\infty(\Omega)$.

In [3], we studied and compared several Finite Volume schemes for the system (1)–(6). They are based on various reformulations of the flux $\mathbf{J}$ using the excess chemical potential $\nu(c) = h(c) - \log(c) = -\log(1-c)$, the activity and the inverse of the activity coefficient respectively defined by $a(c) = e^{h(c)} = \frac{c}{1-c}$, and $\beta(c) = \frac{c}{a(c)} = 1-c$, or the diffusion enhancement $r(c) = -\log(1-c)$ satisfying $r'(c) = ch'(c)$. The flux $\mathbf{J}$, initially defined by (2), can be alternatively rewritten as

$$\begin{aligned} \mathbf{J} &= -\nabla c - c\nabla\left(\Phi + \nu(c)\right), && (7)\\ &= -\beta(c)(\nabla a(c) + a(c)\nabla\Phi), && (8)\\ &= -r'(c)\nabla c - c\nabla\Phi. && (9) \end{aligned}$$

Let us notice that, even $\nu(c) = r(c)$ for our specific choice of $h(c)$, the excess chemical potential and the diffusion enhancement arising respectively in (7) and (9) have a different physical sense so that we keep different notations.

Each formulation (2), (7), (8) and (9) leads to a different scheme that we compared from a numerical analysis point of view. Notice that the flux $\mathbf{J}$ can also be expressed as $\mathbf{J} = -\nabla r(c) - c\nabla\Phi$. This last formulation is used to define a proper notion of weak solution to (1)–(6). In order to state this definition, we introduce the vector space $\mathscr{H}_{\Gamma^D} = \{f \in H^1(\Omega), f_{|\Gamma_D} = 0\}$ and the space-time cylinder $Q_T = (0,T) \times \Omega$.

Definition 1 A couple (c, Φ) is a *weak solution of* (1)–(6) if

- $c \in L^\infty(Q_T; [0,1])$ with $r(c) \in L^2((0,T); H^1(\Omega))$, and $\Phi - \Phi^D \in L^\infty((0,T), \mathscr{H}_{\Gamma^D})$;
- for all $\varphi \in C_c^\infty([0,T) \times \overline{\Omega})$, there holds

$$\iint_{Q_T} c\partial_t\varphi \mathrm{d}\mathbf{x}\mathrm{d}t + \int_\Omega c^0\varphi(0,\cdot)\mathrm{d}\mathbf{x} - \iint_{Q_T} (\nabla r(c) + c\nabla\Phi)\cdot\nabla\varphi \mathrm{d}\mathbf{x}\mathrm{d}t = 0; \quad (10)$$

- for all $\psi \in \mathscr{H}_{\Gamma^D}$ and almost all $t \in (0,T)$, there holds

$$\int_\Omega \nabla\Phi(t,\mathbf{x})\cdot\nabla\psi(\mathbf{x})\mathrm{d}\mathbf{x} = \int_\Omega (c(t,\mathbf{x}) + c^{\mathrm{dop}}(\mathbf{x}))\psi(\mathbf{x})\mathrm{d}\mathbf{x}. \quad (11)$$

We shortly discuss the gradient flow structure of the system (1)–(6). Define the mixing entropy density

$$H(c) = c\log(c) + (1-c)\log(1-c),$$

which is an antiderivative of h, then the electrochemical energy is given by

$$E(c,\Phi) = \int_\Omega \left\{ H(c) + \frac{1}{2}|\nabla\Phi|^2 \right\} \mathrm{d}\mathbf{x} - \int_{\Gamma_D} \Phi^D\nabla\Phi\cdot\mathbf{n}\mathrm{d}\boldsymbol{\gamma}. \quad (12)$$

The electrochemical energy is a Lyapunov functional. Moreover, the dissipation rate for the energy is explicitly given.

Proposition 1 *Let* (c,Φ) *be a smooth solution to (1)–(6), with c bounded away from* 0 *and* 1*, then*

$$\frac{\mathrm{d}}{\mathrm{d}t}E(c,\Phi) + \int_\Omega c\,|\nabla(h(c)+\Phi)|^2\,\mathrm{d}\mathbf{x} = 0.$$

2 TPFA Finite Volume Approximations

For the space discretization, we use the standard notation of an admissible finite volume mesh $\left(\mathscr{T}, \mathscr{E}, (\mathbf{x}_K)_{K\in\mathscr{T}}\right)$, see [3]. Control volumes are denoted by $K \in \mathscr{T}$ with respective measures m_K, whereas edges are denoted by $\sigma \in \mathscr{E}$, their $(d-1)$-dimensional measure being denoted by m_σ. Since our method relies on a two-point flux approximation, we suppose that the mesh satisfies the classical orthogonality condition [6, Chapter 9]. For the time discretization, we consider an increasing finite family of times $0 = t_0 < t_1 < \cdots < t_N = T$. We denote by $\Delta t_n = t_n - t_{n-1}$ for $1 \le n \le N$, by $\boldsymbol{\Delta t} = (\Delta t_n)_{1\le n\le N}$, and by $\overline{\boldsymbol{\Delta t}} = \max_{1\le n\le N} \Delta t_n$.

The initial data c_0 and the doping profile c^{dop} are respectively discretized into $\left(c_K^0\right)_{K\in\mathscr{T}}, \left(c_K^{\text{dop}}\right)_{K\in\mathscr{T}} \in \mathbb{R}^{\mathscr{T}}$ by setting

$$c_K^0 = \frac{1}{m_K}\int_K c^0(\mathbf{x})\mathrm{d}\mathbf{x}, \quad c_K^{\text{dop}} = \frac{1}{m_K}\int_K c^{\text{dop}}(\mathbf{x})\mathrm{d}\mathbf{x}, \qquad \forall K \in \mathscr{T}, \tag{13}$$

Assume that $c^{n-1} = \left(c_K^{n-1}\right)_{K\in\mathscr{T}} \in [0,1]^{\mathscr{T}}$ is given for some $n > 0$. We define how to compute $(c^n, \Phi^n) = \left(c_K^n, \Phi_K^n\right)_{K\in\mathscr{T}}$. For all $K \in \mathscr{T}$ and all $\sigma \in \mathscr{E}_K = \mathscr{E}_{\text{int}} \cup \mathscr{E}_{\text{ext}}$, the set of interior and exterior control volume facets, we define the mirror values $c_{K\sigma}^n$ and $\Phi_{K\sigma}^n$ of c_K^n and Φ_K^n respectively across σ by setting

$$c_{K\sigma}^n = \begin{cases} c_L^n & \text{if } \sigma = K|L \in \mathscr{E}_{\text{int}}, \\ c_K^n & \text{if } \sigma \in \mathscr{E}_{\text{ext}}, \end{cases} \qquad \Phi_{K\sigma}^n = \begin{cases} \Phi_L^n & \text{if } \sigma = K|L \in \mathscr{E}_{\text{int}}, \\ \Phi_K^n & \text{if } \sigma \in \mathscr{E}^N, \\ \Phi_\sigma^n = \frac{1}{m_\sigma}\int_\sigma \Phi^D \mathrm{d}\boldsymbol{\gamma} & \text{if } \sigma \in \mathscr{E}^D. \end{cases}$$

For $\sigma \in \mathscr{E}$, we set $d_\sigma = |\mathbf{x}_K - \mathbf{x}_L|$ if $\sigma = K|L \in \mathscr{E}_{\text{int}}$, $d_\sigma = |\mathbf{x}_K - \mathbf{x}_\sigma|$ if $\sigma \in \mathscr{E}_{\text{ext}}$, and $\tau_\sigma = \frac{m_\sigma}{d_\sigma}$. Given $u = (u_K)_{K\in\mathscr{T}} \in \mathbb{R}^{\mathscr{T}}$, we define the oriented and absolute jumps of u across $\sigma \in \mathscr{E}_K$ by $D_{K\sigma}u = u_{K\sigma} - u_K$, and $D_\sigma u = |D_{K\sigma}u|$.

All the four schemes we consider are based on a backward Euler scheme for the time discretization and a TPFA finite volume scheme for the space discretization. They are written as follows:

$$-\sum_{\sigma\in\mathscr{E}_K} \tau_\sigma D_{K\sigma}\Phi^n = m_K\left(c_K^n + c_K^{\text{dop}}\right), \qquad \forall K \in \mathscr{T}, \tag{14a}$$

$$m_K\frac{c_K^n - c_K^{n-1}}{\Delta t_n} + \sum_{\sigma\in\mathscr{E}_{K,\text{int}}} F_{K\sigma}^n = 0, \qquad \forall K \in \mathscr{T}. \tag{14b}$$

To close the system (14), it remains to define the numerical fluxes $F_{K\sigma}^n$. Due to the no-flux boundary condition, we only have to define the inner fluxes. They are defined with a function $\mathscr{F}$ of the primary unknowns $(c_K^n, c_L^n, \Phi_K^n, \Phi_L^n)$:

$$F_{K\sigma}^n = \tau_\sigma \mathscr{F}(c_K^n, c_L^n, \Phi_K^n, \Phi_L^n), \quad \forall K \in \mathscr{T}, \forall \sigma = K|L. \tag{15}$$

The different schemes considered in this contribution correspond to different choices of $\mathscr{F}$. All of them verify $\mathscr{F}(c_K, c_L, \Phi_K, \Phi_L) = -\mathscr{F}(c_L, c_K, \Phi_L, \Phi_K)$, so that the numerical fluxes are locally conservative. Three of the four schemes are extensions of the Scharfetter-Gummel scheme [12] and feature the Bernoulli function $B(u) = \frac{u}{e^u - 1}$.

The **centred flux** is derived from (2), which suggests the following definition:

$$\mathscr{F}(c_K, c_L, \Phi_K, \Phi_L) = -\frac{c_K + c_L}{2} D_{K\sigma}\left(h(c) + \Phi\right). \tag{C}$$

The associate flux can be seen as a particular case in the TPFA context of the fluxes introduced in [4]. This scheme is not based on the Scharfetter-Gummel scheme.

The **Sedan** flux is named after the SEDAN III simulator [14]. Formula (7) for the flux $\mathbf{J}$ suggests to use a classical Scharfetter-Gummel scheme, but for a modified potential $\Phi + \nu(c)$ instead of only Φ, leading to

$$\mathscr{F}(c_K, c_L, \Phi_K, \Phi_L) = B\Big(D_{K\sigma}(\Phi + \nu(c))\Big)c_K - B\Big(-D_{K\sigma}(\Phi + \nu(c))\Big)c_L. \tag{S}$$

The **activity based flux** is a restriction of the flux introduced in [7]. It relies on the expression (8). With frozen $\beta(c)$, the flux $\mathbf{J}$ is linear w.r.t. $a(c)$. This suggests choosing a particular average for $\beta(c)$—here the arithmetic mean—and applying the Scharfetter-Gummel scheme to approximate $-\nabla a(c) - a(c)\nabla\Phi$, yielding

$$\mathscr{F}(c_K, c_L, \Phi_K, \Phi_L) = \frac{\beta(c_K) + \beta(c_L)}{2}\Big\{B(D_{K\sigma}\Phi)a(c_K) - B(-D_{K\sigma}\Phi)a(c_L)\Big\}. \tag{AB}$$

Formula (9) for the flux $\mathbf{J}$ suggests that, with introducing a variable diffusion coefficient approximating the $r'(c)$ per face, one can use the Scharfetter-Gummel scheme. Following [1], the approximation $\mathfrak{d}r(c_K, c_L)$ of $r'(c)$ is defined as

$$\mathfrak{d}r(c_K, c_L) = \begin{cases} \dfrac{h(c_K) - h(c_L)}{\log(c_K) - \log(c_L)} & \text{if } c_K \neq c_L, \\ r'(c_K) & \text{if } c_K = c_L. \end{cases}$$

This leads to the following definition of the **Bessemoulin-Chatard flux** [1]:

$$\mathscr{F}(c_K, c_L, \Phi_K, \Phi_L) = \mathfrak{d}r(c_K, c_L)\left\{B\left(\frac{D_{K\sigma}\Phi}{\mathfrak{d}r(c_K, c_L)}\right)c_K - B\left(-\frac{D_{K\sigma}\Phi}{\mathfrak{d}r(c_K, c_L)}\right)c_L\right\}. \tag{BC}$$

2.1 Main Results

The energy decay was one of the key properties of the continuous model, cf. Proposition 1. This property is transposed to the discrete setting by all the four discretizations we have considered. The discrete energy functional $E_{\mathscr{T}}$ has to be thought of as a discrete counterpart of the continuous energy functional E, cf. (12). It is defined by:

$$E_{\mathscr{T}}(c^n, \varPhi^n) = \sum_{K \in \mathscr{T}} m_K H(c_K^n) + \frac{1}{2} \sum_{\sigma \in \mathscr{E}} \tau_\sigma \left(D_\sigma \varPhi^n\right)^2 - \sum_{K \in \mathscr{T}} \sum_{\sigma \in \mathscr{E}^D \cap \mathscr{E}_K} \tau_\sigma \varPhi_\sigma^D D_{K\sigma} \varPhi^n.$$

Our first result focuses on the four schemes on a fixed mesh. It states that the nonlinear system corresponding to each scheme admits a solution which preserves the physical bounds on the concentrations and the decay of the energy.

Theorem 1 *Let $(\mathscr{T}, \mathscr{E}, (\mathbf{x}_K)_{K \in \mathscr{T}})$ be an admissible mesh and let c^0 be defined by (13). Then, for all $1 \le n \le N$, the nonlinear system of equations (14)–(15), supplemented either with (C), (S), (AB), or (BC), has a solution $(c^n, \varPhi^n) \in [0, 1]^{\mathscr{T}} \times \mathbb{R}^{\mathscr{T}}$. Moreover, the solution to the scheme satisfies, for all $1 \le n \le N$,*

$$E_{\mathscr{T}}(c^{n,}, \varPhi^n) \le E_{\mathscr{T}}(c^{n-1}, \varPhi^{n-1}) \quad \textit{and} \quad 0 < c_K^n < 1, \quad \forall K \in \mathscr{T}.$$

Once a discrete solution to the scheme $(c^n, \varPhi^n)_{1 \le n \le N}$ at hand, we can define an approximate solution $(c_{\mathscr{T},\boldsymbol{\Delta t}}, \varPhi_{\mathscr{T},\boldsymbol{\Delta t}})$. It is the piecewise constant function defined almost everywhere by

$$c_{\mathscr{T},\boldsymbol{\Delta t}}(t, \mathbf{x}) = c_K^n, \quad \varPhi_{\mathscr{T},\boldsymbol{\Delta t}}(t, \mathbf{x}) = \varPhi_K^n \quad \text{if } (t, \mathbf{x}) \in (t_{n-1}, t_n] \times K.$$

Let $\left(\mathscr{T}_m, \mathscr{E}_m, (\mathbf{x}_K)_{K \in \mathscr{T}_m}\right)_{m \ge 1}$ be a sequence of admissible meshes such that $h_{\mathscr{T}_m}$, $\overline{\boldsymbol{\Delta t}}_m \underset{m \to \infty}{\longrightarrow} 0$ while the mesh regularity remains bounded (see [3] for the definition of the regularity of the mesh). A natural question is the convergence of the associated sequence of approximate solutions $(c_{\mathscr{T}_m,\boldsymbol{\Delta t}_m}, \varPhi_{\mathscr{T}_m,\boldsymbol{\Delta t}_m})_{m \ge 1}$ towards a weak solution to the continuous problem. The convergence result is stated in Theorem 2, only for the centred scheme and the Sedan scheme. The proof is detailed in [3]. It is based on compactness arguments. As far as we know, there is no uniqueness result for the weak solutions, hence the convergence only holds up to a subsequence.

Theorem 2 *For the centred scheme (inner fluxes defined by (15) and (C)) and the Sedan scheme (inner fluxes defined by (15) and (S)), a sequence of approximate solutions $(c_{\mathscr{T}_m,\boldsymbol{\Delta t}_m}, \varPhi_{\mathscr{T}_m,\boldsymbol{\Delta t}_m})_{m \ge 1}$ satisfies, up to a subsequence,*

$$c_{\mathscr{T}_m,\boldsymbol{\Delta t}_m} \underset{m \to \infty}{\longrightarrow} c \quad \textit{a.e. in } Q_T, \qquad \varPhi_{\mathscr{T}_m,\boldsymbol{\Delta t}_m} \underset{m \to \infty}{\longrightarrow} \varPhi \quad \textit{in } L^2(Q_T), \tag{16}$$

where $(c, \varPhi)$ is a weak solution to (1)–(6) in the sense of Definition 1.

3 A Numerical Example

The presented numerical example and those in [3] have been implemented in the Julia language [2] based on the packages `VoronoiFVM.jl` [9] and `ForwardDiff.jl` [11].

The example is a modification of one of the numerical examples in [3]. It considers the problem (1)–(3) in $\Omega = (0, 50)$ with homogeneous Dirichlet boundary conditions for Φ and homogeneous Neumann boundary conditions for c with $c^{\mathrm{dop}} = -0.5$ and $c_0 = 0.7$. We choose a self-consistent initial value Φ_0 for the electrostatic potential such that (3) is fulfilled for c_0.

For this test case, the four schemes behave similarly, as shown in the right picture of Fig. 1. As demonstrated in [3], more extreme examples forcing concentrations to be close to the physical bounds reveal important differences. The left picture of Fig. 1 shows that, for large t, the charge carrier concentration approaches a steady state with two space charge regions. We remark that c stays in the range (0, 1), and that the energy (12) decreases during the time evolution for all four schemes discussed in this paper, as stated in Theorem 1. At the end of the time evolution, an electroneutral region occurs in the center of the domain. At both boundaries, equally charged space regions set up enrichment boundary layers due to the fact that the amount of charge carriers confined to the domain cannot be compensated by the doping.

For the convergence experiment (see Fig. 2) we present results for scheme (S) only, the other schemes discussed perform similarly. For the space discretization, we used 6 levels of refinement building on a subdivision into 100 intervals for the coarsest mesh. Following a suggestion of Gajewski and Gärtner [10], we used an adaptive strategy based on the equidistribution of the energy dissipated per time step for the control of the time step size. We start with $t_1 = 10^{-4}$ and use the following expression to calculate the next time step:

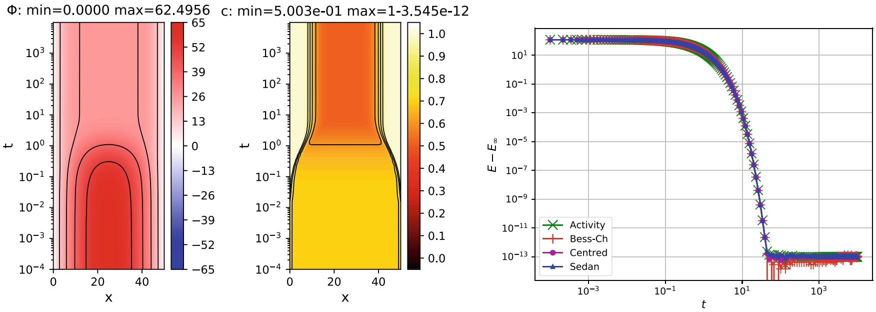


Fig. 1 Left: Time evolution of solution for scheme (S) on domain $\Omega = (0, 50)$ with constant initial value $c = 0.7$, homogeneous Dirichlet boundary conditions for Φ, $c^{\mathrm{dop}} = -\frac{1}{2}$ and homogeneous Neumann boundary conditions for c. Right: Evolution of the relative free energy according to (12) for the different schemes

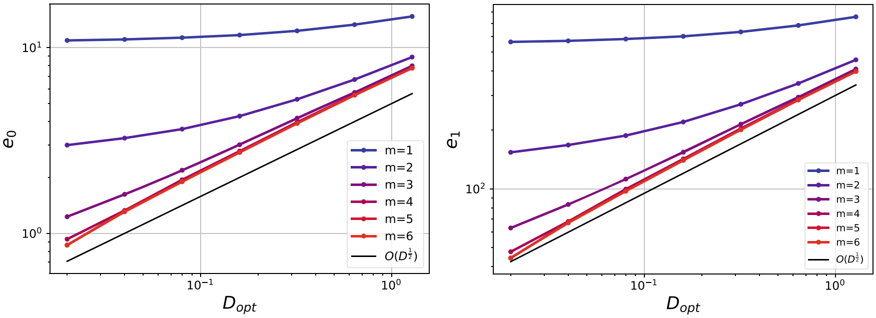


Fig. 2 Error e_0 in $L^2\left((0,T); L^2(\Omega)^2\right)$ (left) and e_1 in $L^2\left((0,T); H^1(\Omega)^2\right)$ on (c, Φ) for scheme (S) versus optimal energy dissipation per time step $\mathscr{D}_{opt}$ for grid step sizes $h = 0.5 * 2^{-m}$ for $m = 1 \ldots 6$

$$t_{n+1} = \min\left\{t_n \cdot 1.2,\ t_n \cdot \frac{\mathscr{D}_{opt}}{\mathscr{D}_n}, 100\right\},$$

where $\mathscr{D}_n = |E^n - E^{n-1}|$ is the change in the free energy during the previous time-step and $\mathscr{D}_{opt}$ is the parameter which controls the time step size. This approach ensures that the energy dissipated per time step remains of the same order as $\mathscr{D}_{opt}$ outside of a start region where the time-step size is ramped up and a final region where the dissipation rate approaches zero.

In Fig. 2, we show for a sequence of meshes the convergence of the $L^2(L^2)$ and $L^2(H^1)$ errors for the approximate solution $(c_{\mathscr{T},\Delta t}, \Phi_{\mathscr{T},\Delta t})$ with respect to a reference solution calculated on a fine space-time grid. For coarse space discretizations, errors are dominated by the spatial error, and decreasing the time step control parameter $\mathscr{D}_{opt}$ does not decrease the overall error. On the other hand, on fine spatial grids, we observe that the errors seem to decrease proportionally to the square root of $\mathscr{D}_{opt}$ which gives rise to a corresponding hypothesis to be investigated in further research.

References

1. Bessemoulin-Chatard, M.: A finite volume scheme for convection-diffusion equations with nonlinear diffusion derived from the Scharfetter-Gummel scheme. Numer. Math. **121**(4), 637–670 (2012)
2. Bezanson, J., Edelman, A.L., Karpinski, S., Shah, V.B.: Julia: a fresh approach to numerical computing. SIAM Rev. **59**(1), 65–98 (2017)
3. Cancès, C., Chainais-Hillairet, C., Fuhrmann, J., Gaudeul, B.: A numerical analysis focused comparison of several finite volume schemes for a unipolar degenerate drift-diffusion model. J. of Num. Anal. https://hal.archives-ouvertes.fr/hal-02194604 (to appear in IMA) (2020)
4. Cancès, C., Guichard, C.: Numerical analysis of a robust free energy diminishing finite volume scheme for parabolic equations with gradient structure. Found. Comput. Math. **17**(6), 1525–1584 (2017)

5. Coehoorn, R., Pasveer, W.F., Bobbert, P.A., Michels, M.A.J.: Charge-carrier concentration dependence of the hopping mobility in organic materials with gaussian disorder. Phys. Rev. B **72**(15), 155206 (2005)
6. Eymard, R., Gallouët, T., Herbin, R.: Finite volume methods. In: Ciarlet, P.G. (ed.) Handbook of Numerical Analysis. North-Holland, Amsterdam, pp. 713–1020 (2000)
7. Fuhrmann, J.: Comparison and numerical treatment of generalised Nernst-Planck models. Comput. Phys. Commun. **196**, 166–178 (2015)
8. Fuhrmann, J.: A numerical strategy for Nernst-Planck systems with solvation effect. Fuel Cells **16**, 12 (2016)
9. Fuhrmann, J.: VoronoiFVM.jl: Solver for coupled nonlinear partial differential equations based on the voronoi finite volume method (2019). https://doi.org/10.5281/zenodo.3529808
10. Gajewski, H., Gärtner, K.: On the discretization of van Roosbroeck's equations with magnetic field. Z. Angew. Math. Mech. **76**(5), 247–264 (1996)
11. Revels, J., Lubin. M., Papamarkou, T.: Forward-mode automatic differentiation in Julia. arXiv:1607.07892 [cs.MS] (2016)
12. Scharfetter, D.L., Gummel, H.K.: Large-signal analysis of a silicon read diode oscillator. IEEE Trans. Electron Dev. **16**(1), 64–77 (1969)
13. Vágner, P., Guhlke, C., Miloš, V., Müller, R., Fuhrmann, J.: A continuum model for yttria-stabilized zirconia incorporating triple phase boundary, lattice structure and immobile oxide ions. J. Solid State Electrochem., pp. 1–20 (2019)
14. Yu, Z., Dutton, R.: SEDAN III. www-tcad.stanford.edu/tcad/programs/sedan3.html (1988)

Non-isothermal Scharfetter–Gummel Scheme for Electro-Thermal Transport Simulation in Degenerate Semiconductors

Markus Kantner and Thomas Koprucki

Abstract Electro-thermal transport phenomena in semiconductors are described by the non-isothermal drift-diffusion system. The equations take a remarkably simple form when assuming the Kelvin formula for the thermopower. We present a novel, non-isothermal generalization of the Scharfetter–Gummel finite volume discretization for degenerate semiconductors obeying Fermi–Dirac statistics, which preserves numerous structural properties of the continuous model on the discrete level. The approach is demonstrated by 2D simulations of a heterojunction bipolar transistor.

Keywords Scharfetter–Gummel scheme · Fermi–Dirac statistics · Electro-thermal transport · Non-isothermal drift-diffusion system · Seebeck effect · Self-heating

MSC (2010) 35K05 · 35K08 · 35Q79 · 65N08 · 80M12 · 82B35 · 82D37

1 Introduction

Self-heating effects are a major concern in modern semiconductor devices, where the on-going miniaturization of feature size leads to increased power loss densities. The optimal design of semiconductor devices relies on numerical simulations, based on thermodynamically consistent models for the coupled electro-thermal transport processes. The standard model for the simulation of self-consistent charge and heat transport processes is the non-isothermal drift-diffusion system [1, 5, 9], which couples the semiconductor device equations to a heat transport equation. The magnitude of the thermoelectric cross effects (Seebeck effect, Thomson–Peltier effect) is governed by the Seebeck coefficient (also *thermopower*), which quantifies the thermoelectric voltage induced by a temperature gradient. Recently [5], the non-isothermal

M. Kantner · T. Koprucki (✉)
Weierstrass Institute (WIAS), Mohrenstr. 39, 10117 Berlin, Germany
e-mail: koprucki@wias-berlin.de

M. Kantner
e-mail: kantner@wias-berlin.de

R. Klöfkorn et al. (eds.), *Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples*, Springer Proceedings in Mathematics & Statistics 323,
https://doi.org/10.1007/978-3-030-43651-3_14

drift-diffusion system has been studied assuming the so-called *Kelvin formula* for the thermopower [8], which has two important implications: First, the Seebeck term in the current density expressions can be entirely absorbed in a temperature-dependent diffusion constant via a generalized Einstein relation. Second, the heat generation rate involves solely the three classically known self-heating effects without any further (transient) contribution. The model equations and its key features are described in Sect. 2. In Sect. 3, we present a finite volume discretization based on a novel, non-isothermal generalization of the Scharfetter–Gummel scheme for the discrete fluxes. The scheme holds for Fermi–Dirac statistics and preserves numerous structural and thermodynamic properties of the continuous system.

2 Non-isothermal Drift-Diffusion System

We consider the non-isothermal drift-diffusion system on $\Omega \subset \mathbb{R}^d$, $d \in \{1, 2, 3\}$,

$$-\nabla \cdot \varepsilon \nabla \Phi = q\,(C + p - n)\,, \tag{1}$$

$$q\partial_t n - \nabla \cdot \mathbf{j}_n = -qR, \tag{2}$$

$$q\partial_t p + \nabla \cdot \mathbf{j}_p = -qR, \tag{3}$$

$$c_V\partial_t T - \nabla \cdot \kappa \nabla T = H. \tag{4}$$

Poisson's Eq. (1) describes the electrostatic potential Φ generated by the electron density n, the density of valence band holes p and the built-in doping profile C. Here, q is the elementary charge and ε is the (absolute) permittivity of the material. The transport and recombination dynamics of the electrons and holes are modeled by the continuity Eqs. (2)–(3), where $\mathbf{j}_{n/p}$ are the electrical current densities and R is the (net-)recombination rate, which comprises several radiative and non-radiative processes [4, 7]. The temperature distribution in the device is described by the heat equation (4), where c_V is the volumetric heat capacity, κ is the thermal conductivity and H is the heat generation rate.

The carrier densities are related with the quasi-Fermi potentials $\varphi_{n/p}$, the electrostatic potential Φ and the (absolute) temperature T via the state equations

$$n = N_c\,(T)\,\mathscr{F}\left(\frac{q(\Phi-\varphi_n)-E_c(T)}{k_B T}\right), \quad p = N_v\,(T)\,\mathscr{F}\left(\frac{E_v(T)-q(\Phi-\varphi_p)}{k_B T}\right), \tag{5}$$

where $N_{c/v}$ are the effective density of states, $E_{c/v}$ are the band edge energies of the conduction and the valence band, respectively, and k_B is Boltzmann's constant. The function $\mathscr{F}$ describes the occupation probability of the electronic states. In the case of non-degenerate semiconductors (Maxwell–Boltzmann statistics), $\mathscr{F}\,(\eta) = \exp\,(\eta)$ is an exponential function. At high carrier densities, where degeneration effects due to the Pauli exclusion principle (Fermi–Dirac statistics) must be taken into account,

$\mathscr{F}$ is typically given by the Fermi–Dirac integral $F_{1/2}$ [4]. The approach outlined below, does not rely on the specific form of $\mathscr{F}$ and is applicable to materials with arbitrary density of states and degenerate or non-degenerate statistics [5].

2.1 Kelvin Formula for the Thermopower

The electrical current densities are modeled as

$$\mathbf{j}_n = -\sigma_n \left(\nabla\varphi_n + P_n \nabla T\right), \qquad \mathbf{j}_p = -\sigma_p \left(\nabla\varphi_p + P_p \nabla T\right), \tag{6}$$

where $\sigma_{n/p}$ are the electrical conductivities and $P_{n/p}$ are the thermopowers of the material. In this paper, we choose the thermopowers according to the Kelvin formula as variational derivatives of the entropy $\mathcal{S}$ with respect to the carrier densities

$$qP_n = -\mathrm{D}_n\mathcal{S}\left(n, p, T\right), \qquad qP_p = +\mathrm{D}_p\mathcal{S}\left(n, p, T\right), \tag{7}$$

where D denotes the Gâteaux derivative. The Kelvin formula is the low frequency and long wavelength limit of the microscopically exact Kubo formula [8]. It was shown to provide a good approximation for several materials at sufficiently high temperature. The entropy is obtained from the free energy $\mathcal{F}\left(n, p, T\right)$ of the system.

We assume the free energy functional [1, 5]

$$\begin{aligned}\mathcal{F}\left(n, p, T\right) = \int_\Omega \mathrm{d}V \Bigg(& k_B T \mathscr{F}^{-1}\left(\frac{n}{N_c}\right) n - k_B T N_c \mathscr{G}\left(\mathscr{F}^{-1}\left(\frac{n}{N_c}\right)\right) + E_c(T) n \\ & + k_B T \mathscr{F}^{-1}\left(\frac{p}{N_v}\right) p - k_B T N_v \mathscr{G}\left(\mathscr{F}^{-1}\left(\frac{p}{N_v}\right)\right) - E_v(T) p \Bigg) \\ & + \int_\Omega \mathrm{d}V\, f_L\left(T\right) + \frac{1}{2}\int_\Omega \mathrm{d}V \int_\Omega \mathrm{d}V'\, G\left(\mathbf{r}, \mathbf{r}'\right) \rho\left(\mathbf{r}\right) \rho\left(\mathbf{r}'\right) + \int_\Omega \mathrm{d}V\, \Phi_{\mathrm{ext}} \rho,\end{aligned} \tag{8}$$

where the first two lines describe the free energy of the non-interacting electron-hole plasma (quasi-free Fermi gas), f_L is the free energy of the lattice phonons (ideal Bose gas), $\mathscr{G}$ is the antiderivative of $\mathscr{F}$ (i.e., $\mathscr{G}'\left(\eta\right) = \mathscr{F}\left(\eta\right)$), $G\left(\mathbf{r}, \mathbf{r}'\right)$ is the Green's function of Poisson's equation and $\rho = q\left(p - n\right)$ is the mobile charge density. The potential Φ_{ext} is generated by the built-in doping-profile and the applied bias.

The free energy (8) recovers the state equations (5) via the variational derivative with respect to the carrier densities $\mathrm{D}_{n/p}\mathcal{F} := \mp q\varphi_{n/p}$, which is the defining relation for the quasi-Fermi potentials, see [5]. The entropy functional is defined as the derivative of the free energy (8) with respect to the temperature: $\mathcal{S}\left(n, p, T\right) = -\partial_T \mathcal{F}\left(n, p, T\right)$. Evaluation of Eq. (7) yields the thermopowers

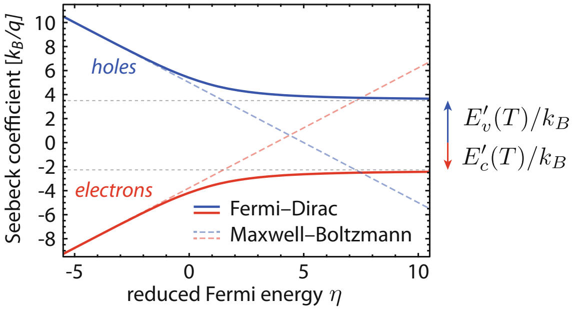


Fig. 1 Thermopowers $P_{n/p}$ according to Eq. (9) as functions of the reduced Fermi energy η (argument of $\mathscr{F}$ in Eq. (5)) in units of k_B/q. The thermopowers are plotted for $\mathscr{F}(\eta) = F_{1/2}(\eta)$ and $N_{c/v} \propto T^{3/2}$. Adapted, with permission, from [5]

$$P_n(n,T) = -\frac{k_B}{q}\left(\frac{TN_c'(T)}{N_c(T)}g\left(\frac{n}{N_c(T)}\right) - \mathscr{F}^{-1}\left(\frac{n}{N_c(T)}\right) - \frac{1}{k_B}E_c'(T)\right), \tag{9a}$$

$$P_p(p,T) = +\frac{k_B}{q}\left(\frac{TN_v'(T)}{N_v(T)}g\left(\frac{p}{N_v(T)}\right) - \mathscr{F}^{-1}\left(\frac{p}{N_v(T)}\right) + \frac{1}{k_B}E_v'(T)\right). \tag{9b}$$

The temperature-dependency of the band edge energies can be modeled using, e.g., the Varshni model [5, 7]. The function

$$g(x) = x\left(\mathscr{F}^{-1}\right)'(x) \tag{10}$$

quantifies the degeneration of the carriers ($g > 1$ for Fermi–Dirac statistics; $g \equiv 1$ for Maxwell–Boltzmann statistics). See Fig. 1 for a plot of the Seebeck coefficients (9).

2.2 *Drift-Diffusion Currents and Heat Generation Rate*

The Kelvin formula has two important implications, which lead to a very simple and appealing form of the thermoelectric cross effects in the system (1)–(4).

First, we rewrite the electrical current densities by passing from the thermodynamic form (6) to the drift-diffusion form. By explicitly evaluating the gradient of the quasi-Fermi potentials using the state equation (5), one observes that the Seebeck terms $\mathbf{j}_{n/p}|_{\text{Seebeck}} = -\sigma_{n/p}P_{n/p}\nabla T$ cancel out *exactly* from the expressions [5]. Using the conductivities $\sigma_n = qM_n n$ and $\sigma_p = qM_p p$ (with mobilities $M_{n/p}$), one arrives at

$$\mathbf{j}_n = -qM_n n\nabla\Phi + qD_n(n,T)\nabla n, \quad \mathbf{j}_p = -qM_p p\nabla\Phi - qD_p(p,T)\nabla p. \tag{11}$$

We emphasize that in Eq. (11)—even though there is no explicit thermal driving force $\propto \nabla T$—the Seebeck effect is fully taken into account via the (temperature-dependent) diffusion coefficients $D_{n/p}$. The latter obey the generalized Einstein relations [6]

$$qD_n = k_B T M_n g\left(n/N_c\left(T\right)\right), \qquad qD_p = k_B T M_p g\left(p/N_v\left(T\right)\right). \tag{12}$$

The flux discretization described in Sect. 3.1 is based on the drift-diffusion form (11).

The second implication of the Kelvin formula concerns the heat generation rate H. The commonly accepted model for H, which was derived by Wachutka [9] from fundamental laws of linear irreversible thermodynamics, takes a particularly simple form, when assuming the Kelvin formula for the thermopower. One obtains

$$H = \sum_{\lambda\in\{n,p\}} \frac{1}{\sigma_\lambda}\|\mathbf{j}_\lambda\|^2 - \sum_{\lambda\in\{n,p\}} T\,\mathbf{j}_\lambda\cdot\nabla P_\lambda + q\left(\varphi_p + TP_p - \varphi_n - TP_n\right)R, \tag{13}$$

which involves solely the three classically known self-heating effects, namely Joule heating (first term), the Thomson–Peltier effect (second term) and recombination heating (last term). Any further (transient) contributions, which necessarily arise for thermopowers different from the Kelvin formula (7), do not occur in the model.

3 Finite Volume Discretization

We assume a boundary conforming Delaunay triangulation of the computational domain $\Omega \subset \mathbb{R}^d$, $d = \{1, 2, 3\}$, and obtain the finite volume discretization [4] of the (stationary) system (1)–(4) by integration over the (restricted) Voronoï cells as

$$-\sum_{L\in N(K)} s_{K,L}\varepsilon\left(\Phi_L - \Phi_K\right) = q|\Omega_K|\left(C_K + p_K - n_K\right), \tag{14a}$$

$$-\sum_{L\in N(K)} s_{K,L}J_{n,K,L} = -q|\Omega_K|R_K, \tag{14b}$$

$$+\sum_{L\in N(K)} s_{K,L}J_{p,K,L} = -q|\Omega_K|R_K, \tag{14c}$$

$$-\sum_{L\in N(K)} s_{K,L}\kappa_{K,L}\left(T_L - T_K\right) = \frac{1}{2}\sum_{L\in N(K)} s_{K,L}\left(H_{J,K,L} + H_{\text{T–P},K,L}\right) + |\Omega_K|H_{R,K}. \tag{14d}$$

Here, $|\Omega_K|$ is the volume of the K-th Voronoï cell, $s_{K,L} = |\partial\Omega_K \cap \partial\Omega_L| / \|\mathbf{r}_L - \mathbf{r}_K\|$ is a geometric factor and $N\left(K\right)$ is the set of adjacent nodes of K. The subscripts K, L indicate evaluation on the respective nodes or edges. The discrete heat sources are

$$H_{J,K,L} = -\sum_{\lambda\in\{n,p\}} J_{\lambda,K,L}\left(\varphi_{\lambda,L} - \varphi_{\lambda,K} + P_{\lambda,K,L}\left(T_L - T_K\right)\right), \tag{15a}$$

$$H_{\text{T–P},K,L} = -\sum_{\lambda\in\{n,p\}} T_{K,L} J_{\lambda,K,L}\left(P_{\lambda,L} - P_{\lambda,K}\right), \tag{15b}$$

$$H_{R,K} = q\left(\varphi_{p,K} + T_K P_{p,K} - \varphi_{n,K} - T_K P_{n,K}\right) R_K, \tag{15c}$$

where we used a technique involving a weakly converging gradient developed in [3] for the discretization of the Joule and Thomson–Peltier terms (see [5] for details).

3.1 Generalized Scharfetter–Gummel Scheme

A robust discretization of the flux projections $J_{n/p,K,L} = (\mathbf{r}_L - \mathbf{r}_K)\cdot\mathbf{j}_{n/p}$ is obtained by integrating Eq. (11) along the edge $\overline{KL} := \{\mathbf{r}\,(x) = x\,\mathbf{r}_L + (1-x)\,\mathbf{r}_K,\ x \in [0,1]\}$, while assuming the electric field, the current density and the mobility to be constant along $\overline{KL}$. The temperature is assumed to be an affine function between adjacent nodes: $T\,(x) = x\,T_L + (1-x)\,T_K$, $x\in[0,1]$. In the case of Fermi–Dirac statistics (with $g\neq 1$), the resulting two-point boundary value problem on $x\in[0,1]$ [5]

$$k_B T(x) g\left(\frac{n(x)}{N_c\left(T(x)\right)}\right)\frac{\mathrm{d}n}{\mathrm{d}x} = q\left(\Phi_L - \Phi_K\right)n(x) + \frac{J_{n,K,L}}{M_{n,K,L}}, \quad n(0) = n_K, \quad n(1) = n_L,$$

can be solved approximately, by freezing the degeneracy factor (10) to a suitable average $g_{n/p,K,L}$ [2, 6]. One obtains the non-isothermal Scharfetter–Gummel scheme

$$J_{n,K,L} = M_{n,K,L} k_B T_{K,L} g_{n,K,L}\left(n_L B\left(X_{n,K,L}\right) - n_K B\left(-X_{n,K,L}\right)\right), \tag{16}$$

(holes analogously) with $X_{n,K,L} = q\left(\Phi_L - \Phi_K\right)/\left(k_B T_{K,L} g_{n,K,L}\right)$ and the Bernoulli function $B\,(x) = x/\left(\exp\,(x) - 1\right)$. The averaged degeneracy factor (consistent with the thermodynamic equilibrium [2, 6]) and the logarithmic mean temperature read

$$g_{n,K,L} = \frac{\eta_{n,L} - \eta_{n,K}}{\log\left(\mathscr{F}\left(\eta_{n,L}\right)/\mathscr{F}\left(\eta_{n,K}\right)\right)}, \quad T_{K,L} = \Lambda\left(T_L, T_K\right) = \frac{T_L - T_K}{\log\left(T_L/T_K\right)}. \tag{17}$$

The scheme (16) is a non-isothermal generalization of the scheme developed in [2, 6].

3.2 Structure-Preserving Properties

The discrete system (14)–(16) has several structure-preserving properties that hold without any smallness assumption. The conservation of charge is immediately guaranteed by the finite volume discretization [4]. Moreover, the scheme (16) is robust in both the drift- and diffusion dominated limits, as it interpolates between the upwind scheme for $X_{n,K,L} \to \pm\infty$ (strong electric field) and a central finite difference scheme for $X_{n,K,L} = 0$ (pure diffusion). The latter involves a discrete analogue of the nonlinear diffusion constant (12) using $g_{n,K,L}$ as in Eq. (17). For the analysis of further properties, which address the consistency with thermodynamics, it is convenient to recast the formula (16) into a discrete analogue of its thermodynamic form (6):

$$J_{n,K,L} = -\sigma_{n,K,L} \left(\varphi_{n,L} - \varphi_{n,K} + P_{n,K,L} \left(T_L - T_K\right)\right). \tag{18}$$

The edge-averaged discrete conductivity, which is implicitly taken by the Scharfetter–Gummel discretization, is a "tilted" logarithmic mean Λ of the carrier densities

$$\sigma_{n,K,L} = \frac{q M_{n,K,L}}{\operatorname{sinhc}\left(\frac{1}{2} X_{n,K,L}\right)} \Lambda \left(n_L \exp\left(-\frac{1}{2} X_{n,K,L}\right), n_K \exp\left(+\frac{1}{2} X_{n,K,L}\right)\right), \tag{19}$$

with $\operatorname{sinhc}(x) = \sinh(x)/x$. The thermopower $P_{n,K,L}$ (required in Eq. (15a)) reads

$$\begin{aligned} P_{n,K,L} = -\frac{k_B}{q} \Bigg[& \log\left(\frac{N_c(T_L)}{N_c(T_K)}\right) \frac{g_{n,K,L}}{\log\left(T_L/T_K\right)} - \frac{1}{k_B} \frac{E_c(T_L) - E_c(T_K)}{T_L - T_K} \\ & - \frac{\left(T_L - T_{K,L}\right)\eta_{n,L} - \left(T_K - T_{K,L}\right)\eta_{n,K}}{T_L - T_K} \Bigg]. \end{aligned} \tag{20}$$

The scheme is manifestly consistent with the thermodynamic equilibrium (no current for $\varphi_{n,K} = \varphi_{n,L}$ and $T_K = T_L$) and the limiting cases of either vanishing chemical ($\varphi_{n,K} = \varphi_{n,L}$: pure Seebeck current) or thermal ($T_K = T_L$: isothermal drift-diffusion) driving forces. The discretization guarantees the non-negativity of the Joule heat term

$$H_{J,K,L} = \sum_{\lambda \in \{n,p\}} \sigma_{\lambda,K,L} \left|\varphi_{\lambda,L} - \varphi_{\lambda,K} + P_{\lambda,K,L}\left(T_L - T_K\right)\right|^2 \geq 0 \tag{21}$$

(using Eqs. (15a) and (18)) and subsequently also the consistency with the 2nd law of thermodynamics [5]. In a 1D case study [5], the scheme (16) was found to be significantly more accurate than the conventional Scharfetter–Gummel-type discretization approach. Both schemes revealed quadratic convergence, but the new scheme (16) saved 1–2 refinement steps to reach the same level of accuracy.

4 Numerical Simulation of a Heterojunction Bipolar Transistor

The approach is demonstrated by numerical simulations of the GaAs/ AlGaAs-based heterojunction bipolar transistor (HBT) shown in Fig. 2a. We assume ideal ohmic contacts with perfect heat sinking ($T_{\mathrm{cont}} = 300\,\mathrm{K}$) and homogeneous Neumann boundary conditions else. The material parameters, including temperature-dependent models for the band edge energies, mobilities and the thermal conductivity, are taken from [7]. The validity of the Kelvin formula for GaAs was studied in [5]. The calculated current-voltage curves (with and without self-heating effects) are shown in Fig. 2b.

The temperature distribution and the heat generation rate are plotted in Fig. 3 for different collector-emitter voltages. The Thomson–Peltier effect is found to cool the AlGaAs/ GaAs heterojunctions (emitter/ emitter cap and emitter/ base junction, blue color in Fig. 3b, d) and heats up the collector/ subcollector junction. With increasing

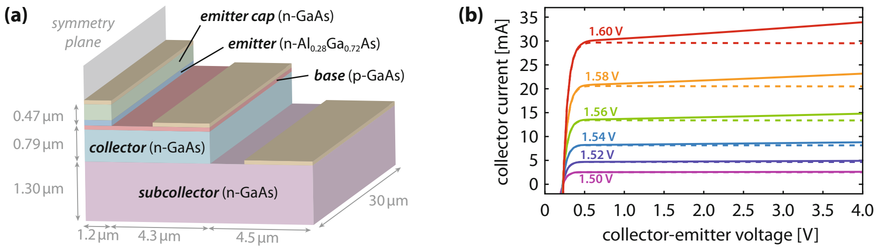


Fig. 2 **a** Sketch of the considered GaAs/ AlGaAs-HBT. Due to symmetry, only half of the device is simulated. The doping densities are: $N_D^+ = 4 \times 10^{19}\,\mathrm{cm}^{-3}$ (emitter cap), $N_D^+ = 2 \times 10^{17}\,\mathrm{cm}^{-3}$ (emitter), $N_A^- = 3 \times 10^{19}\,\mathrm{cm}^{-3}$ (base), $N_D^+ = 2 \times 10^{16}\,\mathrm{cm}^{-3}$ (collector) and $N_D^+ = 5 \times 10^{18}\,\mathrm{cm}^{-3}$ (subcollector). **b** Calculated collector current I_C as a function of the collector-emitter voltage U_{CE} for different base-emitter voltages U_{BE} with (solid lines) and without (dashed) self-heating effects

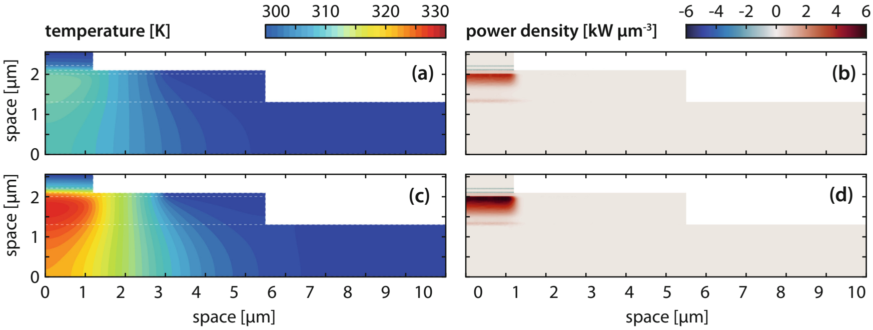


Fig. 3 Simulated temperature distribution and self-heating power density H at stationary operation with **a**, **b** $U_{\mathrm{CE}} = 2\,\mathrm{V}$ and **c**, **d** $U_{\mathrm{CE}} = 4\,\mathrm{V}$. The basis-emitter voltage is $U_{\mathrm{BE}} = 1.6\,\mathrm{V}$ in both cases

current densities (i.e., increasing collector-emitter voltage), the relative importance of Joule heating increases, until it becomes the dominant effect. This leads to a strong temperature increase in the collector region close to the symmetry axis. Recombination processes additionally heat the base region below the base/ emitter junction, but were found to be of minor importance in the present study.

5 Conclusions

The Kelvin formula for the thermopower yields a remarkably simple form of the non-isothermal drift-diffusion system. The specific form of the current density expressions, which contain the thermal driving forces only implicitly, allow for a non-isothermal generalization of the Scharfetter–Gummel scheme for Fermi–Dirac statistics that was previously presented in [2, 6]. The resulting finite volume scheme preserves fundamental thermodynamic properties and relations on the discrete level.

Acknowledgements This work was funded by the German Research Foundation (DFG) under Germany's Excellence Strategy—EXC2046: Math+ (Berlin Mathematics Research Center).

References

1. Albinus, G., Gajewski, H., Hünlich, R.: Thermodynamic design of energy models of semiconductor devices. Nonlinearity **15**(2), 367–383 (2002). https://doi.org/10.1088/0951-7715/15/2/307
2. Bessemoulin-Chatard, M.: A finite volume scheme for convection-diffusion equations with nonlinear diffusion derived from the Scharfetter-Gummel scheme. Numer. Math. **121**(4), 637–670 (2012). https://doi.org/10.1007/s00211-012-0448-x
3. Eymard, R., Gallouët, T.: H-convergence and numerical schemes for elliptic problems. SIAM J. Numer. Anal. **41**(2), 539–562 (2003). https://doi.org/10.1137/s0036142901397083
4. Farrell, P., Rotundo, N., Doan, D.H., Kantner, M., Fuhrmann, J., Koprucki, T.: Drift-Diffusion Models. In: Piprek, J. (ed.) Handbook of Optoelectronic Device Modeling and Simulation: Lasers, Modulators, Photodetectors, Solar Cells, and Numerical Methods, vol. 2, chap. 50, pp. 731–771. CRC Press, Taylor & Francis Group, Boca Raton (2017). https://doi.org/10.4324/9781315152318-25
5. Kantner, M.: Non-isothermal generalization of the Scharfetter–Gummel scheme for degenerate semiconductors using the Kelvin formula for the Seebeck coefficient. J. Comput. Phys. **402**, 109091 (2020). https://doi.org/10.1016/j.jcp.2019.109091 (to appear)
6. Koprucki, T., Rotundo, N., Farrell, P., Doan, D.H., Fuhrmann, J.: On thermodynamic consistency of a Scharfetter-Gummel scheme based on a modified thermal voltage for drift-diffusion equations with diffusion enhancement. Opt. Quantum. Electron. **47**(6), 1327–1332 (2015). https://doi.org/10.1007/s11082-014-0050-9
7. Palankovski, V., Quay, R.: Analysis and Simulation of Heterostructure Devices. Springer, Vienna (2004). https://doi.org/10.1007/978-3-7091-0560-3

8. Peterson, M.R., Shastry, B.S.: Kelvin formula for thermopower. Phys. Rev. B **82**, 195105 (2010). https://doi.org/10.1103/physrevb.82.195105
9. Wachutka, G.K.: Rigorous thermodynamic treatment of heat generation and conduction in semiconductor device modeling. IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. **9**(11), 1141–1149 (1990). https://doi.org/10.1109/43.62751

Entropy Diminishing Finite Volume Approximation of a Cross-Diffusion System

Clément Cancès and Benoît Gaudeul

Abstract We propose a two-point flux approximation finite volume scheme for the approximation of the solutions of a entropy dissipative cross-diffusion system. The scheme is shown to preserve several key properties of the continuous system, among which positivity and decay of the entropy. Numerical experiments illustrate the behaviour of our scheme.

Keywords Cross-diffusion system · Entropy dissipative scheme

MSC (2010) 65M08 · 65M12 · 35K51

1 Finite Volume Approximation of a Cross Diffusion System

The model addressed in this paper is a toy model for the evolution of a material [1] which can be derived thanks to a jump process following the program of [3]. We are interested in the evolution of the composition of the material, which is described by the concentrations $\mathbf{c} = (c_1, \dots, c_I)$ of I different species. The material is represented by an open, connected, bounded, and polyhedral subset Ω of $\mathbb{R}^d$, and the evolution of its composition is prescribed by the following system of partial differential equations. The mass conservation of each species writes for all $i \in \{1, \dots, I\}$

$$\partial_t c_i + \nabla \cdot \mathbf{J}_i = 0 \quad \text{in } \mathbb{R}_+ \times \Omega, \quad \text{with} \quad \mathbf{J}_i = \sum_{j \neq i} \kappa_{ij} \left(c_i \nabla c_j - c_j \nabla c_i \right). \tag{1}$$

C. Cancès (✉)
Inria, Univ. Lille, CNRS, UMR 8524 - Laboratoire Paul Painlevé,
59000 Lille, France
e-mail: clement.cances@inria.fr

B. Gaudeul
Univ. Lille, CNRS, Inria, UMR 8524 - Laboratoire Paul Painlevé,
59000 Lille, France
e-mail: benoit.gaudeul@univ-lille.fr

R. Klöfkorn et al. (eds.), *Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples*, Springer Proceedings in Mathematics & Statistics 323,
https://doi.org/10.1007/978-3-030-43651-3_15

The coefficients κ_{ij} are such that $\kappa_{ij} = \kappa_{ji} \geq 0$. The system is complemented with no-flux boundary conditions $\mathbf{J}_i \cdot \mathbf{n} = 0$ on $\partial\Omega$, and an initial condition $\mathbf{c}^0 = \left(c_1^0, \ldots, c_I^0\right)$ which satisfies $\langle \mathbf{c}, \mathbf{1}\rangle = \sum_{i=1}^I c_i^0 = 1$ in Ω.

This continuous problem has some key-properties that one wants to preserve after discretisation. First, the total mass of each specie is conserved, i.e, $\int_\Omega c_i(t) = \int_\Omega c_i^0$ for all $i \in \{1, \ldots, I\}$ and $t \geq 0$. This follows directly from the local conservation property (1) and the no-flux boundary conditions across $\partial\Omega$. Second, the concentrations remains non-negative, i.e., $c_i(\mathbf{x}, t) \geq 0$. Third, the expression (1) of $\mathbf{J}_i$ and the condition $\kappa_{ij} = \kappa_{ji}$ yield $\sum_{i=1}^I \mathbf{J}_i = \mathbf{0}$, so that $\sum_{i=1}^I c_i(\mathbf{x}, t) = 1$ for all $(\mathbf{x}, t) \in \Omega \times \mathbb{R}_+$. Therefore, $\mathbf{c}(t)$ takes values in the closed convex set

$$\mathscr{A} = \left\{ \mathbf{c} \in L^1(\Omega; \mathbb{R}_+^I) \;\middle|\; \langle \mathbf{c}, \mathbf{1}\rangle = 1 \text{ a.e. in } \Omega \text{ and } \int_\Omega c_i = \int_\Omega c_i^0 \right\}.$$

Finally, the fluxes rewrite $\mathbf{J}_i = -\sum_{j\neq i} \kappa_{ij} c_i c_j \nabla\left(\log(c_i) - \log(c_j)\right)$. Therefore, multiplying (1) by $\log(c_i)$ and integrating over Ω leads to

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathfrak{E}(\mathbf{c}) = -\sum_{\{i,j\}\in\{1,\ldots,I\}^2} \int_\Omega \kappa_{ij} c_i c_j \left|\nabla(\log(c_i) - \log(c_j))\right|^2 \leq 0, \tag{2}$$

where the entropy $\mathfrak{E}$ is the convex functional on $\mathscr{A}$ defined by

$$\mathfrak{E}(\mathbf{c}) = \sum_{i=1}^I \int_\Omega c_i \log(c_i).$$

Then due to the entropy / entropy dissipation relation (2), $t \mapsto \mathfrak{E}(\mathbf{c}(t))$ is non-increasing, and even decreasing unless $\mathbf{c}$ is constant w.r.t. space.

Under appropriate conditions on the coefficients κ_{ij}, the existence of weak solutions to the problem can be established thanks to the so-called entropy method [6, 7]. Strong solutions have been recently investigated in [2].

Our goal is to define a scheme that preserves at the discrete level the above properties, i.e. such that the approximate solution belongs to $\mathscr{A}$ for all time and with a discrete counterpart of (2). To this end, we still need to remark that if the coefficients κ_{ij} are equal to $\kappa^\star > 0$, then $\mathbf{J}_i = -\kappa^\star \nabla c_i$, so that (1) reduces to I decoupled heat equations. Therefore, choosing $\kappa^\star > 0$ and setting $\widetilde{\kappa}_{ij} = \kappa_{ij} - \kappa^\star$, $\mathbf{J}_i$ rewrites as

$$\mathbf{J}_i = -\sum_{j\neq i} \widetilde{\kappa}_{ij} c_i c_j \nabla\left(\log(c_i) - \log(c_j)\right) - \kappa^\star \nabla c_i, \qquad i \in \{1, \ldots, I\}. \tag{3}$$

Our approach consists in approximating the fluxes $\mathbf{J}_i$ under their above form (3). Since it is based on two-point flux approximation (TPFA) finite volumes, it requires the use of a so-called Δ-admissible mesh. Let $(\mathscr{T}, \mathscr{E}, (\mathbf{x}_K)_{K\in\mathscr{T}})$ be a finite volume mesh of Ω fulfilling the classical orthogonality condition required for the consistency of TPFA. Since this notion is classical, we remain sloppy here on the definition and refer to [5, Definition 9.1] or to the companion paper [4] for details. Let us just mention that $\mathcal{T}$ denotes the set of the cells, while only internal edges are considered in the set $\mathcal{E}$, i.e. $\mathcal{E} = \{\sigma = K|L = \partial K \cap \partial L$ for $K, L \in \mathcal{T}\}$. Given $K \in \mathcal{T}$, we denote by $\mathcal{E}_K = \{\sigma \in \mathcal{E} \mid \sigma \subset \partial K\}$ and by m_K the d-dimensional Lebesgue measure of K. For $\sigma = K|L$, we denote by m_σ the $(d-1)$-dimensional Lebesgue measure of σ, by $d_\sigma = |\mathbf{x}_K - \mathbf{x}_L|$ the distance between the cell centers, and by $a_\sigma = \frac{m_\sigma}{d_\sigma}$ the transmissivity of σ. For the time discretisation, we allow for non-uniform time steps $\tau_n = t^n - t^{n-1}$, $n \geq 1$, with $t^0 = 0$. The initial condition is discretised into

$$c_{i,K}^0 = \frac{1}{|K|}\int_K c_i^0, \qquad \forall K \in \mathcal{T},\ i \in \{1,\dots,I\}. \tag{4}$$

In particular, the corresponding piecewise constant reconstruction $\mathbf{c}_{\mathcal{T}}^0 = \left(c_{i,\mathcal{T}}^0\right)_i$, defined by $c_{i,\mathcal{T}}^0(\mathbf{x}) = \sum_{K\in\mathcal{T}} c_{i,K}^0 \chi_K(\mathbf{x})$, belongs to $\mathcal{A}$ provided $\mathbf{c}^0$ does. Now, we assume that $\left(c_{i,K}^{n-1}\right)_{i,K}$ is given and is such that the corresponding piecewise constant reconstruction $\mathbf{c}_{\mathcal{T}}^{n-1}$ belongs to $\mathcal{A}$, then we seek $\left(c_{i,K}^n\right)_{i,K}$ solution of the following nonlinear system. First, the conservation of mass is locally enforced on each cell K:

$$\frac{c_{i,K}^n - c_{i,K}^{n-1}}{\tau_n}|K| + \sum_{\sigma\in\mathcal{E}_K} m_\sigma J_{i,K\sigma}^n = 0, \qquad \forall K \in \mathcal{T},\ i \in \{1,\dots,I\}. \tag{5}$$

No flux boundary conditions translate to $J_{i,K\sigma}^n = 0$ if $\sigma \subset \partial\Omega$. The discretisation of the fluxes $J_{i,K\sigma}^n \simeq \mathbf{J}_i \cdot \mathbf{n}_{K\sigma}$ across the edge $\sigma = K|L$ relies on the expression (3) and writes

$$J_{i,K\sigma}^n = \kappa^\star \frac{c_{i,K}^n - c_{i,L}^n}{d_\sigma} + \sum_{j\neq i} \widetilde{\kappa}_{ij}\left(c_{j,\sigma}^n \frac{c_{i,K}^n - c_{i,L}^n}{d_\sigma} - c_{i,\sigma}^n \frac{c_{j,K}^n - c_{j,L}^n}{d_\sigma}\right) = -J_{i,L\sigma}^n. \tag{6}$$

Finally, the edge concentrations $c_{i,\sigma}^n$ are computed from the cell concentrations $c_{i,K}^n$ and $c_{i,L}^n$ thanks to the continuous formula

$$c_{i,\sigma}^n = \begin{cases} c_{i,K}^n & \text{if } c_{i,K}^n = c_{i,L}^n, \\ \frac{c_{i,K}^n - c_{i,L}^n}{\log(c_{i,K}^n) - \log(c_{i,L}^n)} & \text{if } c_{i,K}^n \neq c_{i,L}^n,\ c_{i,K}^n > 0, c_{i,L}^n > 0, \\ 0 & \text{if } \min(c_{i,K}^n, c_{i,L}^n) \leq 0. \end{cases} \tag{7}$$

The goal of this paper is to show that the scheme (5)–(7) suitably approximates the solutions to (1). This encompasses some mathematical properties of the scheme to be discussed in Sect. 2 and numerical results presented in Sect. 3.

Remark 1 Before going further, let us just highlight why the introduction of the positive parameter $\kappa^\star$ is important. Assume for simplicity that $I = 2$, so that the problem reduces to two uncoupled heat equations on c_1 and $c_2 = 1 - c_1$. Assume that the mesh $\mathscr{T}$ is made of two cells K and L separated by the unique edge σ, and that $c^0_{1,K} = 1$, $c^0_{2,K} = 0$, $c^0_{1,L} = 0$ and $c^0_{2,L} = 1$. Then formula (7) shows that $c^0_{1,\sigma} = c^0_{2,\sigma} = 0$. Therefore, if $\kappa^\star$ is set to 0, then $\mathbf{c}^0_{\mathscr{T}}$ is a steady solution to the scheme, which is not reasonable for the discretisation of the heat equation. The introduction of $\kappa^\star > 0$ annihilates this spurious solution.

2 Some Pieces of Numerical Analysis

Our first statement deals with positivity preservation, mass conservation, the preservation of the constraint $\sum_i c_i = 1$, and with the existence of a solution to the nonlinear system (5)–(7).

Proposition 1 *Given $\mathbf{c}^{n-1}_{\mathscr{T}} \in \mathscr{A}$, then there exists (at least) one approximate solution $\mathbf{c}^n_{\mathscr{T}}$ to the scheme such that $\mathbf{c}^n_{\mathscr{T}} \in \mathscr{A}$.*

Proof In order to carry out the proof, one first needs to replace $c^n_{i\sigma}$ (and $c^n_{j\sigma}$) by $\widetilde{c}^n_{i\sigma} = c^n_{i\sigma} / \max\left(1, \sum_\ell c^n_{\ell\sigma}\right)$ in (6). These two quantities will be shown later on to coincide as $\sum_\ell c^n_{\ell\sigma} \le 1$ on all the internal edges σ.

As a first step to prove that $\mathbf{c}^n_{\mathscr{T}} \in \mathscr{A}$, let us prove by contradiction that $c^n_{i,K} \ge 0$ for all $K \in \mathscr{T}$ and all $i \in \{1, \dots, I\}$. Assume that $\min_L c^n_{i,L} < 0$ for some i, and let K be the cell where $c^n_{i,K} < 0$ is minimum among all $c^n_{i,L}$, $L \in \mathscr{T}$. Then (7) implies that $\widetilde{c}^n_{i,\sigma} = 0$ for all $\sigma \in \mathscr{E}_K$ so that we deduce from (5)–(6) that

$$\sum_{\sigma \in \mathscr{E}_K} a_\sigma \left[\kappa^\star \left(1 - \sum_{j=1}^{I} \widetilde{c}^n_{j,\sigma} \right) \left(c^n_{i,K} - c^n_{i,L}\right) + \sum_{j=1}^{I} \kappa_{ij} \widetilde{c}^n_{j,\sigma} \left(c^n_{i,K} - c^n_{i,L}\right) \right] > 0.$$

Using $\widetilde{c}^n_{j,\sigma} \ge 0$, $\sum_j \widetilde{c}^n_{j,\sigma} \le 1$, and $c^n_{i,K} \le c^n_{i,L}$ in previous inequality yields a contradiction, hence $c^n_{i,K} \ge 0$ for all i and all K.

The fact that $\sum_{K \in \mathscr{T}} c^n_{i,K} m_K = \sum_{K \in \mathscr{T}} c^{n-1}_{i,K} m_K = \int_\Omega c^0_i$ follows directly from the conservativity of the fluxes (6). Finally, one readily checks from (6) that

$$\sum_{i=1}^{I} J^n_{i,K\sigma} = \frac{\kappa^\star}{d_\sigma} \sum_{i=1}^{I} (c^n_{i,K} - c^n_{i,L}), \qquad \forall \sigma = K|L \in \mathscr{E}.$$

So summing (5) over i shows that $s^n_K = \sum_{i=1}^{I} c^n_{i,K}$ satisfies the discrete heat equation

$$\frac{s^n_K - s^{n-1}_K}{\tau_n} m_K + \kappa^\star \sum_{\sigma = K|L \in \mathscr{E}_K} a_\sigma (s^n_K - s^n_L) = 0, \qquad \forall K \in \mathscr{T}.$$

Since $s^{n-1}_K = 1$ for all $K \in \mathscr{T}$, so does $\left(s^n_K\right)_K$. Therefore, $\mathbf{c}^n_{\mathscr{T}}$ belongs to $\mathscr{A}$. Now, it follows from a simple convexity argument that the logarithmic mean $c^n_{i,\sigma}$ of $c^n_{i,K}$ and $c^n_{i,L}$ is smaller than the arithmetic mean, the sum of which over i is equal to 1.

Therefore, $c_{i,\sigma}^n = \widetilde{c}_{i,\sigma}^n$. The existence proof then easily follows from a topological degree argument [8], see our companion paper [4] for details. □

Refining the above proof, one can show that $c_{i,K}^n > 0$ for all $K \in \mathscr{T}$ as soon as $\int_\Omega c_i^0 > 0$. This property is key for the proof of our next statement, and is rigorously established in [4]. Our second statement highlights the energy diminishing character of the scheme, which should be thought as a discrete counterpart of (2).

Proposition 2 *Let $\mathbf{c}_{\mathscr{T}}^n \in \mathscr{A}$ be a solution to the scheme as in Proposition 1, then*

$$\mathfrak{E}(\mathbf{c}_{\mathscr{T}}^n) \leq \mathfrak{E}(\mathbf{c}_{\mathscr{T}}^{n-1}).$$

Proof Without loss of generality, we assume that $\int_\Omega c_i^0 > 0$ for all i (otherwise $c_{i,K}^n = 0$ for all $K \in \mathscr{T}$ thanks to Proposition 1). Since $c_{i,K}^n > 0$, one can multiply (5) by $\log(c_{i,K}^n)$ and to sum over $K \in \mathscr{T}$ and $i \in \{1, \dots, I\}$, which leads to

$$A + B := \sum_{i=0}^{I} \sum_{K\in\mathscr{T}} \frac{c_{i,K}^n - c_{i,K}^{n-1}}{\tau_n} \log(c_{i,K}^n) m_K + \sum_{i=0}^{I} \sum_{K\in\mathscr{T}} \sum_{\sigma=K|L\in\mathscr{E}_K} m_\sigma J_{i,K\sigma}^n \log(c_{i,K}^n) = 0.$$

Thanks to the convexity of $c \mapsto c\log c - c$ and to mass conservation, one has

$$A \geq \frac{1}{\tau_n} \sum_{i=0}^{I} \sum_{K\in\mathscr{T}} \left(c_{i,K}^n \log c_{i,K}^n - c_{i,K}^{n-1} \log c_{i,K}^{n-1}\right) m_K = \frac{\mathfrak{E}(\mathbf{c}_{\mathscr{T}}^n) - \mathfrak{E}(\mathbf{c}_{\mathscr{T}}^{n-1})}{\tau_n}.$$

The particular choice (7) for $c_{i,\sigma}^n$ allows us to rewrite

$$J_{i,K\sigma}^n = \kappa^\star \frac{c_{i,K}^n - c_{i,L}^n}{d_\sigma} + \sum_{j\neq i} \widetilde{\kappa}_{ij} c_{i,\sigma}^n c_{j,\sigma}^n \left(\log(c_{i,K}^n) - \log(c_{j,K}^n) - \log(c_{i,L}^n) + \log(c_{j,L}^n)\right).$$

This implies that

$$\begin{aligned} B = \kappa^\star &\sum_{i=1}^{N} \sum_{\sigma=K|L\in\mathscr{E}} a_\sigma (c_{i,K}^n - c_{i,L}^n) \left(\log(c_{i,K}^n) - \log(c_{i,L}^n)\right) \\ &+ \sum_{\{i,j\}} \sum_{\sigma=K|L\in\mathscr{E}} \widetilde{\kappa}_{ij} a_\sigma c_{i,\sigma}^n c_{j,\sigma}^n \left(\log(c_{i,K}^n) - \log(c_{j,K}^n) - \log(c_{i,L}^n) + \log(c_{j,L}^n)\right)^2. \end{aligned}$$

Since the logarithmic mean $c_{i,\sigma}^n$ of $c_{i,K}^n$ and $c_{i,L}^n$ is smaller than the arithmetic mean, there holds $\sum_i c_{i,\sigma}^n \leq 1$. As a consequence, one has

$$\begin{aligned} &\sum_{i=1}^{N} \sum_{\sigma=K|L\in\mathscr{E}} a_\sigma (c_{i,K}^n - c_{i,L}^n) \left(\log(c_{i,K}^n) - \log(c_{i,L}^n)\right) \\ &\qquad \geq \sum_{\{i,j\}} \sum_{\sigma=K|L\in\mathscr{E}} a_\sigma c_{i,\sigma}^n c_{j,\sigma}^n \left(\log(c_{i,K}^n) - \log(c_{j,K}^n) - \log(c_{i,L}^n) + \log(c_{j,L}^n)\right)^2, \end{aligned}$$

which implies that

$$B \geq \sum_{\{i,j\}} \sum_{\sigma} \kappa_{ij} a_\sigma c^n_{i,\sigma} c^n_{j,\sigma} \left(\log(c^n_{i,K}) - \log(c^n_{j,K}) - \log(c^n_{i,L}) + \log(c^n_{j,L})\right)^2 \geq 0.$$

This concludes the proof of Proposition 2. □

A more involved study allows to show that under classical assumptions on non-degeneracy of the mesh regularity, then $\mathbf{c}_{\mathcal{T},\tau} : (t, \mathbf{x}) \mapsto \sum_{n\geq 1} \mathbf{c}^n_{\mathcal{T}}(\mathbf{x})\xi_{(t^{n-1},t^n]}(t)$ converges in $L^1_{\text{loc}}(\mathbb{R}_+ \times \overline{\Omega})$ towards a weak solution $\mathbf{c}$ to (1) provided $\kappa_{ij} > 0$ for all i, j. The proof relies on the exploitation of the regularization coming from the dissipation (term B in the proof of Proposition 2). We refer to [4] for the details of the convergence proof.

3 Numerical Results

The numerical scheme has been implemented using MATLAB. The nonlinear system corresponding to the scheme is solved thanks to Newton method with stopping criterion $\|\mathbf{c}^{n,k+1} - \mathbf{c}^{n,k}\|_\infty \leq 10^{-12}$. The next iterate $\mathbf{c}^{n,k+1}$ is then "projected" on $\mathcal{A}$ by setting $\mathbf{c}^{n,k+1} = \max(\mathbf{c}^{n,k+1}, 10^{-10}\tau)$, and then $\mathbf{c}^{n,k+1} = \mathbf{c}^{n,k+1}/(\sum_{i=1}^N \mathbf{c}_i^{n,k+1})$. For the first time step, we also make use of a continuation method based on the intermediate diffusion coefficients $\kappa^\lambda_{i,j} = \lambda\kappa_{ij} + (1-\lambda)\kappa^\star$ with $\lambda \in [0, 1]$. The parameter λ is originally set to 1. If the Newton's method does not converge, we let $\lambda = \left(\lambda + \lambda_{\text{prev}}\right)/2$ where λ_{prev} is originally set to 0. If the Newton's method converges, we let $\lambda_{\text{prev}} = \lambda$ and $\lambda = 1$.

Our first test case is devoted to the convergence analysis of the scheme in a one-dimensional setting $\Omega = (0, 1)$. Two different initial conditions are considered: $\mathbf{c}^0_s$ is smooth with coordinates that vanish pointwise at the boundary of Ω, whereas $\mathbf{c}^0_r$ is discontinuous with coordinates vanishing on intervals of Ω:

$$c^0_{1,s}(x) = \frac{1}{4} + \frac{1}{4}\cos(\pi x), \quad c^0_{2,s}(x) = \frac{1}{4} + \frac{1}{4}\cos(\pi x), \quad c^0_{3,s}(x) = \frac{1}{2} - \frac{1}{2}\cos(\pi x),$$
$$c^0_{1,r} = 1_{[\frac{3}{8},\frac{5}{8}]}, \quad c^0_{2,r} = 1_{(\frac{1}{8},\frac{3}{8})} + 1_{(\frac{5}{8},\frac{7}{8})}, \quad c^0_{3,r} = 1_{[0,\frac{1}{8}]} + 1_{[\frac{7}{8},1]}.$$

We also consider two diffusion matrices, one called regular with positive off-diagonal coefficients and an other called singular with a few null off-diagonal coefficients.

$$K^{\text{reg}} = \begin{pmatrix} 0 & 0.2 & 1 \\ 0.2 & 0 & 0.1 \\ 1 & 0.1 & 0 \end{pmatrix} \qquad K^{\text{sing}} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0.1 \\ 1 & 0.1 & 0 \end{pmatrix}$$

For the convergence tests, we have let $\kappa^\star = 0.1$ and the meshes are uniform discretisations of $[0, 1]$ from 2^5 cells to 2^{14} cells. Since we do not have an analytical solution

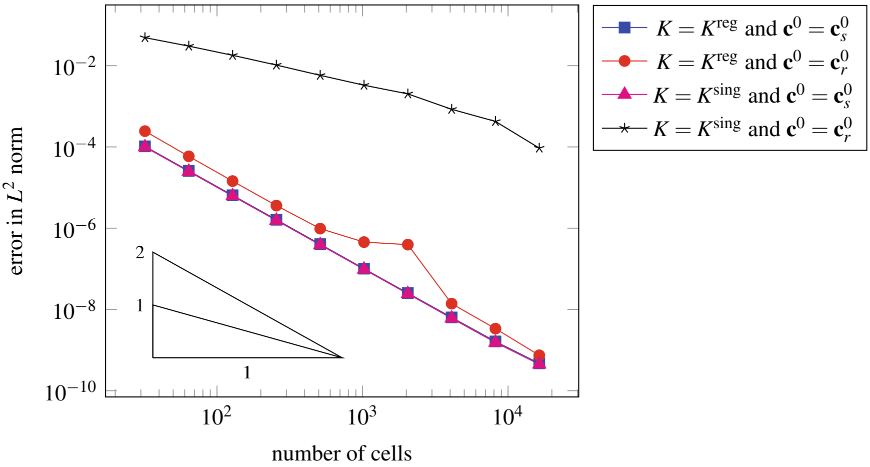
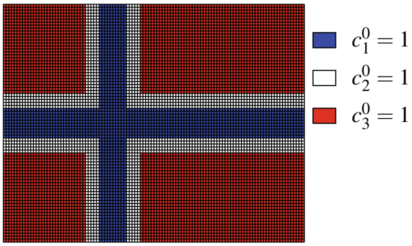


Fig. 1 Error with respect to the solution computed on the finest mesh for 1D settings

Fig. 2 Initial configuration $\mathbf{c}^0$



at hand, the approximate solutions are compared to a reference solution computed on a grid made of 2^{15} cells. The final time is 0.25, and the time discretisation is fixed with a time step of 2^{-18}. Result are summarised in Fig. 1. One notices that our scheme is second order accurate in the setting presented in this paper ($K = K^{\text{reg}}$), but only first order accurate when confronted to what we call non-diffusive discontinuities, i.e., situations where the concentration of species that do not interdiffuse (i.e., $a_{i,j} = 0$) are discontinuous (here c_1 and c_2), and when the concentration of the specie (here c_3) which interacts with both discontinuous species is zero.

Our second test is two-dimensional. We choose K^{sing} as the diffusion matrix, $\kappa^{\star} = 0.1$, $\Omega = [0, 22] \times [0, 16]$, $\tau = 2^{-3}$ and a 2D initial condition $\mathbf{c}^0$ depicted in Fig. 2. The corresponding steady state and long-time limit $\mathbf{c}^{\infty}$ does not depend on $\mathbf{x}$, i.e., $c_i^{\infty}(\mathbf{x}) = \oint c_i^0(\mathbf{y})\mathrm{d}\mathbf{y}$ for all $\mathbf{x} \in \Omega$. The time evolution of the relative energy $\mathfrak{E}(\mathbf{c}) - \mathfrak{E}(\mathbf{c}^{\infty})$ is plotted on Fig. 3, showing exponential decay to the steady state even thought the diffusion matrix is singular. Snapshots showing the evolution of the concentration profiles are presented in Fig. 4.

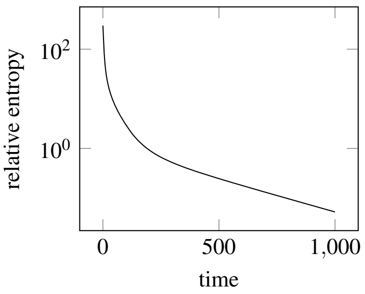
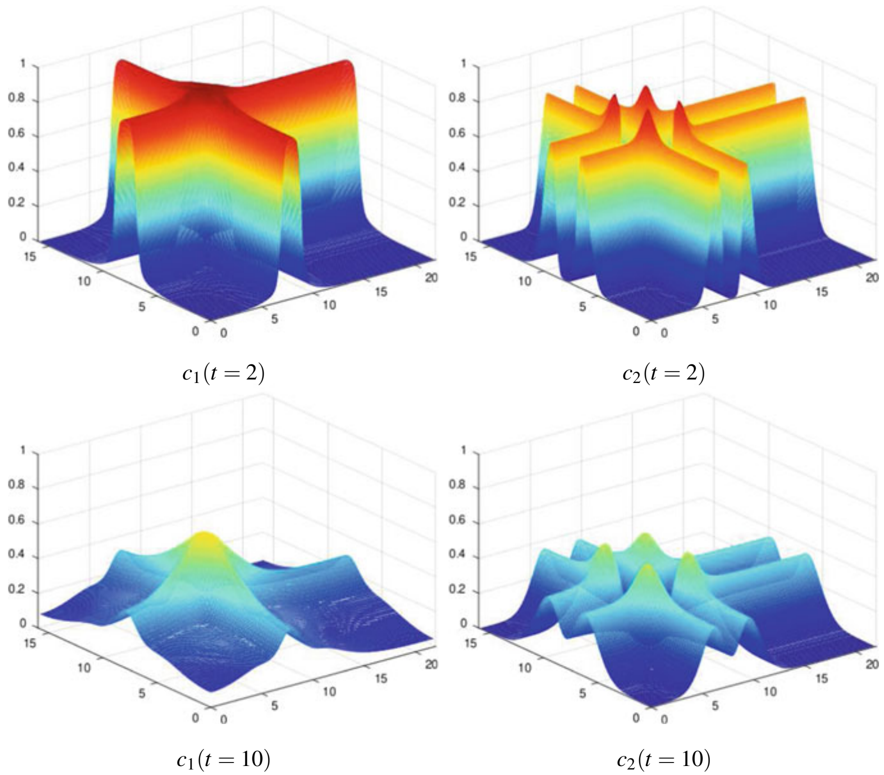


Fig. 3 $\mathfrak{E}(\mathbf{c}) - \mathfrak{E}^{\infty}$ as a function of time



$c_1(t = 2)$

$c_2(t = 2)$

$c_1(t = 10)$

$c_2(t = 10)$

Fig. 4 Concentrations c_1 and c_2 at times $t = 2$ and $t = 10$ (c_3 can be deduced from the relation $c_1 + c_2 + c_3 = 1$)

Acknowledgements The authors acknowledge the support of Labex CEMPI (ANR-11-LABX-0007). Clément Cancès also aknowledges support from the COMODO project (ANR-19-CE46-0002), and he warmly thanks Virginie Ehrlacher and Laurent Monasse for stimulating discussions.

References

1. Bakhta, A., Ehrlacher, V.: Cross-diffusion systems with non-zero flux and moving boundary conditions. ESAIM Math. Model. Numer. Anal. **52**(4), 1385–1415 (2018)
2. Berendsen, J., Burger, M., Ehrlacher, V., Pietschmann, J.F.: Uniqueness of strong solutions and weak-strong stability in a system of cross-diffusion equations. J. Evol. Equ.
3. Burger, M., Di Francesco, M., Pietschmann, J.F., Schlake, B.: Nonlinear cross-diffusion with size-exclusion. SIAM J. Math. Anal. **46**(6), 2842–2871 (2010)
4. Cancès, C., Gaudeul, B.: A convergent entropy diminishing finite volume scheme for a cross-diffusion system (2020). https://arxiv.org/abs/2001.11222
5. Eymard, R., Gallouët, T., Herbin, R.: Finite volume methods. In: Ciarlet, P.G. (ed.) Handbook of Numerical Analysis. North-Holland, Amsterdam, pp. 713–1020 (2000)
6. Jüngel, A.: The boundedness-by-entropy method for cross-diffusion systems. Nonlinearity **28**(6), 1963–2001 (2015)
7. Jüngel, A.: Entropy methods for diffusive partial differential equations. In: Springer Briefs in Mathematics. Springer [Cham] (2016)
8. Leray, J., Schauder, J.: Topologie et équations fonctionnelles. Ann. Sci. École Norm. Sup. **51**((3)), 45–78 (1934)

TPFA Finite Volume Approximation of Wasserstein Gradient Flows

Andrea Natale and Gabriele Todeschi

Abstract Numerous infinite dimensional dynamical systems arising in different fields have been shown to exhibit a gradient flow structure in the Wasserstein space. We construct Two Point Flux Approximation Finite Volume schemes discretizing such problems which preserve the variational structure and have second order accuracy in space. We propose an interior point method to solve the discrete variational problem, providing an efficient and robust algorithm. We present two applications to test the scheme and show its order of convergence.

Keywords Wasserstein gradient flows · Energy diminishing scheme

MSC (2010) 65M08 · 65M12 · 49M29 · 35K65 · 90C51

1 Gradient Flows' Time Discretization

A gradient flow is a process that, starting from an initial point, evolves by maximizing at each instant the rate of decay of a given specific energy. Many problems arising in physics, biology, social sciences, etc., can be recast as infinite dimensional gradient flows. Considering a compact domain $\Omega \subset \mathbb{R}^d$, a finite time horizon $T \in \mathbb{R}^+$, and a real-valued, strictly convex and proper energy functional $\mathscr{E}$, we focus our attention on problems of the form

A. Natale
Université Paris-Sud, 15 Rue Georges Clemenceau, 91400 Orsay, France
e-mail: andrea.natale@u-psud.fr

G. Todeschi (✉)
Inria, Université Paris-Dauphine, UMR CNRS 7534-Ceremade,
2 Rue Simone Iff, 75012 Paris, France
e-mail: gabriele.todeschi@inria.fr

R. Klöfkorn et al. (eds.), *Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples*, Springer Proceedings in Mathematics & Statistics 323,
https://doi.org/10.1007/978-3-030-43651-3_16

$$\begin{cases} \partial_t \rho - \nabla \cdot (\rho \nabla \frac{\delta \mathscr{E}}{\delta \rho}[\rho]) = 0, & \text{in } \Omega \times [0, T], \\ \rho \nabla \frac{\delta \mathscr{E}}{\delta \rho} \cdot \mathbf{n} = 0, & \text{on } \partial\Omega \times [0, T], \\ \rho(0) = \rho^0, & \text{in } \Omega, \end{cases} \tag{1}$$

where $\frac{\delta \mathscr{E}}{\delta \rho}$ denotes the first variation of $\mathscr{E}$, $\rho^0 \in L^1(\Omega; \mathbb{R}^+)$ is a given initial condition and $\mathbf{n}$ is the unit outer normal vector to $\partial\Omega$. Problem (1) denotes the continuity equation of a time evolving non-negative density ρ convected by the velocity field $-\nabla \frac{\delta \mathscr{E}}{\delta \rho}[\rho]$, with no flux across the boundary of the domain, hence preserving its total mass. It is nowadays clear that problems of the form of (1) represent gradient flows of the energy $\mathscr{E}$ with respect to the Wasserstein metric. We refer to [1, 9] for more details on gradient flows and optimal transport.

The underlying variational structure of this type of problems provides useful tools for their study. From the numerical point of view, more robust solvers can be designed by taking it into account. In particular, the property that the energy should decrease as fast as possible at each time step is a useful criterion to assess the goodness and reliability of a numerical solution and it should be preserved. The JKO scheme realizes this by using the variational formulation of the implicit Euler method. For an increasing sequence $(t^n)_{n\in\mathbb{N}} \subset \mathbb{R}$ of time steps such that $\cup_n [t^{n-1}, t^n] = [0, T]$, let $Q^n = \Omega \times [t^{n-1}, t^n]$ and $\partial Q^n = \partial\Omega \times [t^{n-1}, t^n]$. The JKO scheme constructs a sequence $(\rho^n)_{n\in\mathbb{N}}$ as follows: given an approximation ρ^{n-1} of the density at time t^{n-1}, compute $\rho^n = \tilde{\rho}(t^n)$, where $(\tilde{\rho}, \tilde{\mathbf{F}}) : Q^n \to \mathbb{R}^+ \times \mathbb{R}^d$ solves

$$\inf_{(\tilde{\rho}, \tilde{\mathbf{F}})} \int_{Q^n} \frac{|\tilde{\mathbf{F}}|^2}{2\tilde{\rho}} \mathrm{d}\mathbf{x}\mathrm{dt} + \mathscr{E}(\tilde{\rho}(t^n)), \quad \text{where } (\tilde{\rho}, \tilde{\mathbf{F}}) \text{ solve: } \begin{cases} \partial_t \tilde{\rho} + \nabla \cdot \tilde{\mathbf{F}} = 0, & \text{in } Q^n \\ \tilde{\mathbf{F}} \cdot \mathbf{n} = 0, & \text{on } \partial Q^n \\ \tilde{\rho}(t^{n-1}) = \rho^{n-1}. \end{cases} \tag{2}$$

The density ρ^n is computed minimizing the sum of its squared Wasserstein distance from ρ^{n-1} and the energy in ρ^n. The former term corresponds to the total kinetic energy of the curve $\tilde{\rho}$ written in the variables density-momentum, $(\tilde{\rho}, \tilde{\mathbf{F}})$, rather than density-velocity, in order to highlight the convexity of the problem [2]. The sequence of densities $(\rho^n)_{n\in\mathbb{N}}$, meant to be an approximation of the solution at each time step t^n, can be seen as a piecewise constant time-dependent density converging to the flow under suitable assumptions [1, 9]. This time discretization enables to design energy-diminishing schemes that are furthermore robust in the sense that, since (2) is a well-posed convex problem, the solution at step n always exists no matter the time step $\tau^n = t^n - t^{n-1}$.

The Wasserstein distance involved in (2) needs to be further discretized in time. Since the JKO scheme is of order one [6], a first order time discretization is sufficient and leads to a reasonable computational complexity. We can approximate (2) with an LJKO [3]: given an approximation ρ^{n-1} of the density at time t^{n-1}, compute ρ^n solution to

$$\inf_{(\rho,\mathbf{F})} \tau^n \int_\Omega \frac{|\mathbf{F}|^2}{2\rho} \mathrm{d}\mathbf{x} + \mathscr{E}(\rho), \quad \text{where } (\rho,\mathbf{F}) \text{ solve: } \begin{cases} \rho - \rho^{n-1} + \tau^n \, \nabla \cdot \mathbf{F} = 0, & \text{in } \Omega, \\ \mathbf{F} \cdot \mathbf{n} = 0, & \text{on } \partial\Omega, \end{cases} \tag{3}$$

where now $(\rho, \mathbf{F}) : \Omega \to \mathbb{R}^+ \times \mathbb{R}^d$ does not depend on time. The continuity equation is discretized using a single implicit Euler step, whereas the time integral using a right endpoint approximation.

Given the conservative form of the problem, Finite Volume methods appear as natural choices for its discretization. Their relation with optimal transport has been highlighted in, e.g., [5]. Ensuring the positivity of the density is a crucial property for any candidate numerical method, since problems (2) and (3) lose their convexity if the density is negative. In [3] problem (3) is discretized using upwind FV, which provides automatically the positivity for the discrete solution. The problem can then be solved using a Newton scheme. However, this gives an order one space discretization. Moreover, the derived scheme is not particularly robust since small time steps may be required to make the Newton scheme converge. In the present work we propose a more general FV framework, which allows us to consider second order discretizations in space. As a consequence, the positivity constraint on the density needs to be taken into account. To this end, we use an interior point method.

2 Finite Volume Discretization

Assume the domain $\Omega \subset \mathbb{R}^d$ to be polygonal if $d = 2$ or polyhedral if $d = 3$. The specifications for a partitioning of Ω to be admissible for TPFA Finite Volume are classical [4, Definition 9.1]. We denote by $\left(\mathscr{T}, \overline{\Sigma}, (\mathbf{x}_K)_{K\in\mathscr{T}}\right)$ such an admissible mesh, namely the triplet of the set of polyhedral control volumes, the set of faces and the set of cell centers. We use Delaunay triangulations in order to satisfy these assumptions. The Lebesgue measure of $K \in \mathscr{T}$ is denoted by $m_K > 0$. The set $\overline{\Sigma}$ is composed of boundary faces $\Sigma_{ext} = \{\sigma \subset \partial\Omega\}$ and internal faces $\sigma \in \Sigma = \overline{\Sigma} \setminus \Sigma_{ext}$. We denote by $\Sigma_K = \overline{\Sigma}_K \cap \Sigma$ the internal faces belonging to ∂K. For each internal face $\sigma = K|L \in \Sigma$, we refer to the diamond cell Δ_σ as the polyhedron whose edges join $\mathbf{x}_K$ and $\mathbf{x}_L$ to the vertices of σ. Denoting by m_σ the Lebesgue measure of the edge σ and by $d_\sigma = |\mathbf{x}_K - \mathbf{x}_L|$, the measure m_{Δ_σ} of Δ_σ is then equal to $m_\sigma d_\sigma / d$, where d stands for the space dimension. We denote by $d_{K,\sigma}$ the euclidean distance between the cell center $\mathbf{x}_K$ and the midpoint of the edge $\sigma \in \overline{\Sigma}_K$. The size of the mesh is defined by $h_\mathscr{T} = \max_{K\in\mathscr{T}} \operatorname{diam}(K)$.

We introduce the space of discrete conservative fluxes

$$\mathbb{F}_\mathscr{T} = \{\mathbf{F} = (F_{K,\sigma}, F_{L,\sigma})_{\sigma\in\Sigma} \in \mathbb{R}^{2\Sigma} : F_{K,\sigma} + F_{L,\sigma} = 0\}$$

and denote $F_\sigma = |F_{K,\sigma}| = |F_{L,\sigma}|$. We introduce also the spaces of discrete variables on cells $\mathbb{P}_\mathscr{T} = \mathbb{R}^\mathscr{T}$ and diamond cells $\mathbb{P}_\Sigma = \mathbb{R}^\Sigma$, endowed with the two scalar products $\langle\cdot,\cdot\rangle_K : (\mathbf{a}, \mathbf{b}) \in [\mathbb{P}_\mathscr{T}]^2 \mapsto \sum_{K\in\mathscr{T}} a_K b_K m_K$, $\langle\cdot,\cdot\rangle_\sigma : (\mathbf{u}, \mathbf{v}) \in [\mathbb{P}_\Sigma]^2 \mapsto \sum_{\sigma\in\Sigma}$

$u_\sigma v_\sigma m_\sigma d_\sigma$, respectively. We introduce a reconstruction operator from cells to diamond cells $R_\Sigma : \mathbb{P}_{\mathcal{T}} \to \mathbb{P}_\Sigma$. On each edge $\sigma = K|L$, the density on the diamond cell can be reconstructed from the values of the densities ρ_K, ρ_L. To keep the scheme simple, we employ weighted arithmetic averages $\rho_\sigma = \lambda_{K,\sigma}\rho_K + \lambda_{L,\sigma}\rho_L$, with $\lambda_{K,\sigma}, \lambda_{L,\sigma} \in [0, 1]$, $\lambda_{K,\sigma} + \lambda_{L,\sigma} = 1$. Nonetheless, other choices are possible, such as geometric, harmonic and logarithmic averages and all their weighted versions [5]. We consider three possibilities for the weights $(\lambda_{K,\sigma}, \lambda_{L,\sigma})$: $(\frac{1}{2}, \frac{1}{2})$, the standard arithmetic mean; $(\frac{d_{L,\sigma}}{d_\sigma}, \frac{d_{K,\sigma}}{d_\sigma})$, which provides a linear reconstruction of the density at the edge midpoint; $(\frac{d_{K,\sigma}}{d_\sigma}, \frac{d_{L,\sigma}}{d_\sigma})$, which gives a mass weighted arithmetic mean. Thanks to these choices we expect to obtain second order accuracy for the space discretization. We introduce also the adjoint operator of this reconstruction, with respect to the two scalar products, given by $R_{\mathcal{T}} : \boldsymbol{\rho} \in \mathbb{P}_\Sigma \mapsto \left(\sum_{\sigma\in\Sigma_K} \rho_\sigma \lambda_{K,\sigma} \frac{m_\sigma d_\sigma}{m_K}\right)_{K\in\mathcal{T}} \in \mathbb{P}_{\mathcal{T}}$.

Assuming the energy $\mathscr{E}(\rho)$ to be of the form $\int_\Omega E(\rho)\mathrm{d}\mathbf{x}$ for a real valued and strictly convex scalar function E, given the discrete initial density of the form $(\rho_K^0)_{K\in\mathcal{T}} = (\rho^0(\mathbf{x}_K))_{K\in\mathcal{T}} \in \mathbb{P}_{\mathcal{T}}^+$, the discrete LJKO scheme is: given $\boldsymbol{\rho}^{n-1} = (\rho_K^{n-1})_{K\in\mathcal{T}} \in \mathbb{P}_{\mathcal{T}}^+$ approximation of the density at time t^{n-1}, compute $\boldsymbol{\rho}^n$ solution to

$$\inf_{(\boldsymbol{\rho},\mathbf{F})} \tau^n \sum_{\sigma\in\Sigma} \frac{F_\sigma^2}{2(R_\Sigma(\boldsymbol{\rho}))_\sigma} m_\sigma d_\sigma + \sum_{K\in\mathcal{T}} E(\rho_K) m_K, \tag{4}$$

with $(\boldsymbol{\rho}, \mathbf{F}) \in \mathbb{P}_{\mathcal{T}} \times \mathbb{F}_{\mathcal{T}}$ such that $(\rho_K - \rho_K^{n-1})m_K + \tau^n \sum_{\sigma\in\Sigma_K} F_{K,\sigma} m_\sigma = 0$ and $\rho_K \geq 0, \forall K \in \mathcal{T}$. We take as measure of the diamond cell dm_{Δ_σ}, as it is classically done in order to compensate the unidirectional discretization of the momentum [4]. The constraint $\mathbf{F} \cdot \mathbf{n} = 0$ is automatically taken into account disregarding the flux on the boundary edges in the definition of the space of discrete conservative fluxes. The conservation of mass is also automatically enforced thanks to the conservativity of the Finite Volume discretization, i.e. $\sum_{K\in\mathcal{T}} \rho_K^n m_K = \sum_{K\in\mathcal{T}} \rho_K^{n-1} m_K$. Furthermore, the scheme guarantees a discrete energy-dissipation property: given the couple $(\boldsymbol{\rho}^n, \mathbf{F}^n)$ solution to (4), the competitor $(\boldsymbol{\rho}^{n-1}, \mathbf{0})$ provides

$$\tau^n \sum_{\sigma\in\Sigma} \frac{(F_\sigma^n)^2}{2(R_\Sigma(\boldsymbol{\rho}^n))_\sigma} m_\sigma d_\sigma + \sum_{K\in\mathcal{T}} E(\rho_K^n) m_K \leq \sum_{K\in\mathcal{T}} E(\rho_K^{n-1}) m_K.$$

At each step n, (4) is a strictly convex optimization problem with linear constraints. Enforcing the constraints with the multipliers $-\boldsymbol{\phi} \in \mathbb{P}_{\mathcal{T}}$, $\boldsymbol{\lambda} \in \mathbb{P}_{\mathcal{T}}^-$ and using the definition of the conservative fluxes we obtain the saddle point problem

$$\begin{aligned}\inf_{(\boldsymbol{\rho},\mathbf{F})} \sup_{(\boldsymbol{\phi},\boldsymbol{\lambda})} \tau^n \sum_{\sigma\in\Sigma} \frac{(F_\sigma)^2}{2(R_\Sigma(\boldsymbol{\rho}))_\sigma} m_\sigma d_\sigma + \sum_{K\in\mathcal{T}} (\rho_K^{n-1} - \rho_K)\phi_K m_K + \\ + \tau^n \sum_{\sigma\in\Sigma} F_{K,\sigma}\Big(\frac{\phi_L - \phi_K}{d_\sigma}\Big) m_\sigma d_\sigma + \sum_{K\in\mathcal{T}} E(\rho_K) m_K + \sum_{K\in\mathcal{T}} \lambda_K \rho_K m_K.\end{aligned} \tag{5}$$

The solution must satisfy the system of optimality conditions, namely the KKT conditions. Plugging the optimality condition w.r.t. $F_{K,\sigma}$, i.e. $F_{K,\sigma} = -(R_{\Sigma}(\boldsymbol{\rho}))_{\sigma}(\frac{\phi_L-\phi_K}{d_\sigma})$, in (5) and considering that

$$\sum_{\sigma\in\Sigma}(R_{\Sigma}(\rho^n))_{\sigma}\Big(\frac{\phi_L-\phi_K}{d_\sigma}\Big)^2 m_\sigma d_\sigma = \sum_{K\in\mathscr{T}}\rho_K\Big(R_{\mathscr{T}}\Big(\Big(\frac{\phi_L^n-\phi_K^n}{d_\sigma}\Big)^2\Big)\Big)_K m_K,$$

the optimality conditions reduce to the system

$$\begin{cases}(\rho_K^n-\rho_K^{n-1})m_K-\tau^n\sum_{\sigma\in\Sigma_K}(R_{\Sigma}(\rho^n))_{\sigma}(\frac{\phi_L^n-\phi_K^n}{d_\sigma})m_\sigma=0,\\ (\phi_K^n-E'(\rho_K^n)-\lambda_K^n)m_K+\frac{\tau^n}{2}(R_{\mathscr{T}}((\frac{\phi_L^n-\phi_K^n}{d_\sigma})^2))_K m_K=0, \qquad \forall K\in\mathscr{T}.\\ \rho_K^n\geq 0,\ \lambda_K^n\leq 0,\ \rho_K^n\lambda_K^n=0,\end{cases} \tag{6}$$

At each step n of the discrete LJKO, the discrete density $(\rho_K^n)_{K\in\mathscr{T}}$ is completely defined by (6).

System (6) is not easy to solve, the major problem being the non-uniqueness of the multipliers $\boldsymbol{\lambda}$ and $\boldsymbol{\phi}$ whenever the density vanishes. When upwinding is used for the reconstructed density, i.e. $\rho_\sigma=\rho_K$ if $\phi_L>\phi_K$, $\rho_\sigma=\rho_L$ otherwise, the Lagrange multiplier $\boldsymbol{\lambda}$ can be taken equal zero and disregarded [3]. In our framework this is not possible and to avoid dealing explicitly with the positivity constraint we use an interior point method. The constraint is incorporated in the problem by adding to the functional a barrier function of the density which is convex and singular in zero. We use the logarithmic barrier $-\log(\rho)$. In this way the minimizer is automatically repulsed away from zero and the problem can be solved using the Newton scheme. The perturbation introduced by the barrier function can be tuned by multiplying it by a positive coefficient μ. The perturbed version of problem (5) for the n-th step of the discrete LJKO is

$$\begin{aligned}\inf_{(\boldsymbol{\rho},\mathbf{F})}\sup_{\boldsymbol{\phi}}\tau^n\sum_{\sigma\in\Sigma}\frac{(F_\sigma)^2}{2(R_{\Sigma}(\boldsymbol{\rho}))_\sigma}m_\sigma d_\sigma+\sum_{K\in\mathscr{T}}(\rho_K^{n-1}-\rho_K)\phi_K m_K+\\ +\tau^n\sum_{\sigma\in\Sigma}F_{K,\sigma}\Big(\frac{\phi_L-\phi_K}{d_\sigma}\Big)m_\sigma d_\sigma+\sum_{K\in\mathscr{T}}E(\rho_K)m_K-\mu\sum_K\log(\rho_K)m_K,\end{aligned} \tag{7}$$

whose optimality conditions now are

$$\begin{cases}(\rho_K^n-\rho_K^{n-1})m_K-\tau^n\sum_{\sigma\in\Sigma_K}(R_{\Sigma}(\rho^n))_{\sigma}(\frac{\phi_L^n-\phi_K^n}{d_\sigma})m_\sigma=0,\\ (\phi_K^n-E'(\rho_K^n)+s_K)m_K+\frac{\tau^n}{2}(R_{\mathscr{T}}((\frac{\phi_L^n-\phi_K^n}{d_\sigma})^2))_K m_K=0, \qquad \forall K\in\mathscr{T},\\ s_K\rho_K=\mu,\end{cases} \tag{8}$$

where the condition $F_{K,\sigma}=-(R_{\Sigma}(\boldsymbol{\rho}))_{\sigma}(\frac{\phi_L-\phi_K}{d_\sigma})$ has been substituted again. System (8) can be seen as a perturbation of (6), where ρ_K and $s_K=-\lambda_K$ are automatically forced to be positive and the orthogonality is relaxed. For small value of μ it provides an approximation of the solution $(\boldsymbol{\rho}, \boldsymbol{\phi})$ to problem (6). However, the smaller the parameter the more difficult it is to solve problem (8) with a Newton scheme. The idea is then to construct a sequence of solutions to problem (8) for a sequence of coefficients μ decreasing to zero, using the solution corresponding to the previous value of μ as starting point for the Newton scheme. In this way the solver approaches the solution to (6) from the interior of the region of feasibility: the density is always positive. With reference to Algorithm 1, ε_0 and ε_μ are the tolerances for the solution to (6) and (8) respectively, δ_0 and δ_μ denoting a norm of the residues of the two systems of optimality conditions. In practice, it is not necessary to find for each value of μ a precise solution, being interested only in the solution for $\mu = 0$, and relatively big values can be used. Even doing only one Newton step, that is taking $\varepsilon_\mu = \infty$, can be sufficient and extremely effective. Moreover, the behavior of the solver strongly depends also on the initial value μ_0 and the decay ratio $\theta \in (0, 1)$, the difficulty to tune these parameters being its major drawback. We refer to [8] and references therein for more details on interior point methods.

Algorithm 1: Interior point method

```
Given the starting point $\mathbf{x}_0$ and the parameters $\mu_0 > 0, \theta \in (0,1), \varepsilon_0 > 0, \varepsilon_\mu > 0$ ;
while $\delta_0 > \varepsilon_0$ do
    $\mu = \theta\mu$ ;
    while $\delta_\mu > \varepsilon_\mu$ do
        compute Newton direction $\mathbf{d}$ for (8) and a step length $\alpha$;
        update: $\mathbf{x} = \mathbf{x} + \alpha\mathbf{d}$ ;
    end
end
```

As a final remark, note that solving the gradient flow with respect to an energy involving the entropy, i.e. $E(\rho) = \rho \log(\rho)$, automatically prevents the density from becoming negative. However, one cannot control the magnitude of the energy and therefore the interior point method, even if not strictly necessary, helps to get a more robust solver with respect to the Newton scheme. In fact, possible negative values for the density during the iterations of the algorithm could make it diverge, since the problem loses its convexity. The situation is similar when using the upwind technique to enforce the positivity.

3 Numerical Results

One of the most classical example of problems that exhibit a gradient flow structure is the Fokker-Planck equation:

$$\begin{cases} \partial_t \rho = \Delta\rho + \nabla \cdot (\rho \nabla V) & \text{in } \Omega \times [0, T], \\ (\nabla\rho + \rho\nabla V) \cdot \mathbf{n} = 0 & \text{on } \partial\Omega \times [0, T], \end{cases} \tag{9}$$

Table 1 Time-space convergence for the scheme

h_m	τ_m	ε_m^a	Rate	ε_m^b	Rate	ε_m^c	Rate
0.2986	0.0500	3.9382e−02	/	3.9526e−02	/	3.9157e−02	/
0.1493	0.0125	1.0345e−02	1.9286	1.0446e−02	1.9199	1.0246e−02	1.9342
0.0747	0.0031	2.6019e−03	1.9913	2.6367e−03	1.9861	2.5684e−03	1.9962
0.0373	0.0008	6.5090e−04	1.9990	6.6049e−04	1.9971	6.4170e−04	2.0009
0.0187	0.0002	1.6269e−04	2.0003	1.6519e−04	1.9994	1.6033e−04	2.0009

[a]Weights $(\frac{1}{2}, \frac{1}{2})$. [b]Weights $(\frac{d_L}{d_\sigma}, \frac{d_K}{d_\sigma})$. [c]Weights $(\frac{d_K}{d_\sigma}, \frac{d_L}{d_\sigma})$.

complemented with a positive initial condition, with $V \in W^{1,\infty}(\Omega)$ a Lipschitz continuous exterior potential. Equation (9) has been one of the first equations to be recasted as a gradient flow in the Wasserstein space with respect to the energy $\mathscr{E}(\rho) = \int_\Omega (\rho \log(\rho) + \rho V)\mathrm{d}\mathbf{x}$ [6]. This example gives us the possibility to test the convergence of scheme (4). Consider indeed the density $\rho_s(\mathbf{x}, t) = \exp(-(\pi^2 + \frac{g^2}{4})t + \frac{g}{2}x)(\pi \cos(\pi x) + \frac{g}{2} sin(\pi x)) + \pi \exp(g(x - \frac{1}{2}))$, which is a solution to (9) in the domain $[0, 1]^2 \times [0, 0.25]$ with potential $V(\mathbf{x}) = -gx$. Consider a sequence of meshes $\left(\mathscr{T}_m, \overline{\Sigma}_m, (\mathbf{x}_K)_{K\in\mathscr{T}_m}\right)$ with decreasing mesh size $h_m = h_{\mathscr{T}_m}$, and a sequence of decreasing time steps τ_m such that $(\frac{\tau_{m+1}}{\tau_m}) = (\frac{h_{m+1}}{h_m})^2$. We solve problem (9) with scheme (4) using this sequence of meshes and using as discrete initial condition $\rho_K^0 = \rho_s(\mathbf{x}_K, 0)$. For each solution we compute the mesh-dependent $L^1((0, T); L^1(\Omega))$ error $\varepsilon_m = \sum_n \tau_m \sum_{K\in\mathscr{T}_m} |\rho_K^n - \rho_s(\mathbf{x}_K, n\tau_m)| m_K$. In Table 1 are listed the errors for each m together with the convergence rate $\sqrt{\frac{\varepsilon_{m-1}}{\varepsilon_m}}$ for the three different weighted arithmetic averages. The scheme is first order accurate in time and second order accurate in space.

As second application, we consider a gradient flow of an energy which is not singular in zero. On the domain $\Omega = [-1.5, 1.5]^2$, for a time interval $[0, T]$, consider the porous medium equation,

$$\partial_t \rho = \Delta \rho^\gamma + \nabla \cdot (\rho \nabla V)$$

which has been proven in [7] to be a gradient flow in the Wasserstein space with respect to the energy $\mathscr{E}(\rho) = \int_\Omega \frac{1}{\gamma-1}\rho^\gamma + \rho V$, for a given γ strictly greater than one. We consider the confining potential $V(\mathbf{x}) = \frac{1}{2}||\mathbf{x}||_2^2$ which forces the density to concentrate at the origin. In (1) the evolution of an initial cross shaped density is shown for the case $\gamma = 2$. As expected, the solution converges towards the Barenblatt profile $\rho^\infty(\mathbf{x}) = \max((\frac{M}{2\pi})^{\frac{\gamma-1}{\gamma}} - \frac{\gamma-1}{2\gamma}||\mathbf{x}||^2, 0)^{\frac{1}{\gamma-1}}$, with M being the total mass of the initial condition (Figs. 1, 2).

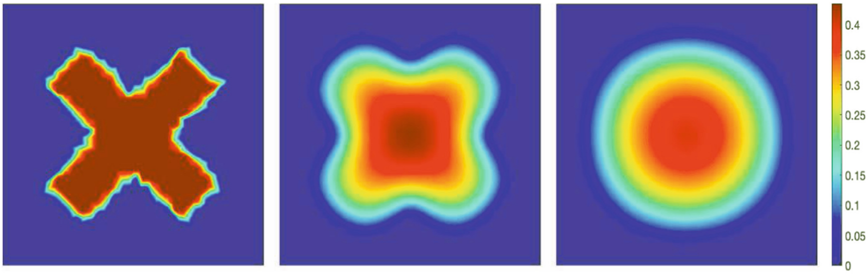


Fig. 1 Convergence towards the Barenblatt solution ($\gamma = 2$). Time steps $t = 0$, $t = 0.1$ and $t = 0.7$

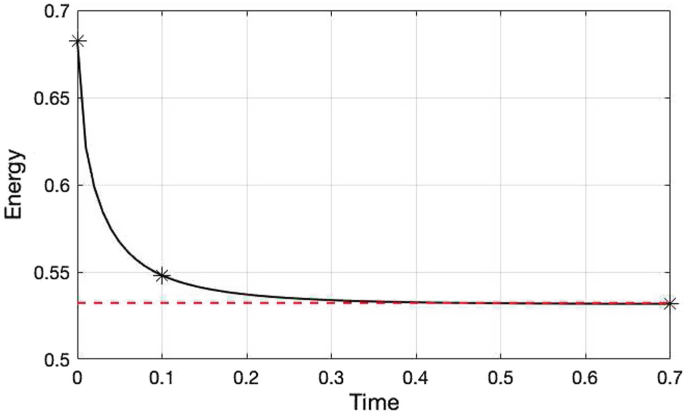


Fig. 2 Exponential decay profile of the discrete energy $\sum_{K \in \mathcal{T}} E(\rho_K) m_K$ (black), with the three values corresponding to Fig. 1, compared to the value of the energy for the Barenblatt equilibrium solution (red)

Acknowledgements GT acknowledges that this project has received funding from the European Unions Horizon 2020 research and innovation programme under the Marie Skodowska-Curie grant agreement No 754362.

References

1. Ambrosio, L., Gigli, N., Savaré, G.: Gradient flows in metric spaces and in the space of probability measures, 2nd edn. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel (2008)
2. Benamou, J.-D., Brenier, Y.: A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. Numer. Math. **84**(3), 375–393 (2000)
3. Cancès, C., Gallouët, T., Todeschi, G.: A variational finite volume scheme for Wasserstein gradient flows. https://arxiv.org/abs/1907.08305. Preprint
4. Eymard, R., Gallouët, T., Herbin, R.: Finite volume methods. In: Ciarlet, P.G. (ed.) Handbook of Numerical Analysis. North-Holland, Amsterdam, pp. 713–1020 (2000)
5. Gladbach, P., Kopfer, E., Maas, J.: Scaling limits of discrete optimal transport. https://arxiv.org/abs/1809.01092. Preprint
6. Jordan, R., Kinderlehrer, D., Otto, F.: The variational formulation of the Fokker-Planck equation. SIAM J. Math. Anal. **29**(1), 1–17 (1998)
7. Otto., F.: The geometry of dissipative evolution equations: the porous medium equation. Comm. Partial Differ. Eqn. **26**(1-2), 101–174 (2001)
8. Plik, I., Terlaky, T.: Interior point methods for nonlinear optimization. In: Di Pillo, G., Schoen, F. (eds.) Nonlinear Optimization. Springer Berlin Heidelberg, Berlin, Heidelberg, vol. 1989, pp. 215–276 (2010)
9. Santambrogio, F.: Optimal transport for applied mathematicians: calculus of variations, PDEs, and modeling. In: Progress in Nonlinear Differential Equations and Their Applications 87. Birkhäuser Basel, 1 edn (2015)

Free Energy Diminishing Discretization of Darcy-Forchheimer Flow in Poroelastic Media

Jakub W. Both, Jan M. Nordbotten, and Florin A. Radu

Abstract In this paper, we develop a discretization for the non-linear coupled model of classical Darcy-Forchheimer flow in deformable porous media, an extension of the quasi-static Biot equations. The continuous model exhibits a generalized gradient flow structure, identifying the dissipative character of the physical system. The considered mixed finite element discretization is compatible with this structure, which gives access to a simple proof for the existence, uniqueness, and stability of discrete approximations. Moreover, still within the framework, the discretization allows for the development of finite volume type discretizations by lumping or numerical quadrature, reducing the computational cost of the numerical solution.

Keywords Biot · Poroelasticity · Darcy-Forchheimer · Gradient flow · Mixed finite element method · Multipoint flux approximation · Mass lumping

MSC (2010) 65M08 · 65N12 · 65N30 · 76S05

1 Introduction

Flow in deformable porous media has been of increased interest in the recent past. Applications of societal and industrial relevance range from geotechnical to biomedical engineering, including the consolidation of the subsurface due to fluid production and the deformation of fluid-filled soft tissue.

Regarding slow viscous flow in linearly poroelastic media, the quasi-static linear Biot equations are often chosen as mathematical model, essentially coupling equations of linear elasticity and single-phase flow. For applications with significantly faster flow rates, Darcy's law is not further applicable. Instead the classical non-linear Darcy-Forchheimer law [8] is often utilized, cf., e.g., [2, 13].

J. W. Both (✉) · J. M. Nordbotten · F. A. Radu
University of Bergen, Postbox 7803, 5020 Bergen, Norway
e-mail: jakub.both@uib.no

R. Klöfkorn et al. (eds.), *Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples*, Springer Proceedings in Mathematics & Statistics 323,
https://doi.org/10.1007/978-3-030-43651-3_17

In this paper, we study the discretization of Darcy-Forchheimer flow in poroelastic media. The basis for this will be a (mixed) finite element method—widely used in the context of flow in porous media since being locally mass conservative. However, typically it suffers from larger algebraic systems to be solved, compared to, e.g., cell-centered finite volume discretizations. To circumvent this, for Darcy flow in non-deformable media, mass lumping [3] or approximate numerical quadrature techniques resulting in the (symmetric) multipoint flux mixed finite element method [17] have been developed allowing for local discrete flux pressure relationships. These are related to finite volume schemes employing respectively two-point and multipoint flux approximations [3, 11, 12]. Moreover, the resulting linear systems involve block-diagonal mass matrices, which allow for efficient solution. Recently, these techniques have been also applied in the context of deformable media [1, 10]. Regarding Darcy-Forchheimer flow in non-deformable media, especially the mixed finite element method on unstructured meshes [9, 14] and (similar to the above efforts) a block-centered finite difference method on rectangular grids [16] have been developed and studied in more detail including their well-posedness and theoretical convergence.

Motivated by all these advances, we propose a combination of the mixed finite element method and similar localization techniques in the context of Darcy-Forchheimer flow in deformable media. We particularly emphasize the inherent gradient flow structure of the continuous model, quantifying the dissipation of free energy over time. By construction, the numerical schemes considered here mimic a similar structure. Remarkably, it gives access to simple well-posedness and stability analyses of the numerical schemes.

The outline of the remaining paper is as follows. In Sect. 2, the mathematical model is described. In Sect. 3, the numerical method is presented, for which theoretical properties are discussed in Sect. 4.

Not part of this paper, but in the future, a numerical study will be conducted with focus on assessing the potential accuracy loss of the localization techniques, and efficiency gain regarding the algebraic solution. In particular, the exploitation of the block-diagonal nature of the flux mass term will be combined with robust splitting schemes as in [5, 6], benefiting from the linear character of the elasticity equations.

2 Model for Darcy-Forchheimer Flow in Poroelastic Media

The mathematical model couples the balance of linear momentum and the conservation of mass for a poroelastic medium, here modeled as an open, connected domain $\Omega \subset \mathbb{R}^d$, $d \in \{2, 3\}$. In addition, constitutive relations are considered: the medium is assumed to be linearly elastic and fully saturated with a slightly compressible fluid, with fluid flow described by the classical Darcy-Forchheimer law [8]. The solid-fluid interaction is governed by the so-called effective stress. Finally, the system of governing equations reads

$$-\boldsymbol{\nabla} \cdot (\mathbb{C}\boldsymbol{\varepsilon}(\boldsymbol{u}) - \alpha p\,\mathbf{I}) = \boldsymbol{f} \qquad \text{on } \Omega, \tag{1a}$$

$$\partial_t \left(\frac{1}{M} p + \alpha \boldsymbol{\nabla} \cdot \boldsymbol{u}\right) + \boldsymbol{\nabla} \cdot \boldsymbol{q} = 0 \qquad \text{on } \Omega, \tag{1b}$$

$$\mu \boldsymbol{\kappa}^{-1} \boldsymbol{q} + \rho\beta |\boldsymbol{q}| \boldsymbol{q} = -\boldsymbol{\nabla} p \qquad \text{on } \Omega, \tag{1c}$$

where the primal variables are the displacement $\boldsymbol{u} : \Omega \to \mathbb{R}^d$, the fluid pressure $p : \Omega \to \mathbb{R}$, and the volumetric flux $\boldsymbol{q} : \Omega \to \mathbb{R}^d$. Moreover, $\mathbb{C}$ denotes the (symmetric positive definite) fourth-order stiffness tensor, $\boldsymbol{\varepsilon}(\cdot)$ denotes the linear strain tensor, $\alpha \in (0, 1]$ is the Biot coefficient, $\boldsymbol{f}$ is a an external force, ∂_t denotes the derivative in time, $M \geq 0$ is the modulus accounting for the compressibility of the fluid and solid grains, $\mu > 0$ is the fluid viscosity, $\boldsymbol{\kappa}$ is the (symmetric positive definite) permeability tensor, $\rho > 0$ is the reference fluid density, and $\beta \geq 0$ is the Forchheimer index; the case $\beta = 0$ simplifies to Darcy's law. Ultimately, (1c) can be viewed as non-linear Darcy law with a direction dependent mobility.

For the sake of brevity, all material parameters are assumed to be constant. Furthermore, no external body or surface sources for the volume content are considered. However, we note that corresponding extensions are possible.

The system is closed with initial conditions $\theta(0) = \theta^\circ$ for the volume content $\theta = \frac{1}{M} p + \alpha \boldsymbol{\nabla} \cdot \boldsymbol{u}$ as well as boundary conditions for the displacement and flux. Here, for simplicity we choose $\boldsymbol{u}_{|\partial\Omega} = \mathbf{0}$ and $(\boldsymbol{q} \cdot \boldsymbol{n})_{|\partial\Omega} = 0$, where $\boldsymbol{n}$ denotes the outward normal on $\partial\Omega$.

2.1 The Gradient Flow Structure of the Model

As presented in [6], the weak formulation of model (1) exhibits a generalized gradient flow structure, in the sense of the lecture notes by Peletier [15]. Short, one can define the standard poroelastic Helmholtz free energy and a non-quadratic dissipation potential extending the classical potential corresponding to linear Darcy flow

$$\mathscr{E}(\boldsymbol{u}, \theta) = \frac{1}{2} \langle \mathbb{C}\boldsymbol{\varepsilon}(\boldsymbol{u}), \boldsymbol{\varepsilon}(\boldsymbol{u}) \rangle + \frac{M}{2} \|\theta - \alpha \boldsymbol{\nabla} \cdot \boldsymbol{u}\|^2_{L^2(\Omega)} - \langle \boldsymbol{f}, \boldsymbol{u} \rangle,$$

$$\mathscr{D}(\boldsymbol{q}) = \frac{\mu}{2} \left\langle \boldsymbol{\kappa}^{-1} \boldsymbol{q}, \boldsymbol{q} \right\rangle + \frac{\beta\rho}{3} \|\boldsymbol{q}\|^3_{L^3(\Omega)},$$

with $\|\cdot\|_{L^p(\Omega)}$ and $\langle \cdot, \cdot \rangle$ denoting the standard $L^p(\Omega)$ norm and $L^2(\Omega)$ scalar product, respectively. In the following, $H^1(\Omega)^d$ denotes the Sobolev space consisting of $L^2(\Omega)^d$ functions with weak derivatives in $L^2(\Omega)^{d\times d}$, and $H(\mathrm{div}; \Omega)$ requires solely the divergence to be in $L^2(\Omega)$. Moreover, let $\frac{\delta}{\delta(\cdot)}$ denote the Fréchet differential.

In the absence of dissipation of energy due to solid deformation, weak solutions to (1) are alternatively characterized by the degenerate generalized gradient flow

$$\boldsymbol{u} = \underset{\boldsymbol{v}\in H^1(\Omega)^d}{\arg\min}\ \mathscr{E}(\boldsymbol{v},\theta),$$

$$(\partial_t\theta, \boldsymbol{q}) = \underset{(s,\boldsymbol{w})\in L^2(\Omega)\times L^3(\Omega)^d\cap H(\mathrm{div};\Omega)}{\arg\min} \left\{ \mathscr{D}(\boldsymbol{w}) + \left\langle \frac{\delta\mathscr{E}}{\delta\theta}(\boldsymbol{u},\theta), s \right\rangle \right\}$$

$$\text{subj. to} \quad \begin{cases} \boldsymbol{v} = \boldsymbol{0} & \text{on } \partial\Omega, \\ s + \nabla\cdot\boldsymbol{w} = 0, & \text{on } \Omega, \\ \boldsymbol{w}\cdot\boldsymbol{n} = 0 & \text{on } \partial\Omega, \end{cases}$$

such that the flux $\boldsymbol{q}$ governs $\partial_t\theta$. The governing equations (1) can be recovered as the optimality conditions. The fluid pressure p enters as a Lagrange variable associated to mass conservation, as well as a dual variable $p = \frac{\delta\mathscr{E}}{\delta\theta}(\boldsymbol{u},\theta)$ and $p = \frac{\delta\mathscr{E}}{\delta\nabla\cdot\boldsymbol{u}}(\boldsymbol{u},\theta)$.

For sufficiently smooth solutions, employing the chain rule and the convexity of $\mathscr{D}$ yield the following energy–dissipation relation

$$\frac{d}{dt}\mathscr{E}(\boldsymbol{u},\theta) = \left\langle \frac{\delta\mathscr{E}(\boldsymbol{u},\theta)}{\delta(\boldsymbol{u},\theta)}, (\partial_t\boldsymbol{u}, \partial_t\theta) \right\rangle = -\langle p, \nabla\cdot\boldsymbol{q}\rangle = -\left\langle \frac{\delta\mathscr{D}}{\delta\boldsymbol{q}}(\boldsymbol{q}), \boldsymbol{q} \right\rangle \leq -\mathscr{D}(\boldsymbol{q}).$$

Remark 1 (*Incompressible case*) In the incompressible case, i.e., $M = \infty$, the energy degenerates and becomes merely sub-differentiable, as then (here for $\boldsymbol{f} = \boldsymbol{0}$)

$$\mathscr{E}(\boldsymbol{u},\theta) = \frac{1}{2}\langle\mathbb{C}\boldsymbol{\varepsilon}(\boldsymbol{u}), \boldsymbol{\varepsilon}(\boldsymbol{u})\rangle + \chi\left(\|\theta - \alpha\nabla\cdot\boldsymbol{u}\|_{L^2(\Omega)}\right), \quad \text{with} \quad \chi(p) = \begin{cases} 0, & p = 0, \\ \infty, & \text{else.} \end{cases}$$

3 Numerical Discretization

We present discretizations mimicking the free energy dissipating character of the continuous problem formulation. Focussing on the non-linear character of the problem, we limit the discussion to equidistant time-stepping and simplicial grids.

3.1 Semi-discrete Approximation in Variational Form

For the discretization in time, we utilize the minimizing movement scheme, which for (1) is equivalent with the implicit Euler method. Given a time interval of interest $[0, T]$, with $T > 0$ denoting the final time, we consider an equidistant partition $0 = t_0 < t_1 < \ldots < t_N = T$ with time step size τ. Furthermore, let $\mathscr{J}(\boldsymbol{u},\theta,\boldsymbol{q}) := \tau\mathscr{D}(\boldsymbol{q}) + \mathscr{E}(\boldsymbol{u},\theta)$. The discretization at time step $n \geq 1$ reads: *given* $\theta^{n-1} \in L^2(\Omega)$, *find* $(\boldsymbol{u}^n, \theta^n, \boldsymbol{q}^n) \in H^1(\Omega)^d \times L^2(\Omega) \times L^3(\Omega)^d \cap H(\mathrm{div};\,\Omega)$ *satisfying*

$$(\boldsymbol{u}^n, \theta^n, \boldsymbol{q}^n) = \underset{(\boldsymbol{u},\theta,\boldsymbol{q})}{\arg\min}\ \mathscr{J}(\boldsymbol{u}, \theta, \boldsymbol{q}),\ \text{subj. to}\ \begin{cases} \boldsymbol{u} = \boldsymbol{0}, & \text{on } \partial\Omega, \\ \frac{\theta-\theta^{n-1}}{\tau} + \nabla \cdot \boldsymbol{q} = 0, & \text{on } \Omega, \\ \boldsymbol{q} \cdot \boldsymbol{n} = 0, & \text{on } \partial\Omega. \end{cases}$$

3.2 Fully Discrete Approximation in Variational Form

For the discretization in space, we utilize the (mixed) finite element method with the possibility for reduced computational complexity using lumping or appropriate numerical quadrature. Starting with the semi-discrete formulation, we introduce a mesh-dependent version $\mathscr{J}_h$ of the original objective function $\mathscr{J}$ for this.

The mesh. Assume the physical domain Ω is polygonal and can be partitioned by simplices. Let $\mathscr{T}_h$ denote such a simplicial mesh with elements $K \in \mathscr{T}_h$ and faces $e \in \mathscr{F}_h$; let $\{\boldsymbol{r}_i\}_{i=1,..,d+1}$ denote the corners of $K \in \mathscr{T}_h$.

The finite element spaces. We consider classical conforming approximation spaces (including essential boundary conditions): $\mathscr{V}_h$, piecewise linear elements for the displacement; $\mathscr{Q}_h$, piecewise constant elements for the fluid content and the fluid pressure; and either $\mathscr{W}_h^{\mathrm{RT}}$, lowest order Raviart-Thomas, or $\mathscr{W}_h^{\mathrm{BDM}}$, lowest order Brezzi-Douglas-Marini elements, for the flux, depending on the subsequent choice for $\mathscr{J}_h$. If the particular choice is not crucial, we write $\mathscr{W}_h$ and allow for $\mathscr{W}_h^{\mathrm{RT}}$ and $\mathscr{W}_h^{\mathrm{BDM}}$. For detailed introduction of the finite element spaces, we refer to [4].

Objective function for MFEM. Instead of pursuing the Galerkin method for deriving fully discrete approximations, a mesh-dependent approximation of the objective function $\mathscr{J}$ is additionally considered

$$\mathscr{J}_h(\boldsymbol{u}_h, \theta_h, \boldsymbol{q}_h) := \tau \mathscr{D}_h(\boldsymbol{q}_h) + \mathscr{E}_h(\boldsymbol{u}_h, \theta_h).$$

The canonical mixed finite element discretization of (1) results in particular for

$$\begin{aligned} \mathscr{D}_h(\boldsymbol{q}_h) &:= \mathscr{D}(\boldsymbol{q}_h), \\ \mathscr{E}_h(\boldsymbol{u}_h, \theta_h) &:= \frac{1}{2} \langle \mathbb{C}\boldsymbol{\varepsilon}(\boldsymbol{u}_h), \boldsymbol{\varepsilon}(\boldsymbol{u}_h) \rangle + \frac{M}{2} \left\| \Pi_{\mathscr{Q}_h} (\theta_h - \alpha \nabla \cdot \boldsymbol{u}_h) \right\|_{L^2(\Omega)}^2, \end{aligned}$$

for $\boldsymbol{q}_h \in \mathscr{W}_h$ and $(\boldsymbol{u}_h, \theta_h) \in \mathscr{V}_h \times \mathscr{Q}_h$, where $\Pi_{\mathscr{Q}_h}$ denotes the $L^2(\Omega)$ projection onto $\mathscr{Q}_h$, allowing for measuring the fluid energy in the 'units' of the fluid volume. In the incompressible case ($M = \infty$), $\mathscr{E}_h$ is defined using an indicator function, as $\mathscr{E}$.

Definition of the method. Given a suitable approximation $\theta_h^0 \in \mathscr{Q}_h$ of the initial datum $\theta^0 \in L^2(\Omega)$, the fully discrete approximation at time step $n \geq 1$ reads: *given* $\theta_h^{n-1} \in \mathscr{Q}_h$*, find* $(\boldsymbol{u}_h^n, \theta_h^n, \boldsymbol{q}_h^n) \in \mathscr{V}_h \times \mathscr{Q}_h \times \mathscr{W}_h$ *satisfying*

$$(\boldsymbol{u}_h^n, \theta_h^n, \boldsymbol{q}_h^n) = \underset{(\boldsymbol{u}_h,\theta_h,\boldsymbol{q}_h)}{\arg\min}\ \tau \mathscr{D}_h(\boldsymbol{q}_h) + \mathscr{E}_h(\boldsymbol{u}_h, \theta_h) \tag{2a}$$

$$\text{subj. to } |K| \left(\theta_K - \theta_K^{n-1}\right) + \tau \int_K \nabla \cdot \boldsymbol{q}_h \, dx = 0 \quad \forall K \in \mathscr{T}_h. \tag{2b}$$

Saddle point formulation. The optimality conditions corresponding to (2) are obtained by introducing a Lagrange multiplier, which eventually can be identified as the discrete fluid pressure $p_h \in \mathscr{Q}_h$. We skip the calculations and directly present the final discrete system to be solved at time step $n \geq 1$: *given* $(\boldsymbol{u}_h^{n-1}, p_h^{n-1}) \in \mathscr{V}_h \times \mathscr{Q}_h$, *find* $(\boldsymbol{u}_h^n, p_h^n, \boldsymbol{q}_h^n) \in \mathscr{V}_h \times \mathscr{Q}_h \times \mathscr{W}_h$ *satisfying for all test functions* $(\boldsymbol{v}_h, \boldsymbol{w}_h) \in \mathscr{V}_h \times \mathscr{W}_h$ *and elements* $K \in \mathscr{T}_h$

$$\left\langle \mathbb{C}\boldsymbol{\varepsilon}(\boldsymbol{u}_h^n), \boldsymbol{\varepsilon}(\boldsymbol{v}_h) \right\rangle - \alpha \left\langle p_h^n, \nabla \cdot \boldsymbol{v}_h \right\rangle = \langle \boldsymbol{f}, \boldsymbol{v}_h \rangle, \tag{3a}$$

$$m(\boldsymbol{q}_h^n; \boldsymbol{q}_h^n, \boldsymbol{w}_h) - \left\langle p_h^n, \nabla \cdot \boldsymbol{w}_h \right\rangle = 0, \tag{3b}$$

$$\frac{|K|}{M}(p_K^n - p_K^{n-1}) + \alpha \int_K \nabla \cdot (\boldsymbol{u}_h^n - \boldsymbol{u}_h^{n-1}) \, dx + \tau \int_K \nabla \cdot \boldsymbol{q}_h^n \, dx = 0, \tag{3c}$$

where the non-linear form is given by $m(\boldsymbol{u}; \boldsymbol{v}, \boldsymbol{w}) = \int_\Omega \left(\mu \boldsymbol{\kappa}^{-1} + \rho\beta|\boldsymbol{u}|\mathbf{I}\right) \boldsymbol{v} \cdot \boldsymbol{w} \, dx$. The volume content at time step n can be recovered by $\theta_h^n = \frac{1}{M} p_h^n + \alpha \Pi_{\mathscr{Q}_h} \nabla \cdot \boldsymbol{u}_h^n$.

Localized dissipation potential. One drawback of the above formulation is the non-local relation between fluxes and pressure gradients. From a computational point of view, this results in an involved numerical solution. Instead, motivated by efforts for linear problems, we consider two choices: (i) mass lumping as in [3, 10] resulting in a two-point flux type approximation for $\boldsymbol{q}_h \in \mathscr{W}_h^{\mathrm{RT}}$

$$\mathscr{D}_h^{\mathrm{ML}}(\boldsymbol{q}_h) = \sum_{e \in \mathscr{F}_h} \omega_e \left[\frac{\mu}{2\kappa} \left| \int_e \boldsymbol{q}_h \cdot \boldsymbol{n}_e \, ds \right|^2 + \frac{\beta}{3} \left| \int_e \boldsymbol{q}_h \cdot \boldsymbol{n}_e \, ds \right|^3 \right]$$

where $\boldsymbol{n}_e$ is a uniquely defined normal on $e \in \mathscr{F}_h$ and ω_e is a suitable weight involving distances and measures of geometric entities [3]; and (ii) trapezoidal quadrature as in [17] resulting in a multipoint flux type approximation for $\boldsymbol{q}_h \in \mathscr{W}_h^{\mathrm{BDM}}$

$$\mathscr{D}_h^{\mathrm{Q}}(\boldsymbol{q}_h) = \sum_{K \in \mathscr{T}_h} \frac{|K|}{d+1} \sum_{i=1}^{d+1} \left[\frac{\mu}{2} \boldsymbol{\kappa}_{|K}^{-1}(\boldsymbol{r}_i) \boldsymbol{q}_{h|K}(\boldsymbol{r}_i) \cdot \boldsymbol{q}_{h|K}(\boldsymbol{r}_i) + \frac{\beta\rho}{3} \left| \boldsymbol{q}_{h|K}(\boldsymbol{r}_i) \right|^3 \right].$$

Lumping is only a sufficient approximation for scalar permeabilities [3]. Also one has to be aware of the fact that only the normal component of the flux contributes here, which is inconsistent with the constitutive Darcy-Forchheimer law.

The corresponding optimality conditions read as (3) but with an approximation m_h of m, which after suitable linearization results in a block-diagonal matrix. Thereby, fluxes may be explicitly represented in terms of pressure values at cell centers. This finally enables the development of efficient numerical solvers.

4 Existence, Uniqueness, and Stability

The scheme (3) satisfies local mass conservation at nonlinear and linearized level—independent of the particular choices for $\mathscr{W}_h$ and $\mathscr{D}_h$. In the following, we additionally emphasize a free energy dissipating character, naturally deduced from the inherited minimization structure (2). This structure in particular also guarantees unique solutions to (3). For the main theoretical result, we require consistency of the original and approximate dissipation potentials; motivated by [3, 17], we expect that there exists a broad class of grids and parameters for which the following holds.

Assumption 1 There exist constants $0 < c \leq C < \infty$ such that $c\,\mathscr{D}(\boldsymbol{q}_h) \leq \mathscr{D}_h(\boldsymbol{q}_h) \leq C\,\mathscr{D}(\boldsymbol{q}_h)$ for all $\boldsymbol{q}_h \in \mathscr{W}_h$.

Theorem 1 (Existence, uniqueness and stability) *Let $\mathscr{D}_h$ satisfy Assumption 1. Let $M < \infty$ and the remaining material parameters be as introduced in Sect. 2. In addition, let $m_h^0 \in \mathscr{Q}_h$ and $\boldsymbol{u}_h^0 \in \mathscr{V}_h$, then for all time steps $n \geq 1$, the numerical scheme (3) has a unique solution $(\boldsymbol{u}_h^n, p_h^n, \boldsymbol{q}_h^n) \in \mathscr{V}_h \times \mathscr{Q}_h \times \mathscr{W}_h$. It furthermore satisfies the following discrete energy dissipation inequality: for all $N > 0$ it holds*

$$\left\langle \mathbb{C}\boldsymbol{\varepsilon}(\boldsymbol{u}_h^N), \boldsymbol{\varepsilon}(\boldsymbol{u}_h^N) \right\rangle + \frac{1}{M} \left\| p_h^N \right\|^2 + \sum_{n=0}^{N} \tau \mathscr{D}_h(\boldsymbol{q}_h^n) \leq \left\langle \mathbb{C}\boldsymbol{\varepsilon}(\boldsymbol{u}_h^0), \boldsymbol{\varepsilon}(\boldsymbol{u}_h^0) \right\rangle + \frac{1}{M} \left\| p_h^0 \right\|^2 .$$

Proof The proof follows by induction with the induction step for $n \geq 1$ reading as follows. The existence and uniqueness result follows from classical convex analysis [7] applied to the minimization formulation (2).

The discrete function spaces $\mathscr{V}_h, \mathscr{Q}_h, \mathscr{W}_h$ are reflexive Banach spaces, equipped with natural norms: the $H^1(\Omega)$ semi-norm $\|\cdot\|_{\mathscr{V}} := |\cdot|_{H^1(\Omega)}$ on $\mathscr{V}_h$, the $L^2(\Omega)$ norm $\|\cdot\|_{\mathscr{Q}} := \|\cdot\|_{L^2(\Omega)}$ for $\mathscr{Q}_h$, and $\|\cdot\|_{\mathscr{W}}$ on $\mathscr{W}_h$, defined by

$$\|\boldsymbol{w}_h\|_{\mathscr{W}} := \|\boldsymbol{w}_h\|_{L^2(\Omega)} + \beta \|\boldsymbol{w}_h\|_{L^3(\Omega)} + \|\nabla \cdot \boldsymbol{w}_h\|_{L^2(\Omega)}, \qquad \boldsymbol{w}_h \in \mathscr{W}_h.$$

The objective function $\mathscr{J}_h$ is strictly convex on $\mathscr{V}_h \times \mathscr{Q}_h \times \mathscr{W}_h$, which follows directly from the definition and Assumption 1 ensuring the definiteness of $\mathscr{D}_h$; we emphasize that $\|\theta_h\|^2$ can be isolated by hiding the coupling term in the elastic energy. Furthermore, $\mathscr{J}_h$ is lower semi-continuous, and proper on

$$C := \left\{ (\boldsymbol{u}_h, \theta_h, \boldsymbol{q}_h) \in \mathscr{V}_h \times \mathscr{Q}_h \times \mathscr{W}_h \,\middle|\, \text{(2b) is satisfied} \right\}$$

since $\mathscr{J}_h(\boldsymbol{0}, \theta_h^{n-1}, \boldsymbol{0}) < \infty$ with $(\boldsymbol{0}, \theta_h^{n-1}, \boldsymbol{0}) \in C$. The feasible set C is by that not only convex and closed, but also non-empty. Lastly, again under Assumption 1, $\mathscr{J}_h$ is coercive over C; for this, we particularly note that $(\boldsymbol{u}_h, p_h, \boldsymbol{q}_h) \in C$ with $\|\nabla \cdot \boldsymbol{q}_h\|_{L^2(\Omega)} \to \infty$ implies $\|\theta_h\|_{L^2(\Omega)} \to \infty$, which eventually results in $J(\boldsymbol{u}_h, \theta_h, \boldsymbol{q}_h) \to \infty$. Ultimately, existence and uniqueness of a solution

$(\boldsymbol{u}_h^n, \theta_h^n, \boldsymbol{q}_h^n) \in C$ to (2) follows from classical convex analysis [7]. By comparing the solution with the feasible competitor $(\boldsymbol{u}_h^{n-1}, \theta_h^{n-1}, \mathbf{0}) \in C$, stability can be concluded

$$\mathscr{J}_h(\boldsymbol{u}_h^n, \theta_h^n, \boldsymbol{q}_h^n) = \mathscr{E}_h(\boldsymbol{u}_h^n, \theta_h^n) + \tau \mathscr{D}_h(\boldsymbol{q}_h^n) \leq \mathscr{E}_h(\boldsymbol{u}_h^{n-1}, \theta_h^{n-1}) = \mathscr{J}_h(\mathbf{0}, \theta_h^{n-1}, \mathbf{0}).$$

Existence, uniqueness, and stability of a solution $(\boldsymbol{u}_h^n, p_h^n, \boldsymbol{q}_h^n) \in \mathscr{V}_h \times \mathscr{Q}_h \times \mathscr{W}_h$ to the saddle point formulation (3) then follows from the equivalence of (2) and (3) and the identification $\theta_h^n = \frac{1}{M} p_h^n + \alpha \Pi_{\mathscr{Q}_h} \nabla \cdot \boldsymbol{u}_h^n$. The equivalence follows due to the well-known inf-sup stability of $\mathscr{W}_h \times \mathscr{Q}_h$ [4].

Remark 2 (*Incompressible case*) In the case of an incompressible medium, i.e., $M = \infty$, both problem formulations (2) (modified by the indicator function, cf. Rem. 1) and (3) are still equivalent, if also $\mathscr{V}_h \times \mathscr{Q}_h$ is inf-sup stable wrt. the divergence operator. This is not the case for $\mathscr{V}_h$ as defined above [4].

Remark 3 (*Convergence*) The convergence of the numerical approximation towards the continuous solution for decreasing mesh and time step size is a delicate subject—in particular regarding the localization techniques. For instance, mass lumping (similar to linear TPFA) is well-known to be prone for errors even in the linear case, e.g., for anisotropic permeability. A further study will be conducted in the future.

Acknowledgements This work is supported in part by the Research Council of Norway Project 250223, as well as the FracFlow project funded by Equinor through Akademiaavtalen.

References

1. Ambartsumyan, I., Khattatov, E., Nordbotten, J.M., Yotov, I.: A multipoint stress mixed finite element method for elasticity on simplicial grids. arXiv:1805.09920, accepted for SINUM
2. Atik, Y., Kabir, H., Vallet, G.: On a nonlinear system of BiotForchheimer type. Complex Var. Elliptic 1–23 (2019). https://doi.org/10.1080/17476933.2019.1695786
3. Baranger, J., Maitre, J.F., Oudin, F.: Connection between finite volume and mixed finite element methods. ESAIM-Math. Model Num. **30**(4), 445–465 (1996)
4. Boffi, D., Brezzi, F., Fortin, M.: Mixed Finite Element Methods and Applications, vol. 44. Springer (2013)
5. Both, J.W., Borregales, M., Nordbotten, J.M., Kumar, K., Radu, F.A.: Robust fixed stress splitting for Biot's equations in heterogeneous media. Appl. Math. Lett. **68**, 101–108 (2017)
6. Both, J.W., Kumar, K., Nordbotten, J.M., Radu, F.A.: The gradient flow structures of thermo-poro-visco-elastic processes in porous media. arXiv e-prints arXiv:1907.03134 (2019)
7. Ekeland, I., Temam, R.: Convex Analysis and Variational Problems, vol. 28. SIAM (1999)
8. Forchheimer, P.: Wasserbewegung durch boden. Z. Ver. Deutsch, Ing. **45**, 1782–1788 (1901)
9. Girault, V., Wheeler, M.F.: Numerical discretization of a Darcy-Forchheimer model. Numer. Math. **110**(2), 161–198 (2008)
10. Hu, X., Rodrigo, C., Gaspar, F.J., Zikatanov, L.T.: A nonconforming finite element method for the Biot's consolidation model in poroelasticity. J. Comput. Appl. Math. **310**, 143–154 (2017)
11. Klausen, R.A., Radu, F.A., Eigestad, G.T.: Convergence of MPFA on triangulations and for Richards' equation. Int. J. Numer. Meth. Fl. **58**(12), 1327–1351 (2008). https://doi.org/10.1002/fld.1787

12. Klausen, R.A., Winther, R.: Convergence of multipoint flux approximations on quadrilateral grids. Numer. Meth. Part D E **22**(6), 1438–1454 (2006)
13. Markert, B.: A constitutive approach to 3-d nonlinear fluid flow through finite deformable porous continua. Transport Porous Med. **70**(3), 427 (2007)
14. Park, E.J.: Mixed finite element methods for generalized Forchheimer flow in porous media. Numer. Meth. Part D E **21**(2), 213–228 (2005). https://doi.org/10.1002/num.20035
15. Peletier, M.A.: Variational Modelling: Energies, Gradient Flows, and Large Deviations. arXiv preprint arXiv:1402.1990 (2014)
16. Rui, H., Pan, H.: A block-centered finite difference method for the Darcy-Forchheimer model. SIAM J. Numer. Anal. **50**(5), 2612–2631 (2012). https://doi.org/10.1137/110858239
17. Wheeler, M.F., Yotov, I.: A multipoint flux mixed finite element method. SIAM J. Numer. Anal. **44**(5), 2082–2106 (2006)

Energy Stable Discretization for Two-Phase Porous Media Flows

Clément Cancès and Flore Nabet

Abstract We propose a $\mathbb{P}_1$ finite-element scheme with mass-lumping for a model of two incompressible and immiscible phases in a porous media flow. We prove the dissipation of the free energy and the existence of a solution to the nonlinear scheme. We also present numerical simulations to illustrate the behavior of the scheme.

Keywords Two-phase porous media flows · Energy stable finite-elements

MSC (2010) 65M60 · 65M12 · 35K65

1 Immiscible Two-Phase Flows in Porous Media

We are interested in the numerical approximations of the equations governing an immiscible incompressible two-phase flow in a porous medium. Let $\Omega \subset \mathbb{R}^d$ ($d = 2, 3$) be an open bounded polyhedral subset with Lipschitz boundary condition and let $t_{\mathrm{f}} > 0$ be an arbitrary finite time horizon. Then the conservation of the wetting (subscript w) and non-wetting phases (subscript n) are given by

$$\phi \partial_t s_\alpha - \nabla \cdot (\eta_\alpha(s_\alpha) \boldsymbol{\Lambda} \nabla p_\alpha) = q_\alpha(s_\alpha), \qquad \alpha \in \{\mathrm{w}, \mathrm{n}\}, \tag{1}$$

where the unknowns are the phase saturations s_α, which satisfy

$$s_{\mathrm{n}} + s_{\mathrm{w}} = 1, \tag{2}$$

C. Cancès
Inria, Univ. Lille, CNRS, UMR 8524 - Laboratoire Paul Painlevé, 59000 Lille, France
e-mail: clement.cances@inria.fr

F. Nabet (✉)
CMAP, Ecole Polytechnique, CNRS, I.P. Paris, 91128 Palaiseau, France
e-mail: flore.nabet@polytechnique.edu

R. Klöfkorn et al. (eds.), *Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples*, Springer Proceedings in Mathematics & Statistics 323,
https://doi.org/10.1007/978-3-030-43651-3_18

and the phase pressures p_α. The porosity $\phi \in (0, 1)$ is given, as well as the intrinsic permeability $\boldsymbol{\Lambda}$, which is assumed to be symmetric and uniformly elliptic. The mobility $\eta_\alpha : [0, 1] \to \mathbb{R}$ is assumed to be continuous and strictly increasing, with $\eta_\alpha(0) = 0$ and $\eta_\alpha(s) > 0$ if $s > 0$. They are extended to the whole $\mathbb{R}$ by $\eta_\alpha(s) = 0$ if $s < 0$ and $\eta_\alpha(s) = \eta_\alpha(1)$ if $s > 1$. The sources q_α are such that

$$q_\alpha(\mathbf{x}, s_\alpha) = q_{\text{inj}}(\mathbf{x}) \frac{\eta_\alpha(c_\alpha)}{\eta_{\text{w}}(c_{\text{w}}) + \eta_{\text{n}}(c_{\text{n}})} - q_{\text{sink}}(\mathbf{x}) \frac{\eta_\alpha(s_\alpha)}{\eta_{\text{w}}(s_{\text{w}}) + \eta_{\text{n}}(s_{\text{n}})}, \tag{3}$$

where $c_{\text{w}} \in (0, 1]$ and $c_{\text{n}} = 1 - c_{\text{w}}$ is the prescribed composition of the injected mixture, and where $q_{\text{inj}}, q_{\text{sink}} \in L^\infty(\Omega)$ are nonnegative, bounded, and such that $\int_\Omega q_{\text{inj}} = \int_\Omega q_{\text{sink}}$. The phase pressures are linked by the capillary pressure relation

$$p_{\text{n}} - p_{\text{w}} = \gamma(s_{\text{n}}), \tag{4}$$

where $\gamma \in L^1(0, 1)$ is strictly increasing, nonnegative, and blows up as s_{n} tends to 1. This function is extended for $s < 0$ by $\gamma(s) = \gamma(0) + 2s$. We further assume that $s \mapsto \eta_{\text{w}}(1 - s)\gamma(s) \in L^\infty(0, 1)$ and $s \mapsto \eta_{\text{w}}(1 - s)\gamma'(s) \in L^1(0, 1)$. These assumptions are satisfied by the usual models of the literature (see for instance [1]). The system is complemented with no-flux boundary conditions and initial conditions $s_\alpha^{\text{ini}} \in L^\infty(\Omega; [0, 1])$ that are compatible with (2). Note that since $\gamma \in L^1(0, 1)$, then $\Gamma : s \mapsto \int_0^s \gamma(a)\text{d}a$ is bounded on $[0, 1]$. The phase pressures being only defined up to a constant, we enforce additionally that $\int_\Omega p_{\text{n}} = 0$.

Multiplying (1) by p_α, summing over $\alpha \in \{\text{n}, \text{w}\}$, integrating over Ω, and using (2) and (4) yields

$$\frac{\text{d}}{\text{d}t} \int_\Omega \phi \Gamma(s_{\text{n}}) + \int_\Omega \sum_{\alpha \in \{\text{n},\text{w}\}} \eta_\alpha(s_\alpha) \boldsymbol{\Lambda} \nabla p_\alpha \cdot \nabla p_\alpha = \int_\Omega \sum_{\alpha \in \{\text{n},\text{w}\}} q_\alpha(s_\alpha) p_\alpha. \tag{5}$$

Following [6], we define the global pressure P by $P = p_{\text{n}} - r(s_{\text{n}})$ with $r : s_{\text{n}} \mapsto \int_0^{s_{\text{n}}} \frac{\eta_{\text{w}}(1-a)}{\eta_{\text{n}}(a)+\eta_{\text{w}}(1-a)} \gamma'(a)\text{d}a$. The definition of P yields

$$\sum_\alpha \eta_\alpha(s_\alpha) |\nabla p_\alpha|^2 = (\eta_{\text{n}}(s_{\text{n}}) + \eta_{\text{w}}(s_{\text{w}})) |\nabla P|^2 + \frac{\eta_{\text{n}}(s_{\text{n}})\eta_{\text{w}}(s_{\text{w}})}{\eta_{\text{n}}(s_{\text{n}}) + \eta_{\text{w}}(s_{\text{w}})} |\nabla \gamma(s_{\text{n}})|^2. \tag{6}$$

In view of the particular form (3) of the source terms,

$$\sum_{\alpha \in \{\text{n},\text{w}\}} q_\alpha(s_\alpha) p_\alpha \le \left(q_{\text{inj}} - q_{\text{sink}}\right) (P + r(s_{\text{n}})) + q_{\text{sink}} k(s_{\text{n}}), \tag{7}$$

with $k(s_{\text{n}}) = \frac{\eta_{\text{w}}(1-s_{\text{n}})}{\eta_{\text{w}}(1-s_{\text{n}})+\eta_{\text{n}}(s_{\text{n}})} \gamma(s_{\text{n}})$. Since $\eta_{\text{w}}(1 - \cdot)\gamma' \in L^1(0, 1)$ and $\eta_{\text{w}}(1 - \cdot)\gamma \in L^\infty(0, 1)$, both r and k are bounded on $(0, 1)$. Moreover, the extensions outside $(0, 1)$ of η_α and γ ensure that for all $\varepsilon > 0$, there exists C_ε such that

$$|s| + |k(s)| + |r(s)| \le \varepsilon \Gamma(s) + C_\varepsilon, \quad \forall s \in (-\infty, 1). \tag{8}$$

Combining (7) with (6) in (5) together with the uniform ellipticity of $\boldsymbol{\Lambda}$, $\eta_\mathrm{n}(s) + \eta_\mathrm{w}(1-s) \ge \delta > 0$ for all s, and (8) we get that

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega \phi \Gamma(s_\mathrm{n}) + \int_\Omega \left(|\nabla P|^2 + \sum_{\alpha \in \{\mathrm{n},\mathrm{w}\}} \eta_\alpha(s_\alpha)|\nabla p_\alpha|^2 \right) \le C. \tag{9}$$

This estimate is enough to establish the existence of a weak solution. In this paper, our goal is to show that this stability is still encoded in very natural numerical schemes. For the sake of simplicity, we present our analysis in the framework of $\mathbb{P}_1$ finite-elements with mass-lumping, but our approach can be extended to a wide family of schemes having the structure highlighted in [3, Sect. 3]. We show here how to transpose estimate (9) to the discrete setting and to infer the existence of discrete solutions therefrom. A full convergence study will be carried out in a forthcoming contribution. While deeply inspired from [7], the goal of this paper is to exploit more finely the energy estimate which allows to relax some stringent conditions on the anisotropy, on the mesh and the non-linearities presented in [7].

2 An Energy Stable Finite-Element Scheme

We study the problem (1)–(4) using a $\mathbb{P}_1$ conforming finite-element scheme with mass-lumping for the space discretization. Let $\mathscr{T}$ be a conforming simplicial discretization of Ω. We denote by $T \in \mathscr{T}$ a simplex, $\mathscr{V}_\mathscr{T}$ is the set of all the vertices $\mathbf{a}$ and $\mathscr{V}_T \subset \mathscr{V}_\mathscr{T}$ the set of the $(d+1)$ vertices $\mathbf{a}_0, \ldots, \mathbf{a}_d$ of the simplex T. We also denote by $V_h = \{u_h \in C(\overline{\Omega}) : u_h|_T \text{ is affine for all } T \in \mathscr{T}\}$ the usual conforming $\mathbb{P}_1$ finite-element space corresponding to the mesh $\mathscr{T}$ and by $(\varphi_\mathbf{a})_{\mathbf{a}\in\mathscr{V}_\mathscr{T}}$ the basis of V_h. In order to deal with the mass-lumping procedure, for any vertex $\mathbf{a} \in \mathscr{V}_\mathscr{T}$, we define the set $\mathfrak{s}_\mathbf{a}$, the boundary $\partial\mathfrak{s}_\mathbf{a}$ of which being defined by the hyperplanes joining the centers of mass of the simplices, edges (and faces if $d = 3$) sharing $\mathbf{a}$ as a vertex. We can now define the functional space $X_h := \{u \in L^\infty(\Omega) : u|_{\mathfrak{s}_\mathbf{a}} \text{ is constant for all } \mathbf{a} \in \mathscr{V}_\mathscr{T}\}$, and the linear mappings $\pi_X : C(\overline{\Omega}) \to X_h$ and $\pi_V : C(\overline{\Omega}) \to V_h$ by $\pi_\ell u(\mathbf{a}) = u(\mathbf{a})$, for any $\mathbf{a} \in \mathscr{V}_\mathscr{T}$, for any $u \in C(\overline{\Omega})$, $\ell = X, V$. In order to lighten the notations, for any $u_h \in V_h$ we write $\pi_X u_h = \overline{u}_h$. We will use the following Poincaré inequality that can be established as in [2]: there exist $C_1, C_2 > 0$ depending only on the mesh regularity such that for any $u_h \in V_h$,

$$\left\| \overline{u}_h - \frac{1}{|\Omega|}\int_\Omega u_h \right\|_{L^2(\Omega)} \le C_1 \left\| u_h - \frac{1}{|\Omega|}\int_\Omega u_h \right\|_{L^2(\Omega)} \le C_2 \|\nabla u_h\|_{L^2(\Omega)}. \tag{10}$$

Before detailing the numerical scheme, we have to define the discrete tensor field $\boldsymbol{\Lambda}_h : \Omega \to \mathbb{R}^{d\times d}$ almost everywhere by $\boldsymbol{\Lambda}_h(x) := \boldsymbol{\Lambda}_T := \frac{1}{|T|}\int_T \boldsymbol{\Lambda}$ if $x \in T$. From there, we define the matrix $\mathbf{A}_T := (\alpha^T_{i,j})_{1\le i,j\le d} \in \mathbb{R}^{d\times d}$ by

$$\alpha^T_{i,j} = \alpha^T_{j,i} := \int_T \boldsymbol{\Lambda}_T \nabla\varphi_{\mathbf{a}_i} \cdot \nabla\varphi_{\mathbf{a}_j}$$

and for any $u_h, v_h \in V_h$ one has,

$$\int_T \boldsymbol{\Lambda}_T \nabla u_h \cdot \nabla v_h = \begin{pmatrix} v_{\mathbf{a}_1} - v_{\mathbf{a}_0} \\ \dots \\ v_{\mathbf{a}_d} - v_{\mathbf{a}_0} \end{pmatrix} \cdot \mathbf{A}_T \begin{pmatrix} u_{\mathbf{a}_1} - u_{\mathbf{a}_0} \\ \dots \\ u_{\mathbf{a}_d} - u_{\mathbf{a}_0} \end{pmatrix}. \tag{11}$$

Following [4], we can prove that there exists $C_3 > 0$ depending on the regularity of the mesh and on the anisotropy ratio of $\boldsymbol{\Lambda}$ and $C_4 > 0$ depending, in addition, on d such that for any $T \in \mathscr{T}$ the matrix $\mathbf{A}_T$ satisfies

$$\operatorname{cond}_2(\mathbf{A}_T) \le C_3 \ \text{ and } \ \sum_{i=1}^{d} \left(\sum_{j=1}^{d} |\alpha^T_{i,j}| \right) (u_{\mathbf{a}_i})^2 \le C_4 \mathbf{u} \cdot \mathbf{A}_T \mathbf{u}, \ \forall \mathbf{u} = (u_{\mathbf{a}_i}) \in \mathbb{R}^d. \tag{12}$$

We are now in a position to give the numerical scheme using a backward Euler scheme for the time discretization. Let $(t^n)_{n=0,\dots,N}$ be a partition of the interval $[0, t_\text{f}]$ and for $n = 1, \dots, N$ we denote by $\tau_n = t^n - t^{n-1}$ the time step. We define the discrete initial data by $s^0_{\alpha,h} := \sum_{\mathbf{a}\in\mathscr{V}_\mathscr{T}} s^0_{\alpha,\mathbf{a}} \varphi_\mathbf{a} \in V_h$ with $s^0_{\alpha,\mathbf{a}} = \frac{1}{|\mathfrak{s}_\mathbf{a}|}\int_{\mathfrak{s}_\mathbf{a}} s^{\text{ini}}_\alpha$.

Let $s^{n-1}_\alpha \in V_h$ be given, we search for s^n_α, $p^n_\alpha \in V_h$ such that for any $v_{\alpha,h} \in V_h$ with $\alpha = (\text{n}, \text{w})$ one has,

$$\phi \int_\Omega \frac{\overline{s}^n_{\alpha,h} - \overline{s}^{n-1}_{\alpha,h}}{\tau_n} \overline{v}_{\alpha,h} + \int_\Omega \eta^n_{\alpha,h} \boldsymbol{\Lambda}_h \nabla p^n_{\alpha,h} \cdot \nabla v_{\alpha,h} = \int_\Omega \overline{q}_\alpha(\overline{s}^n_{\alpha,h}) \overline{v}_{\alpha,h}, \tag{13a}$$

$$s^n_{\text{n},h} + s^n_{\text{w},h} = 1, \tag{13b}$$

$$p^n_{\text{n},h} - p^n_{\text{w},h} = \gamma^n_{\text{n},h}, \tag{13c}$$

$$\int_\Omega p^n_{\text{n},h} = 0. \tag{13d}$$

We have denoted by $\eta^n_{\alpha,h} = \pi_V \eta(s^n_{\alpha,h})$, $\gamma^n_{\text{n},h} = \pi_V \gamma(s^n_{\text{n},h})$ and,

$$\overline{q}_\alpha(\overline{s}^n_{\alpha,h}) = \overline{q}_\text{inj} \frac{\eta_\alpha(c_\alpha)}{\eta_\text{w}(c_\text{w}) + \eta_\text{n}(c_\text{n})} - \overline{q}_\text{sink} \frac{\eta_\alpha(\overline{s}^n_{\alpha,h})}{\eta_\text{w}(\overline{s}^n_{\text{w},h}) + \eta_\text{n}(\overline{s}^n_{\text{n},h})}.$$

Mimicking the continuous case, we define the discrete global pressure and we can obtain the discrete counterpart of (6).

Proposition 1 *Let $s^n_{\alpha,h}$, $p^n_{\alpha,h} \in V_h$ be a solution to the scheme (13). Then there exists $C_5 > 0$ depending on the regularity of the mesh, on the anisotropy ratio of Λ, on δ and d such that*

$$\int_{\Omega} \boldsymbol{\Lambda}_h \nabla P_h \cdot \nabla P_h \le C_5 \left(\int_{\Omega} \eta^n_{\mathrm{n},h} \boldsymbol{\Lambda}_h \nabla p^n_{\mathrm{n},h} \cdot \nabla p^n_{\mathrm{n},h} + \int_{\Omega} \eta^n_{\mathrm{w},h} \boldsymbol{\Lambda}_h \nabla p^n_{\mathrm{w},h} \cdot \nabla p^n_{\mathrm{w},h} \right),$$

where $P^n_h = p^n_{\mathrm{n},h} - \pi_V r(s^n_{\mathrm{n},h}) \in V_h$.

Proof We define the functions

$$f_{\mathrm{n}}(s) = \frac{\eta_{\mathrm{n}}(s)}{\eta_{\mathrm{n}}(s) + \eta_{\mathrm{w}}(1-s)} \quad \text{and} \quad f_{\mathrm{w}}(s) = \frac{\eta_{\mathrm{w}}(1-s)}{\eta_{\mathrm{n}}(s) + \eta_{\mathrm{w}}(1-s)}.$$

Then, noting that $f_{\mathrm{n}} + f_{\mathrm{w}} = 1$ and using Eq. (13c), for any $T \in \mathscr{T}$ and for any vertices $\mathbf{a}_0, \mathbf{a}_i \in \mathscr{V}_T$, there exists $s^n_i \in [\min(\mathbf{a}_0, \mathbf{a}_i), \max(\mathbf{a}_0, \mathbf{a}_i)]$ such that,

$$P^n_{\mathbf{a}_0} - P^n_{\mathbf{a}_i} = f_{\mathrm{n}}(s^n_i)\left(p^n_{\mathrm{n},\mathbf{a}_0} - p^n_{\mathrm{n},\mathbf{a}_i}\right) - f_{\mathrm{w}}(s^n_i)\left(p^n_{\mathrm{w},\mathbf{a}_0} - p^n_{\mathrm{w},\mathbf{a}_i}\right).$$

Since η_α is strictly increasing, for any $T \in \mathscr{T}$ with $\mathbf{a}_0, \dots, \mathbf{a}_d$ as vertices

$$\eta^n_{\alpha,T} := \frac{1}{d+1} \sum_{i=0}^{d} \eta_\alpha(s^n_{\alpha,\mathbf{a}_i}) \ge \frac{1}{d+1} \max_{\mathbf{x}\in T} \eta_\alpha(\mathbf{x}) \ge \frac{1}{d+1} \eta_\alpha(s^n_i). \tag{14}$$

Thus using that $f_{\mathrm{n}}, f_{\mathrm{w}} \le 1$ and $\eta_{\mathrm{n}}(s) + \eta_{\mathrm{w}}(1-s) \ge \delta > 0$ we obtain,

$$\frac{\delta}{2(d+1)} \sum_{i=0}^{d} \left|P^n_{\mathbf{a}_0} - P^n_{\mathbf{a}_i}\right|^2 \le \eta^n_{\mathrm{n},T} \sum_{i=0}^{d} \left|p^n_{\mathrm{n},\mathbf{a}_0} - p^n_{\mathrm{n},\mathbf{a}_i}\right|^2 + \eta^n_{\mathrm{w},T} \sum_{i=0}^{d} \left|p^n_{\mathrm{w},\mathbf{a}_0} - p^n_{\mathrm{w},\mathbf{a}_i}\right|^2.$$

Since for any v_1, v_2, w satisfying $|v_1|^2 + |v_2|^2 \ge \mathrm{cond}_2(\mathbf{A}_T)|w|^2$ one has

$$v_1 \cdot \mathbf{A}_T v_1 + v_2 \cdot \mathbf{A}_T v_2 \ge w \cdot \mathbf{A}_T w,$$

we use equality (11) associated with the fact that the condition number of $\mathbf{A}_T$ is bounded, cf. (12). Then summing the resulting estimate over $T \in \mathscr{T}$ and noting that the Lagrange vertex-quadrature formula is exact on $\mathbb{P}_1$ (see [5, Remark 2.2]) we obtain the claim. □

Proposition 2 *Let* $s^{n-1}_{\alpha,h} \in V_h$ *be given and* $s^n_\alpha, p^n_\alpha \in V_h$ *be a solution to the scheme (13). There exists* $C_6 > 0$ *depending on the data of the continuous problem but neither on the mesh* $\mathscr{T}$ *or nor the time step* τ_n *such that,*

$$\phi \int_{\Omega} \Gamma(\overline{s}^n_{\mathrm{n},h}) + \tau_n \sum_{\alpha\in\{\mathrm{n},\mathrm{w}\}} \int_{\Omega} \eta^n_{\alpha,h} \boldsymbol{\Lambda}_h \nabla p^n_{\alpha,h} \cdot \nabla p^n_{\alpha,h} + \tau_n \int_{\Omega} \nabla P^n_h \cdot \nabla P^n_h$$
$$\le C_6 \left(1 + \phi \int_{\Omega} \Gamma(\overline{s}^{n-1}_{\mathrm{n},h})\right).$$

Proof Let us choose $v_{\alpha,h} = p^n_{\alpha,h}$ as test function in Eq. (13a) and then add the resulting equations. Then, since Γ is convex, thanks to relation (13c) we obtain

$$\phi \int_\Omega \Gamma(\overline{s}^n_{\mathrm{n},h}) + \sum_{\alpha\in\{\mathrm{n},\mathrm{w}\}} \tau_n \int_\Omega \eta^n_{\alpha,h} \boldsymbol{\Lambda}_h \nabla p^n_{\alpha,h} \cdot \nabla p^n_{\alpha,h} \\ \le \phi \int_\Omega \Gamma(\overline{s}^{n-1}_{\mathrm{n},h}) + \tau_n \int_\Omega \sum_{\alpha\in\{\mathrm{n},\mathrm{w}\}} \overline{q}_\alpha(\overline{s}^n_{\alpha,h})\overline{p}^n_{\alpha,h}. \quad (15)$$

As for the continuous case, one has

$$\sum_{\alpha\in\{\mathrm{n},\mathrm{w}\}} \overline{q}_\alpha(\overline{s}^n_{\alpha,h})\overline{p}^n_{\alpha,h} \le \left(\overline{q}_{\mathrm{inj}} - \overline{q}_{\mathrm{sink}}\right)\left(\overline{P}^n_h + r(\overline{s}^n_{\mathrm{n},h})\right) + \overline{q}_{\mathrm{sink}} k(\overline{s}^n_{\mathrm{n},h}). \quad (16)$$

Using the definition of the discrete global pressure P^n_h and Eq. (13d), combined with the discrete Poincaré inequality (10) and (8) give

$$\|\overline{P}^n_h\|_{L^1(\Omega)} \le C_2|\Omega|^{1/2}\|\nabla P^n_h\|_{L^2(\Omega)} + \int_\Omega \left|r(\overline{s}^n_{\mathrm{n},h})\right| \\ \le \varepsilon\|\nabla P^n_h\|^2_{L^2(\Omega)} + \varepsilon \left\|\Gamma(\overline{s}^n_{\mathrm{n},h})\right\|_{L^1(\Omega)} + C_\varepsilon. \quad (17)$$

Since $q_{\mathrm{inj}}, q_{\mathrm{sink}} \in L^\infty(\Omega)$, the use of the above inequality and of (8) in (16) leads to

$$\int_\Omega \sum_{\alpha\in\{\mathrm{n},\mathrm{w}\}} \overline{q}_\alpha(\overline{s}^n_{\alpha,h})\overline{p}^n_{\alpha,h} \le \varepsilon\|\nabla P^n_h\|^2_{L^2(\Omega)} + \varepsilon \left\|\Gamma(\overline{s}^n_{\mathrm{n},h})\right\|_{L^1(\Omega)} + C_\varepsilon \quad (18)$$

whatever $\varepsilon > 0$. Using (18) together with Proposition 1 in (15) provides the expected bound. □

Thanks to Eqs. (13b) and (13c) we see that the saturations and the pressures of the wetting and non-wetting phases are linked. Thus we can choose the pressure of the wetting phase and the capillary pressure as main unknowns. Choosing $v_{\alpha,h} = \varphi_{\mathbf{a}}$ as test functions in Eq. (13a) we can rewrite the scheme (13) as a nonlinear system of $2\#\mathscr{V}_\mathscr{T}$ algebraic equations $\mathscr{F}^n((\gamma(s^n_{\mathrm{n},\mathbf{a}}), p^n_{\mathrm{w},\mathbf{a}})_{\mathbf{a}\in\mathscr{V}_\mathscr{T}}) = 0$. Since $\gamma(1) = +\infty$, the function $\mathscr{F}^n$ is continuous but non uniformly continuous. However, we prove in the following lemma that this situation is avoided for a solution to the scheme (13).

Proposition 3 *Let $s^{n-1}_{\alpha,h} \in V_h$ be such that $\int_\Omega \overline{s}^{n-1}_{\mathrm{w},h} \ge 0$ and $s^n_{\alpha,\mathrm{h}}, p^n_{\alpha,\mathrm{h}} \in V_h$ be a solution the scheme (13). There exists $\sigma_{\tau_n,\mathscr{T}}, \varepsilon_{\tau_n,\mathscr{T}} > 0$ depending on the data of the continuous problem, $\mathscr{T}$, τ_n and $s^{n-1}_{\mathrm{n},h}$ such that,*

$$-\sigma_{\tau_n,\mathscr{T}} \le s^n_{\mathrm{n},\mathbf{a}} \le 1 - \varepsilon_{\tau_n,\mathscr{T}}, \quad \forall \mathbf{a} \in \mathscr{V}_\mathscr{T}.$$

Proof First of all, thanks to the extension of γ for $s < 0$, the energy estimate given in Proposition 2 yields $\int_\Omega ((\overline{s}^n_{\mathrm{n},h})^-)^2 \le C^{n-1}$, which provides the lower bound.

Then we prove a bound on the pressure of the non-wetting phase $p_{\mathrm{n},h}$. Thanks to inequality (17) and the definition of P^n_h one has

$$\|\overline{p}^n_{\mathrm{n},h}\|_{L^1(\Omega)} \le C^{n-1} \quad \Rightarrow \quad |p^n_{\mathrm{n},\mathbf{a}}| \le \frac{C^{n-1}}{|\mathfrak{s}_\mathbf{a}|}, \ \forall \mathbf{a} \in \mathscr{V}_\mathscr{T}. \tag{19}$$

Now, let us note that proving the upper bound is equivalent to proving that there exists $\gamma^\star_{\tau_n,\mathscr{T}}$ such that for any $\mathbf{a} \in \mathscr{V}_\mathscr{T}$, $\gamma(s_{\mathrm{n},\mathbf{a}}) \le \gamma^\star_{\tau_n,\mathscr{T}}$.

We choose $v_{\mathrm{w},h} = 1$ as test function in Eq. (13a), then since q_{inj} is nonnegative, $\eta_\mathrm{n}(s) + \eta_\mathrm{w}(1-s) \ge \delta > 0$ and $c_\mathrm{w} > 0$ (and so $\eta_\mathrm{w}(c_\mathrm{w}) > 0$), one has

$$s^n_{\mathrm{w},\mathbf{a}_\mathrm{i}} \ge \frac{1}{|\Omega|}\int_\Omega \overline{s}^n_{\mathrm{w},h} > \frac{1}{|\Omega|}\left(\int_\Omega \overline{s}^{n-1}_{\mathrm{w},h} - \frac{\tau_n}{\delta}\|q_{\mathrm{sink}}\|_{L^\infty(\Omega)}\int_\Omega \eta_\mathrm{w}(\overline{s}^n_{\mathrm{n},h})\right).$$

Note that we proved here by induction that $\int_\Omega \overline{s}^n_{\mathrm{w},h} \ge 0$. Since $s \mapsto (s + \eta_\mathrm{w}(s))^{-1}$ is Lipschitz, there exists $\mathbf{a}_\mathrm{i} \in \mathscr{V}_\mathscr{T}$ such that $s^n_{\mathrm{w},\mathbf{a}_\mathrm{i}} > 0$ that is there exists $\mathbf{a}_\mathrm{i} \in \mathscr{V}_\mathscr{T}$ such that $s^n_{\mathrm{n},\mathbf{a}_\mathrm{i}} < 1$.

Let $\mathbf{a}_\mathrm{f} \in \mathscr{V}_\mathscr{T}$ be arbitrary and $(\mathbf{a}_q)_{q=0,\cdots,\ell}$ be a path from $\mathbf{a}_\mathrm{i}$ to $\mathbf{a}_\mathrm{f}$. Let $q \in \{0,\ldots,\ell-1\}$. Using the property (12) of the matrix $\mathbf{A}_T$ and since the quadrature formula is exact on $\mathbb{P}_1$, Proposition 2 gives

$$\sum_{T\in\mathscr{T}} \eta^n_{\mathrm{w},T} \sum_{i=1}^d \left(\sum_{j=1}^d |\alpha^T_{i,j}|\right) \left(p^n_{\mathrm{w},h}(\mathbf{a}_i) - p^n_{\mathrm{w},h}(\mathbf{a}_0)\right)^2 \le \frac{C_4C_6}{\tau_n}\left(1 + \phi\int_\Omega \Gamma(\overline{s}^{n-1}_{\mathrm{n},h})\right).$$

We assume by induction that there exists $\varepsilon_{\tau_n,\mathscr{T}} > 0$ such that $s^n_{\mathrm{n},\mathbf{a}_q} < 1 - \varepsilon_{\tau_n,\mathscr{T}}$ that is $s^n_{\mathrm{w},\mathbf{a}_q} > \varepsilon_{\tau_n,\mathscr{T}}$. Thus, if T is a simplex with $\mathbf{a}_q$, $\mathbf{a}_{q+1}$ as vertices, the definition (14) of $\eta^n_{\mathrm{w},T}$ yields $\eta^n_{\mathrm{w},T} \ge \frac{\eta(s^n_{\mathrm{w},\mathbf{a}_q})}{d+1} \ge \varepsilon'_{\tau_n,\mathscr{T}}$. Thanks to Eqs. (13c) and (19) it follows that,

$$\left|\gamma(s^n_{\mathrm{n},\mathbf{a}_q}) - \gamma(s^n_{\mathrm{n},\mathbf{a}_{q+1}})\right| - \left|p^n_{\mathrm{n},\mathbf{a}_q} - p^n_{\mathrm{n},\mathbf{a}_{q+1}}\right| \le C_{\tau_n,\mathscr{T}} \quad \Rightarrow \quad \gamma(s^n_{\mathrm{n},\mathbf{a}_{q+1}}) \le \gamma^{\star\star}_{\tau_n,\mathscr{T}}.$$

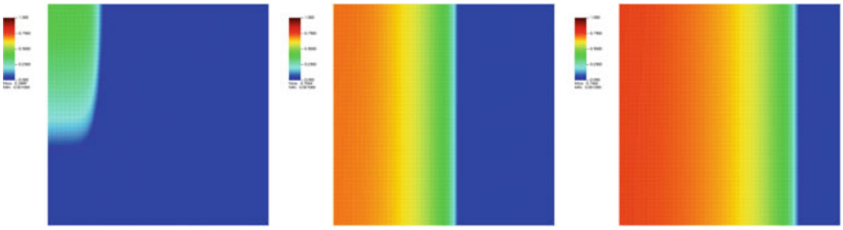
We conclude the proof by induction along the path. □

The bound on the saturation associated with the definition (14) on $\eta^n_{\mathrm{w},T}$ yields $\eta^n_{\mathrm{w},T} \ge \eta_\mathrm{w}(\varepsilon_{\tau_n,\mathscr{T}})$. This, combined with the Poincaré inequality (10) and since $\gamma(s^n_{\mathrm{n},\mathbf{a}}) \le \gamma(1 - \varepsilon_{\tau_n,\mathscr{T}})$ for any $\mathbf{a} \in \mathscr{V}_\mathscr{T}$, allows us to obtain a discrete bound on the pressure.

Proposition 4 *There exists $p^\star_{\tau_n,\mathscr{T}} > 0$ depending on the data of the continuous problem, $\mathscr{T}$, τ_n and $s^{n-1}_{\mathrm{n},h}$ such that $\int_\Omega |p^n_{\mathrm{w},h}|^2 \le p^\star_{\tau_n,\mathscr{T}}$.*

Thanks to the material introduced above, it is possible to prove the existence of a solution to the discrete problem using the topological degree theory.

Theorem 1 (Existence of a solution) *Let $s^{n-1}_{\mathrm{n},h} \in V_h$ be given, there exists at least one solution to the scheme (13).*



(a) $t = 0.002$ (b) $t = 0.01$ (c) $t = 0.015$

Fig. 1 Approximate saturation $s_{\mathrm{n},h}$ in Ω for different times t

3 Numerical Results

We present here numerical results obtained with the software FreeFem [8] in the two-dimension case by choosing as main unknowns the saturation of the non-wetting phase and the pressure of the wetting phase. To solve the nonlinear system we use a Newton method with a stopping criteria on the ℓ^∞-norm between two successive iterations. The computational domain is the unit square $\Omega = [0, 1]^2$ and the mesh is made up of triangles whose mesh size is approximately equal to 0.028. The final time is $t_\mathrm{f} = 0.015$ and the time step is constant $\tau_n = 10^{-3}$. We choose the porosity $\phi = 0.3$, the permeability tensor field $\boldsymbol{\Lambda} = \begin{pmatrix} 1 & 0 \\ 0 & 100 \end{pmatrix}$ and $c_\mathrm{w} = 0.2$. For $s \in [0, 1]$ we define the mobility functions by $\eta_\mathrm{n}(s) = s^2$ and $\eta_\mathrm{w}(s) = 2s$, the capillary pressure by $\gamma(s) = \frac{1}{\sqrt{1-s}}$ and the source functions are defined by $q_\mathrm{inj} = 40.1_{[0,0.2]\times[0.8,1]}$ and $q_\mathrm{sink} = 40.1_{[0.8,1]\times[0,0.2]}$. We plot in Fig. 1 the approximate saturation of the non-wetting phase.

One observes from the outset of the simulation the influence on the injection well q_inj and of the anisotropy ratio in the longitudinal direction. Moreover we can see that the maximum does not exceed $c_\mathrm{n} = 0.8$.

Acknowledgements This work was supported by the French National Research Agency (ANR) through Grant ANR-13-JS01-0007-01 (GEOPOR project). Clément Cancès acknowledges the support of Labex CEMPI (ANR-11-LABX-0007).

References

1. Bear, J., Bachmat, Y.: Introduction to Modeling of Transport Phenomena in Porous Media. Kluwer Academic Publishers, Dordrecht, The Netherlands (1990)
2. Brenner, K., Masson, R.: Convergence of a vertex centered discretization of two-phase darcy flows on general meshes. Int. J. Finite **10**, 1–37 (2013)
3. Cancès, C.: Energy stable numerical methods for porous media flow type problems. Oil Gas Sci. Technol.-Rev. IFPEN **73**, 1–18 (2018)

4. Cancès, C., Guichard, C.: Numerical analysis of a robust free energy diminishing finite volume scheme for parabolic equations with gradient structure. Found. Comput. Math. **17**(6), 1525–1584 (2017)
5. Cancès, C., Nabet, F., Vohralík, M.: Convergence and a posteriori error analysis for energy-stable finite element approximations of degenerate parabolic equations (2018). HAL: hal-01894884
6. Chavent, G., Jaffré, J.: Mathematical Models and Finite Elements for Reservoir Simulation, vol. 17, stud. math. appl. edn. North-Holland, Amsterdam (1986)
7. Eymard, R., Herbin, R., Michel, A.: Mathematical study of a petroleum-engineering scheme. M2AN Math. Model. Numer. Anal. **37**(6), 937–972 (2003)
8. Hecht, F.: New development in FreeFem++. J. Numer. Math. **20**(3–4), 251–265 (2012)

A Finite-Volume Scheme for a Cross-Diffusion Model Arising from Interacting Many-Particle Population Systems

Ansgar Jüngel and Antoine Zurek

Abstract A finite-volume scheme for a cross-diffusion model arising from the mean-field limit of an interacting particle system for multiple population species is studied. The existence of discrete solutions and a discrete entropy production inequality is proved. The proof is based on a weighted quadratic entropy that is not the sum of the entropies of the population species.

Keywords Finite volume scheme · Cross-diffusion system · Entropy method

MSC (2010) 35K51 · 35K55 · 35Q92 · 65M08

1 Introduction

1.1 Presentation of the Model

We consider the following cross-diffusion system:

$$\partial_t u_i + \operatorname{div}\big(- \delta \nabla u_i - u_i \nabla p_i(u)\big) = 0, \quad p_i(u) = \sum_{j=1}^{n} a_{ij} u_j \quad \text{in } \Omega,\ t > 0, \tag{1}$$

where $i = 1, \ldots, n$ with $n \geq 2$, $\Omega \subset \mathbb{R}^2$ is an open bounded polygonal domain, and $\delta > 0$, $a_{ij} > 0$. We impose the initial and no-flux boundary conditions

$$u_i(0) = u_i^0 \geq 0 \quad \text{in } \Omega, \quad \nabla u_i \cdot \nu = 0 \quad \text{on } \partial\Omega,\ t > 0,\ i = 1, \ldots, n, \tag{2}$$

A. Jüngel · A. Zurek (✉)
Institute for Analysis and Scientific Computing, Vienna University of Technology,
Wiedner Hauptstraße 8–10, 1040 Wien, Austria
e-mail: antoine.zurek@tuwien.ac.at

A. Jüngel
e-mail: juengel@tuwien.ac.at

R. Klöfkorn et al. (eds.), *Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples*, Springer Proceedings in Mathematics & Statistics 323,
https://doi.org/10.1007/978-3-030-43651-3_19

where ν is the exterior unit normal vector on $\partial\Omega$. We write $u := (u_1, \dots, u_n)$ and $u^0 := (u_1^0, \dots, u_n^0)$. Equations (1) are derived from a weakly interacting stochastic many-particle system in the mean-field limit [7]. It can be seen as a simplification of the Shigesada–Kawasaki–Teramoto (SKT) population model [12], where the diffusion is reduced to $\delta\nabla u_i$. The two-species system was analyzed first in [3], but up to now, no analytical or numerical results are available for the n-species system. The diffusion matrix associated to (1) is neither symmetric nor positive definite but we show below that system (1) possesses an entropy structure [10] yielding gradient estimates that are the basis for the numerical analysis.

We assume that $A := (a_{ij}) \in \mathbb{R}^{n\times n}$ is positively stable (i.e., all eigenvalues of A have positive real part) and that the detailed-balance condition holds, i.e., there exist numbers $\pi_1, \dots, \pi_n > 0$ such that

$$\pi_i a_{ij} = \pi_j a_{ji} \quad \text{for all } i, j = 1, \dots, n. \tag{3}$$

We refer to [6] for an interpretation of this condition and its connection to Markov chains. Note that for the two-species model this condition is always satisfied, just set $\pi_1 = a_{21}$ and $\pi_2 = a_{12}$. Since $A_1 := \mathrm{diag}(\pi_i^{-1})$ is symmetric, positive definite and $A_2 := (\pi_i a_{ij})$ is symmetric, by [11, Prop. 6.1], the number of positive eigenvalues of $A = A_1 A_2$ equals that for A_2. Thus, A_2 has only positive eigenvalues, which together with the symmetry means that A_2 is symmetric, positive definite.

Our (numerical) analysis is based on the observation that system (1) possesses an entropy structure with a weighted quadratic entropy that has not been observed before in cross-diffusion systems:

$$H[u] = \int_\Omega h(u)dx, \quad \text{where } h(u) := \frac{1}{2\delta}\sum_{i,j=1}^n \pi_i a_{ij} u_i u_j \; = \frac{1}{2\delta} u^T A_2 u,$$

where $(A_2)_{ij} = \pi_i a_{ij}$. Interestingly, this entropy is not of the form $\sum_{i=1}^n h_i(u_i)$, but it mixes the species. A formal computation shows that

$$\frac{dH}{dt} + \sum_{i,j=1}^n \pi_i a_{ij} \int_\Omega \nabla u_i \cdot \nabla u_j dx + \frac{1}{\delta}\sum_{i=1}^n \pi_i \int_\Omega u_i |\nabla p_i(u)|^2 dx = 0.$$

With $\lambda > 0$ being the smallest eigenvalue of A_2, we conclude the following entropy production inequality:

$$\frac{dH}{dt} + \lambda \sum_{i=1}^n \int_\Omega |\nabla u_i|^2 dx + \frac{1}{\delta}\sum_{i=1}^n \pi_i \int_\Omega u_i |\nabla p_i(u)|^2 dx \le 0.$$

Our aim is to prove this inequality for the finite-volume solutions.

1.2 The Numerical Scheme

A mesh of Ω is given by a set $\mathscr{T}$ of open polygonal control volumes, a set $\mathscr{E}$ of edges, and a set $\mathscr{P}$ of points $(x_K)_{K\in\mathscr{T}}$. We assume that the mesh is admissible in the sense of Definition 9.1 in [9]. We distinguish in $\mathscr{E}$ the interior edges $\sigma = K|L$ and the exterior edges such that $\mathscr{E} = \mathscr{E}_{\rm int} \cup \mathscr{E}_{\rm ext}$. For a given control volume $K \in \mathscr{T}$, we denote by $\mathscr{E}_K$ the set of its edges. This set splits into $\mathscr{E}_K = \mathscr{E}_{{\rm int},K} \cup \mathscr{E}_{{\rm ext},K}$. For any $\sigma \in \mathscr{E}$, there exists at least one cell $K \in \mathscr{T}$ such that $\sigma \in \mathscr{E}_K$ and we denote this cell by K_σ. When σ is an interior edge, $\sigma = K|L$, K_σ can be either K or L. For all $\sigma \in \mathscr{E}$, we define $\mathrm{d}_\sigma = \mathrm{d}(x_K, x_L)$ if $\sigma = K|L \in \mathscr{E}_{\rm int}$ and $\mathrm{d}_\sigma = \mathrm{d}(x_K, \sigma)$ if $\sigma \in \mathscr{E}_{{\rm ext},K}$. Then the transmissibility coefficient is defined by $\tau_\sigma = \mathrm{m}(\sigma)/\mathrm{d}_\sigma$ for all $\sigma \in \mathscr{E}$. We assume that the mesh satisfies the following regularity constraint:

$$\exists \xi > 0,\ \forall K \in \mathscr{T},\ \forall \sigma \in \mathscr{E}_K :\ \mathrm{d}(x_K, \sigma) \geq \xi \mathrm{d}_\sigma. \tag{4}$$

The size of the mesh is denoted by $\Delta x = \max_{K\in\mathscr{T}} \mathrm{diam}(K)$. Let $N_T \in \mathbb{N}$ be the number of time steps, $\Delta t = T/N_T$ be the time step size, and $t_k = k\Delta t$ for $k = 0, \ldots, N_T$.

Let $\mathscr{H}_\mathscr{T}$ be the linear space of functions $\Omega \to \mathbb{R}$ which are constant on each $K \in \mathscr{T}$. For $v \in \mathscr{H}_\mathscr{T}$, we introduce

$$D_{K,\sigma} v = v_{K,\sigma} - v_K, \quad D_\sigma v = |D_{K,\sigma} v| \quad \text{for all } K \in \mathscr{T},\ \sigma \in \mathscr{E}_K,$$

where $v_{K,\sigma}$ is either v_L $(\sigma = K|L)$ or v_K $(\sigma \in \mathscr{E}_{{\rm ext},K})$. Finally, we define the (squared) discrete H^1 norm

$$\|v\|_{1,2,\mathscr{T}}^2 = \sum_{\sigma\in\mathscr{E}} \tau_\sigma (D_\sigma v)^2 + \sum_{K\in\mathscr{T}} \mathrm{m}(K) v_K^2.$$

For all $K \in \mathscr{T}$ and $i = 1, \ldots, n$, $u_{i,K}^0$ denotes the mean value of u_i^0 over K. The finite-volume scheme for (1) reads as

$$\frac{\mathrm{m}(K)}{\Delta t}(u_{i,K}^k - u_{i,K}^{k-1}) + \sum_{\sigma\in\mathscr{E}_K} \mathscr{F}_{i,K,\sigma}^k = 0, \quad i = 1, \ldots, n, \tag{5}$$

$$\mathscr{F}_{i,K,\sigma}^k = -\tau_\sigma\big(\delta D_{K,\sigma} u_i^k + u_{i,\sigma}^k D_{K,\sigma} p_i(u^k)\big) \quad \text{for all } K \in \mathscr{T},\ \sigma \in \mathscr{E}_K, \tag{6}$$

with $u^k = (u_1^k, \ldots, u_n^k)$ and $u_{i,\sigma}^k := \min\{u_{i,K}^k, u_{i,K,\sigma}^k\}$. As in [1], this definition of $u_{i,\sigma}^k$ allows us to prove the nonnegativity of $u_{i,K}^k$. This property can be also obtained by an upwind approximation of $u_i \nabla p_i(u)$ in (1).

1.3 Main Result

The main result of this work is the existence of nonnegative solutions to scheme (5)–(6), which preserve the entropy production inequality.

Theorem 1 (Existence of discrete solutions) *Assume that $u^0 \in L^2(\Omega)^n$ with $u_i^0 \geq 0$, $\delta > 0$, $a_{ij} > 0$, A is positively stable, and* (3) *holds. Then there exists a solution $(u_K^k)_{K\in\mathscr{T},k=0,\ldots,N_T}$ with $u_K^k = (u_{1,K}^k, \ldots, u_{n,K}^k)$ to scheme* (5)–(6) *satisfying $u_{i,K}^k \geq 0$ for all $K \in \mathscr{T}$, $i = 1, \ldots, n$, and $k = 0, \ldots, N_T$. Moreover, the following discrete entropy production inequality holds:*

$$\sum_{K\in\mathscr{T}} \mathrm{m}(K)h(u_K^k) + \Delta t\lambda \sum_{i=1}^n \sum_{\sigma\in\mathscr{E}} \tau_\sigma (D_\sigma u_i^k)^2 \\ + \frac{\Delta t}{\delta} \sum_{i=1}^n \sum_{\sigma\in\mathscr{E}} \tau_\sigma \pi_i u_{i,\sigma}^k (D_\sigma p_i(u^k))^2 \leq \sum_{K\in\mathscr{T}} \mathrm{m}(K)h(u_K^{k-1}), \quad (7)$$

where λ denotes the smallest eigenvalue of A_2.

We expect that the detailed-balance condition (3) can be replaced by a weak cross-diffusion condition as in [6]. The positive stability of A implies the parabolicity of (1) in the sense of Petrovskii. Indeed, A_2, defined by $(A_2)_{ij} = \pi_i a_{ij}$, and $A_3 = \mathrm{diag}(u_i/\pi_i)$ are symmetric, positive definite matrices for $u \in (0,\infty)^n$. Thus, its product $(u_i a_{ij})$ has only positive eigenvalues [4, Theorem 7] which proves the claim. The assumption that the diffusion coefficient δ is the same for all species is a simplification needed to conclude that $h(u)$ is coercive, $h(u) \geq (\lambda/2\delta)|u|^2$ for $u \in \mathbb{R}^n$. It can be removed by exploiting the Shannon entropy to show first that u_i is nonnegative, but this requires more technical effort which will be detailed in a future work.

2 Proof of Theorem 1

We proceed by induction. For $k = 0$, we have $u_i^0 \geq 0$ by assumption. Assume that there exists a solution u^{k-1} for some $k \in \{1, \ldots, N_T\}$ such that $u_i^{k-1} \geq 0$ in Ω, $i = 1, \ldots, n$. The construction of a solution u^k is split in several steps.

Step 1: Definition of a linearized problem. Let $R > 0$, we set

$$Z_R := \big\{ w = (w_1, \ldots, w_n) \in (\mathscr{H}_\mathscr{T})^n \ : \ \|w_i\|_{1,2,\mathscr{T}} < R \quad \text{for } i = 1, \ldots, n \big\},$$

and let $\varepsilon > 0$ be given. We define the mapping $F_\varepsilon : Z_R \to \mathbb{R}^{\theta n}$ by $F_\varepsilon(w) = w^\varepsilon$, with $\theta = \#\mathscr{T}$, where $w^\varepsilon = (w_1^\varepsilon, \ldots, w_n^\varepsilon)$ is the solution to the linear problem

$$\varepsilon\left(-\sum_{\sigma\in\mathscr{E}_K}\tau_\sigma D_{K,\sigma}(w_i^\varepsilon)+\mathrm{m}(K)w_{i,K}^\varepsilon\right)=-\left(\frac{\mathrm{m}(K)}{\Delta t}(u_{i,K}-u_{i,K}^{k-1})+\sum_{\sigma\in\mathscr{E}_K}\mathscr{F}_{i,K,\sigma}^{+}\right),\tag{8}$$

for $K\in\mathscr{T}$, $i=1,\ldots,n$, and $\mathscr{F}_{i,K,\sigma}^{+}$ is defined in (6) with $u_{i,\sigma}$ replaced by $\bar{u}_{i,\sigma}=\min\{u_{i,K}^{+},u_{i,K,\sigma}^{+}\}$, where $z^+=\max\{0,z\}$. Here, $u_{i,K}$ is a function of $w_{1,K},\ldots,w_{n,K}$, defined by the entropy variables

$$w_{i,K}=\frac{\pi_i}{\delta}p_i(u_K)=\sum_{j=1}^{n}\frac{\pi_i a_{ij}}{\delta}u_j\quad\text{for all } K\in\mathscr{T},\ i=1,\ldots,n.\tag{9}$$

This is a linear system with the invertible coefficient matrix A_2/δ, and so, the function $u_K=u(w_K)$ is well-defined. The existence of a unique solution w_i^ε to the linear scheme (8)–(9) is now a consequence of [9, Lemma 3.2].

Step 2: Continuity of F_ε. We fix $i\in\{1,\ldots,n\}$. Multiplying (8) by $w_{i,K}^\varepsilon$ and summing over $K\in\mathscr{T}$, we obtain, after discrete integration by parts,

$$\varepsilon\|w_i^\varepsilon\|_{1,2,\mathscr{T}}^2=-\sum_{K\in\mathscr{T}}\frac{\mathrm{m}(K)}{\Delta t}(u_{i,K}-u_{i,k}^{k-1})w_{i,K}^\varepsilon+\sum_{\substack{\sigma\in\mathscr{E}_{\mathrm{int}}\\ \sigma=K|L}}\mathscr{F}_{i,K,\sigma}^{+}D_{K,\sigma}w_i^\varepsilon=:J_1+J_2.$$

By the Cauchy–Schwarz inequality and the definition of $\mathscr{F}_{i,K,\sigma}^{+}$, we find that

$$|J_1|\le\frac{1}{\Delta t}\left(\sum_{K\in\mathscr{T}}\mathrm{m}(K)(u_{i,K}-u_{i,K}^{k-1})^2\right)^{1/2}\left(\sum_{K\in\mathscr{T}}\mathrm{m}(K)(w_{i,K}^\varepsilon)^2\right)^{1/2}$$

$$|J_2|\le\left(\sum_{\sigma\in\mathscr{E}}\tau_\sigma\big(\delta D_\sigma u_i+\bar{u}_{i,\sigma}D_\sigma p_i(u)\big)^2\right)^{1/2}\left(\sum_{\sigma\in\mathscr{E}}\tau_\sigma(D_\sigma w_i^\varepsilon)^2\right)^{1/2}.$$

Hence, since u_i is a linear combination of $(w_1,\ldots,w_n)\in Z_R$, there exists a constant $C(R)>0$ which is independent of w^ε such that $|J_1|+|J_2|\le C(R)\|w_i^\varepsilon\|_{1,2,\mathscr{T}}$. Inserting these estimations, it follows that $\varepsilon\|w_i^\varepsilon\|_{1,2,\mathscr{T}}\le C(R)$.

We turn to the proof of the continuity of F_ε. Let $(w^m)_{m\in\mathbb{N}}\subset Z_R$ be such that $w^m\to w$ as $m\to\infty$. The previous estimate shows that $w^{\varepsilon,m}:=F_\varepsilon(w^m)$ is bounded uniformly in $m\in\mathbb{N}$. Thus, there exists a subsequence of $(w^{\varepsilon,m})$, which is not relabeled, such that $w^{\varepsilon,m}\to w^\varepsilon$ as $m\to\infty$. Passing to the limit $m\to\infty$ in scheme (8)–(9) and taking into account the continuity of the nonlinear functions, we see that w_i^ε is a solution to (8)–(9) for $i=1,\ldots,n$ and $w^\varepsilon=F_\varepsilon(w)$. Because of the uniqueness of the limit function, the whole sequence converges, which proves the continuity.

Step 3: Existence of a fixed point. We claim that the map F_ε admits a fixed point. We use a topological degree argument [8], i.e., we prove that $\deg(I-F_\varepsilon,Z_R,0)=1$, where deg is the Brouwer topological degree. Since deg is invariant by homotopy, it is sufficient to prove that any solution $(w^\varepsilon,\rho)\in\overline{Z}_R\times[0,1]$ to the fixed-point equation $w^\varepsilon = \rho F_\varepsilon(w^\varepsilon)$ satisfies $(w^\varepsilon, \rho) \notin \partial Z_R \times [0,1]$ for sufficiently large values of $R > 0$. Let (w^ε, ρ) be a fixed point and $\rho \neq 0$, the case $\rho = 0$ being clear. Then w_i^ε solves

$$\varepsilon\left(-\sum_{\sigma\in\mathscr{E}_K}\tau_\sigma D_{K,\sigma}(w_i^\varepsilon) + \mathrm{m}(K)w_{i,K}^\varepsilon\right) = -\rho\bigg(\frac{\mathrm{m}(K)}{\Delta t}(u_{i,K}^\varepsilon - u_{i,K}^{k-1}) + \sum_{\sigma\in\mathscr{E}_K}\mathscr{F}_{i,K,\sigma}^{+,\varepsilon}\bigg), \tag{10}$$

for all $K \in \mathscr{T}$, $i = 1, \ldots, n$, and $\mathscr{F}_{i,K,\sigma}^{+,\varepsilon}$ is defined as in (6) with u replaced by u^ε which is related to w^ε by (9). The following discrete entropy production inequality is the key argument.

Lemma 1 (Discrete entropy production inequality) *Let the assumptions of Theorem 1 hold. Then, for any* $\rho \in (0, 1]$ *and* $\varepsilon > 0$*,*

$$\begin{aligned}\rho\sum_{K\in\mathscr{T}}\mathrm{m}(K)h(u_K^\varepsilon) + \varepsilon\Delta t\sum_{i=1}^n\|w_i^\varepsilon\|_{1,2,\mathscr{T}}^2 + \rho\Delta t\lambda\sum_{i=1}^n\sum_{\sigma\in\mathscr{E}}\tau_\sigma(D_\sigma u_i^\varepsilon)^2 \\ + \rho\frac{\Delta t}{\delta}\sum_{i=1}^n\sum_{\sigma\in\mathscr{E}}\tau_\sigma\pi_i\bar{u}_{i,\sigma}^\varepsilon(D_\sigma p_i(u^\varepsilon))^2 \le \rho\sum_{K\in\mathscr{T}}\mathrm{m}(K)h(u_K^{k-1}),\end{aligned} \tag{11}$$

with $\lambda > 0$ *being the smallest eigenvalue of* A_2 *and obvious notations for* $\bar{u}_{i,\sigma}^\varepsilon$.

Proof We multiply (10) by $\Delta t w_{i,K}^\varepsilon$ and sum over i and $K \in \mathscr{T}$. This gives, after discrete integration by parts, $\varepsilon\Delta t\sum_{i=1}^n\|w_i^\varepsilon\|_{1,2,\mathscr{T}}^2 + J_3 + J_4 + J_5 = 0$, where

$$\begin{aligned}J_3 &= \rho\sum_{i=1}^n\sum_{K\in\mathscr{T}}\mathrm{m}(K)(u_{i,K}^\varepsilon - u_{i,K}^{k-1})w_{i,K}^\varepsilon, \\ J_4 &= -\rho\Delta t\sum_{i=1}^n\sum_{\substack{\sigma\in\mathscr{E}_{\mathrm{int}}\\ \sigma=K|L}}\tau_\sigma\delta D_{K,\sigma}u_i^\varepsilon w_{i,K}^\varepsilon, \\ J_5 &= \rho\Delta t\sum_{i=1}^n\sum_{\substack{\sigma\in\mathscr{E}_{\mathrm{int}}\\ \sigma=K|L}}\tau_\sigma\bar{u}_{i,\sigma}^\varepsilon D_{K,\sigma}p_i(u^\varepsilon)D_{K,\sigma}w_{i,K}^\varepsilon.\end{aligned}$$

To estimate J_3, we use the convexity of h; for J_4, we take into account the symmetry of τ_σ with respect to $\sigma = K|L$, definition (9) of w_i^ε and the positive definiteness of A_2; and for J_5, we employ definition (9) of w_i^ε:

$$\begin{aligned}J_3 &\ge \rho\sum_{K\in\mathscr{T}}\mathrm{m}(K)\big(h(u_K^\varepsilon) - h(u_K^{k-1})\big), \\ J_4 &= \rho\Delta t\sum_{i,j=1}^n\sum_{\substack{\sigma\in\mathscr{E}_{\mathrm{int}}\\ \sigma=K|L}}\tau_\sigma\pi_i a_{ij}D_{K,\sigma}u_i^\varepsilon D_{K,\sigma}u_j^\varepsilon \ge \rho\Delta t\lambda\sum_{i=1}^n\sum_{\sigma\in\mathscr{E}}\tau_\sigma(D_\sigma u_i^\varepsilon)^2,\end{aligned}$$

$$J_5 = \rho \frac{\Delta t}{\delta} \sum_{i=1}^{n} \sum_{\sigma \in \mathscr{E}} \tau_\sigma \pi_i \bar{u}^{\varepsilon}_{i,\sigma} (D_\sigma p_i(u^\varepsilon))^2.$$

Putting all the estimations together completes the proof. □

We proceed with the topological degree argument. Lemma 1 implies that

$$\varepsilon \Delta t \sum_{i=1}^{n} \|w_i^\varepsilon\|^2_{1,2,\mathscr{T}} \le \rho \sum_{K \in \mathscr{T}} \mathrm{m}(K) h(u_K^{k-1}) \le \sum_{K \in \mathscr{T}} \mathrm{m}(K) h(u_K^{k-1}).$$

Then, if we define $R := (\varepsilon \Delta t)^{-1/2} (\sum_{K \in \mathscr{T}} \mathrm{m}(K) h(u_K^{k-1}))^{1/2} + 1$, we conclude that $w^\varepsilon \notin \partial Z_R$ and $\deg(I - F_\varepsilon, Z_R, 0) = 1$. Thus, F_ε admits a fixed point.

Step 4: Limit $\varepsilon \to 0$. Recall that $h(u_K) \ge \lambda/(2\delta)|u_K|^2$ (note that $u_{i,K} \in \mathbb{R}$ at this point). Thus, by Lemma 1, there exists a constant $C > 0$ depending only on the mesh but not on ε such that for all $K \in \mathscr{T}$ and $i = 1, \ldots, n$,

$$|u^\varepsilon_{i,K}| \le C(\lambda) \left(\sum_{K \in \mathscr{T}} \mathrm{m}(K) h(u_K^{k-1}) \right)^{1/2}.$$

Thus, up to a subsequence, for $i = 1, \ldots, n$ and for all $K \in \mathscr{T}$, we infer the existence of $u_{i,K} \in \mathbb{R}$ such that $u^\varepsilon_{i,K} \to u_{i,K}$ as $\varepsilon \to 0$. We deduce from (11) that there exists a subsequence (not relabeled) such that $\varepsilon w^\varepsilon_{i,K} \to 0$ for any $K \in \mathscr{T}$ and $i = 1, \ldots, n$. Hence, the limit $\varepsilon \to 0$ in (8) yields the existence of a solution to (8) with $\varepsilon = 0$.

Let $i \in \{1, \ldots, n\}$ and $K \in \mathscr{T}$ such that $u_{i,K} = \min_{L \in \mathscr{T}} u_{i,L}$. We multiply (8) with $\varepsilon = 0$ by $\Delta t u^-_{i,K}$ with $z^- = \min\{0, z\}$ and use the induction hypothesis:

$$\mathrm{m}(K)(u^-_{i,K})^2 - \Delta t \sum_{\sigma \in \mathscr{E}_K} \tau_\sigma (\delta + a_{ii} \bar{u}_{i,\sigma}) D_{K,\sigma}(u_i) u^-_{i,K} \\ - \Delta t \sum_{j \ne i} \sum_{\sigma \in \mathscr{E}_K} \tau_\sigma a_{ij} \bar{u}_{i,\sigma} D_{K,\sigma}(u_j) u^-_{i,K} = 0.$$

The second term is nonpositive since $\bar{u}_{i,\sigma} \ge 0$ and $D_{K,\sigma}(u_i) \ge 0$, by the choice of K. The last term vanishes since $\bar{u}_{i,\sigma} u^-_{i,K} = u^+_{i,K} u^-_{i,K} = 0$, by the definition of $\bar{u}_{i,\sigma}$. This shows that $u_{i,L} \ge u_{i,K} \ge 0$ for all $L \in \mathscr{T}$ and $i = 1, \ldots, n$. Passing to the limit $\varepsilon \to 0$ in (11) yields inequality (7), which completes the proof of Theorem 1.

3 Convergence Analysis and Perspectives

In this section, we sketch the proof of the convergence of the scheme and possible extensions of the method presented in this paper.

- Let us give the main features of the proof of convergence. First, thanks to the a priori estimates given by (7) and assumption (4), we prove the existence of a constant $C > 0$ independent of Δx and Δt such that for all $i = 1, \dots, n$ and $\phi \in C_0^\infty(Q_T)$, where $Q_T := \Omega \times (0, T)$,

$$\sum_{k=1}^{N_T} \sum_{K\in\mathscr{T}} \mathrm{m}(K)(u_{i,K}^k - u_{i,K}^{k-1})\phi(x_K, t_k) \le C\|\nabla\phi\|_{L^\infty(Q_T)}. \tag{12}$$

  Next, we consider a sequence of admissible meshes $(\mathscr{T}_\eta, \Delta t_\eta)_{\eta>0}$ of Q_T, indexed by the size $\eta = \{\Delta x, \Delta t\}$, satisfying (4) uniformly in η. For any $\eta > 0$, we denote by $u_\eta = (u_{1,\eta}, \dots, u_{n,\eta})$ the piecewise constant (in time and space) finite-volume solution constructed in Theorem 1. We deduce, thanks to [2, Theorem 3.9] and (12), that there exist nonnegative functions $u_1, \dots, u_n$ such that, up to a subsequence,

$$u_{i,\eta} \to u_i \quad \text{a.e. in } Q_T \text{ as } \eta \to 0, \quad i = 1, \dots, n.$$

  Moreover, we conclude from (7) that $u_{i,\eta}$ is uniformly bounded in $L^\infty(0, T; L^2(\Omega))$ and $L^2(0, T; L^p(\Omega))$ for $p < \infty$ thanks to (7) and Sobolev embedding. We deduce from the Riesz–Thorin theorem that $(u_{i,\eta})$ is bounded in $L^r(Q_T)$ for some $2 < r < 4$ and thus, it is equi-integrable. Thus, applying the Vitali convergence theorem, we infer that, up to a subsequence, $u_{i,\eta} \to u_i$ strongly in $L^r(Q_T)$ for all $r < 4$ as $\eta \to 0$, $i = 1, \dots, n$. The discrete entropy production inequality yields a uniform bound of the discrete gradient ∇^η of $u_{i,\eta}$ in $L^2(Q_T)$; see [5] for a definition of ∇^η. It follows from [5, Lemma 4.4] that, up to a subsequence,

$$\nabla^\eta u_{i,\eta} \rightharpoonup \nabla u_i \quad \text{weakly in } L^2(Q_T) \text{ as } \eta \to 0,\ i = 1, \dots, n.$$

  Finally, following the method developed in [5], we prove that the limit function $u = (u_1, \dots, u_n)$ is a weak solution to (1)–(2).
- We already mentioned that system (1) can be interpreted as a simplification of the SKT model. In a future work, we will analyze a structure-preserving finite-volume approximation of the full SKT model. Such a discretization was analyzed in [1], but only for positive definite diffusion matrices associated to (1). We will extend the analysis of [1] without this assumption.

Acknowledgements The authors acknowledge partial support from the Austrian Science Fund (FWF), grants P30000, P33010, W1245, and F65 and from the bilateral Amadée Program of the Austrian ÖAD. The authors also thank the referees for their careful reading.

References

1. Andreianov, B., Bendahmane, M., Baier, R.: Finite volume method for a cross-diffusion model in population dynamics. Math. Models Meth. Appl. Sci. **21**, 307–344 (2011)
2. Andreianov, B., Cancès, C., Moussa, A.: A nonlinear time compactness result and applications to discretization of degenerate parabolic-elliptic pdes. J. Funct. Anal. **273**, 3633–3670 (2017)
3. Bertsch, B., Gurtin, M., Hilhorst, D., Peletier, L.: On interacting populations that disperse to avoid crowding: preservation of segregation. J. Math. Biol. **23**, 1–13 (1985)
4. Bosch, A.: Note on the factorization of a square matrix into two Hermitian or symmetric matrices. SIAM Rev. **29**, 463–468 (1987)
5. Chainais-Hillairet, C., Liu, J.G., Peng, Y.J.: Finite volume scheme for multi-dimensional drift-diffusion equations and convergence analysis. ESAIM: Math. Model. Numer. Anal. **37**, 319–338 (2003)
6. Chen, X., Daus, E., Jüngel, A.: Global existence analysis of cross-diffusion population systems for multiple species. Arch. Rational Mech. Anal. **227**, 715–747 (2018)
7. Chen, L., Daus, E., Jüngel, A.: Rigorous mean-field limit and cross-diffusion. Z. Angew. Math. Phys. **70**, article 122, 21 pages (2019)
8. Deimling, K.: Nonlinear Functional Analysis. Springer, Berlin (1985)
9. Eymard, R., Gallouët, T., Herbin, R.: Finite volume methods. In: Handbook of Numerical Analysis, vol. VII, North-Holland, pp. 713–1020 (2000)
10. Jüngel, A.: The boundedness-by-entropy method for cross-diffusion systems. Nonlinearity **28**, 1963–2001 (2015)
11. Serre, D.: Matrices. Theory and Applications, 2nd edn. Springer, New York (2010)
12. Shigesada, N., Kawasaki, K., Teramoto, E.: Spatial segregation of interacting species. J. Theor. Biol. **79**, 83–99 (1979)

Finite Volume Method for a System of Continuity Equations Driven by Nonlocal Interactions

Anissa El Keurti and Thomas Rey

Abstract We present a new finite volume method for computing numerical approximations of a system of nonlocal transport equation modeling interacting species. This method is based on the work [F. Delarue, F. Lagoutière, N. Vauchelet, *Convergence analysis of upwind type schemes for the aggregation equation with pointy potential*, Ann. Henri. Lebesgue 2019], where the nonlocal continuity equations are treated as conservative transport equations with a nonlocal, nonlinear, rough velocity field. We analyze some properties of the method, and illustrate the results with numerical simulations.

Keywords Upwind finite volume method · System of aggregation equations · Population dynamics · Continuity equations · Measure-valued solutions

MSC (2010) 45K05 · 65M08 · 65L20 · 92D25

1 A Nonlocal Predator-Prey Model

We consider a system of nonlocal equations modeling the swarming dynamics of species which interact with each others through attractive/repulsive potentials (such as predators and preys). The system is an extension of the well-known aggregation equation [1], and can be written in the following form:

$$\begin{cases} \partial_t \rho_1 + \operatorname{div}(\rho_1(\nabla W_1 * \rho_1 + \nabla K * \rho_2)) = 0, & \rho_1(0,\cdot) = \rho_1^{in}, \\ \partial_t \rho_2 + \operatorname{div}(\rho_2(\nabla W_2 * \rho_2 - \beta \nabla K * \rho_1)) = 0, & \rho_2(0,\cdot) = \rho_2^{in}, \end{cases} \tag{1}$$

A. El Keurti · T. Rey (✉)
Inria–Laboratoire Paul Painlevé, CNRS, UMR 8524, Université de Lille,
59000 Lille, France
e-mail: thomas.rey@univ-lille.fr

A. El Keurti
e-mail: elkeurti.anissa@gmail.com

R. Klöfkorn et al. (eds.), *Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples*, Springer Proceedings in Mathematics & Statistics 323,
https://doi.org/10.1007/978-3-030-43651-3_20

where $\rho_1(t,x)$ and $\rho_2(t,x)$ are probability measures that model the density of species 1 and 2 (respectively predators and preys), for $x \in \mathbb{R}^d$, $t \in \mathbb{R}$. This model was introduced in [5], where it was derived from a system of N interacting particles. It has since been mathematically studied in [2, 6].

The functions $W_\alpha,\ K : \mathbb{R}^d \to \mathbb{R}_+,\ \alpha \in \{1,2\}$ denote respectively the *intra-specific* interaction potentials of the species α, and the *inter-specific* interaction potential. The intra-specific potential W_α can be of *attractive* (namely radial with a nonnegative derivative) or *repulsive* type (radial with a nonpositive derivative), depending on the gregarious behavior of species α. The potential K is of attractive type, modeling the fact that species 2 flees species 1 whereas species 1 is attracted by species 2. The parameter $\beta \in [0,1)$ expresses the mobility of species 1.

2 Cauchy Theory

Definition 1 A function $W : \mathbb{R}^d \to \mathbb{R}$ is called a *pointy* potential if it satisfies the following properties:

1. W is Lipschitz continuous, symmetric and $W(0) = 0$;
2. W is λ-convex for some $\lambda \leq 0$ (namely $W - \frac{\lambda}{2}|\cdot|^2$ is convex);
3. $W \in \mathscr{C}^1(\mathbb{R}^d \setminus \{0\})$.

Let us assume that $W_\alpha, \alpha \in \{1,2\}$, and K are pointy potentials as in Definition 1. These potentials being Lipschitz, there exist $\omega_{\alpha,\infty}$ and κ_∞ such that for all $x \neq 0$:

$$|\nabla W_\alpha(x)| \leq \omega_{\alpha,\infty}, \quad |\nabla K(x)| \leq \kappa_\infty. \tag{2}$$

Let us also define the *macroscopic velocities* $\widehat{a_{\rho_1}}$ and $\widehat{a_{\rho_2}}$ as

$$\widehat{a_{\rho_1}}(t,x) := -\int_{\mathbb{R}^d} \left(\widehat{\nabla W}_\alpha(x-y)\ \rho_1(t,y) + \widehat{\nabla K}(x-y)\ \rho_2(t,y)\right) dy, \tag{3}$$

$$\widehat{a_{\rho_2}}(t,x) := -\int_{\mathbb{R}^d} \left(\widehat{\nabla W}_\alpha(x-y)\ \rho_2(t,y) - \beta\widehat{\nabla K}(x-y)\ \rho_1(t,y)\right) dy, \tag{4}$$

where we denoted for a pointy potential W the following extension:

$$\widehat{\nabla W}(x) = \begin{cases} \nabla W_\alpha(x) \text{ for } x \neq 0, \\ 0 \text{ for } x = 0. \end{cases}$$

Existence theory for problem (1) has been studied in [5] in the case of $\mathscr{C}^1$ pointy potentials. Uniqueness was obtained in [7] by introducing duality solutions. This approach will allow to prove the convergence of our numerical scheme (7). Using the theory of Filippov characteristics, one can also prove the following general result:

Theorem 1 (From [3]) *Let W_α, $\alpha \in \{1,2\}$, and K be pointy potential that satisfy (2), and $\rho_\alpha^{in} \in \mathscr{P}_2(\mathbb{R}^d)$. There exist unique probability measures ρ_α that are global distributional solutions to the following system of transport equations:*

$$\begin{cases} \partial t \rho_1 + div(\widehat{a_{\rho_1}}\rho_1) = 0, & \rho_1(0,\cdot) = \rho_1^{in}, \\ \partial t \rho_2 + div(\widehat{a_{\rho_2}}\rho_2) = 0, & \rho_2(0,\cdot) = \rho_2^{in}. \end{cases} \tag{5}$$

3 Numerical Scheme

We shall now apply the numerical scheme introduced in [4] for approximating solutions to the classical (single species) aggregation equation to the system (1). Let us introduce a cartesian mesh $(C_J)_{J\in\mathbb{Z}^d}$ of $\mathbb{R}^d$, with step Δx_i in the direction $i \in \{1,\dots,d\}$, and $\Delta x = \max \Delta x_i$. The center of a given cell C_J will then be defined by $x_j := (J_1\Delta x_1, \dots, J_d \Delta x_d)$. Let also $e_i := (0,\dots,0,1,0,\dots,0)$ be the ith vector of the canonical basis.

For an initial probability measure $\rho_\alpha^{in} \in \mathscr{P}_2(\mathbb{R}^d)$, $\alpha \in \{1,2\}$, we define $\rho_{\alpha,J}^0$ as the cell average values of ρ_α^{in} over the cell C_J :

$$\rho_{\alpha,J}^0 = \frac{1}{m\,(C_J)} \int_{C_J} \rho_1^{ini}(dx) \geq 0. \tag{6}$$

Given an approximation $({\rho_\alpha}_J^n)_{J\in\mathbb{Z}^d}$ of the cell averages of $\rho_\alpha(t^n,\cdot)$ at a given time $t^n = n\Delta t$, we compute ${\rho_\alpha}_J^{n+1}$ as:

$$\begin{cases} {\rho_1}_J^{n+1} = {\rho_1}_J^n - \displaystyle\sum_{i=1}^d \frac{\Delta t}{\Delta x_i}\Big(& ({a_1}_{i\,J}^n)^+ {\rho_1}_J^n - ({a_1}_{i\,J+e_i}^n)^- {\rho_1}_{J+e_i}^n \\ & - ({a_1}_{i\,J-e_i}^n)^+ {\rho_1}_{J-e_i}^n + ({a_1}_{i\,J}^n)^- {\rho_1}_J^n \Big), \\ {\rho_2}_J^{n+1} = {\rho_2}_J^n - \displaystyle\sum_{i=1}^d \frac{\Delta t}{\Delta x_i}\Big(& ({a_2}_{i\,J}^n)^+ {\rho_2}_J^n - ({a_2}_{i\,J+e_i}^n)^- {\rho_2}_{J+e_i}^n \\ & - ({a_2}_{i\,J-e_i}^n)^+ {\rho_2}_{J-e_i}^n + ({a_2}_{i\,J}^n)^- {\rho_2}_J^n \Big). \end{cases} \tag{7}$$

where the discrete macroscopic velocities are defined as

$$\begin{cases} {a_1}_{i\,J}^n = -\sum_{L\in\mathbb{Z}^d} \left({\rho_1}_L^n D_i {W_1}_J^L + {\rho_2}_L^n D_i K_J^L \right), \\ {a_2}_{i\,J}^n = -\sum_{L\in\mathbb{Z}^d} \left({\rho_2}_L^n D_i {W_2}_J^L - \beta {\rho_1}_K^n D_i K_J^L \right), \end{cases} \tag{8}$$

with $D_i W_J^K := \partial_{x_i} \widehat{W}(x_J - x_K)$ for a pointy potential W.

Lemma 1 *If W_α, $\alpha \in \{1, 2\}$, and K are pointy potentials and the following CFL condition holds:*

$$(\omega_{\alpha,\infty} + \kappa_\infty) \sum_{i=1}^{d} \frac{\Delta t}{\Delta x_i} \leq 1, \tag{9}$$

one has the following properties for the scheme (7):

1. *For $\rho_\alpha^{in} \in \mathscr{P}_2(\mathbb{R}^d)$ and $\rho_{\alpha,J}^0$ given by (6), the sequences $(\rho_{\alpha,J}^n)_{J\in\mathbb{Z}^d, n\in\mathbb{N}}$ and $(a_{\alpha}{}_{i\,J}^{n})_{J\in\mathbb{Z}^d, n\in\mathbb{N}, i=1,\ldots,d}$ satisfy:*

$$\rho_{\alpha,J}^n \geq 0, \quad |a_{\alpha}{}_{i\,J}^{n}| \leq (\omega_{\alpha,\infty} + \kappa_\infty), \quad i = 1, \ldots, d,$$

and for all $n \in \mathbb{N}$,

$$\sum_{J\in\mathbb{Z}^d} \rho_{\alpha,J}^n m(C_J) = \int_{\mathbb{R}} \rho_\alpha^{in}(dx).$$

2. *Conservation of the weighted center of mass:*

$$\sum_{J\in\mathbb{Z}^d} x_J(\beta\rho_{1,J}^n + \rho_{2,J}^n) = \sum_{J\in\mathbb{Z}^d} x_J(\beta\rho_{1,J}^0 + \rho_{2,J}^0).$$

Proof 1. By summing the two equations of (7) over all $J \in \mathbb{Z}^d$, one obtains the mass conservation. Then, writing both identities in (7) as:

$$\rho_{\alpha}{}_{J}^{n+1} = \rho_{\alpha}{}_{J}^{n}\Big[1 - \sum_{i=1}^{d} |a_{\alpha}{}_{i\,J}^{n}|\Big] + \sum_{i=1}^{d} \frac{\Delta t}{\Delta x_i}(a_{\alpha}{}_{i\,J+e_i}^{n})^- \rho_{\alpha}{}_{J+e_i}^{n}$$
$$+ \sum_{i=1}^{d} \frac{\Delta t}{\Delta x_i}(a_{\alpha}{}_{i\,J-e_i}^{n})^+ \rho_{\alpha}{}_{J-e_i}^{n},$$

one proves by induction on n that $\rho_{\alpha,J}^n \geq 0$ for all $J \in \mathbb{Z}^d$, $n \in \mathbb{N}$ under the CFL condition (9). Indeed, by using the definition (8), one has

$$|a_{\alpha_{i,J}}^n| \leq (\omega_\infty + \kappa_\infty) \sum_{J\in\mathbb{Z}^d} \rho_{\alpha,J}^n = (\omega_\infty + \kappa_\infty) \sum_{J\in\mathbb{Z}^d} \rho_{\alpha,J}^0, \quad i \in \{1, \ldots, d\},$$

which concludes the proof by a convexity argument.

2. Using a discrete integration by parts and (7), one has:

$$\sum_{J\in\mathbb{Z}^d} x_J \rho_{\alpha,J}^{n+1} = \sum_{J\in\mathbb{Z}^d} x_J \rho_{\alpha,J}^n - \sum_{i=1}^{d} \frac{\Delta t}{\Delta x_i} \sum_{J\in\mathbb{Z}^d} \Big((a_{\alpha}{}_{i\,J}^{n})^+ \rho_{\alpha}{}_{J}^{n}(x_J - x_{J+e_i})$$
$$- (a_{\alpha}{}_{i\,J}^{n})^- \rho_{\alpha}{}_{J}^{n}(x_{J-e_i} - x_J)\Big).$$

Since x_J denote the cell centers, one has

$$\sum_{J\in\mathbb{Z}^d} x_J\left(\beta\rho_{1,J}^{n+1}+\rho_{2,J}^{n+1}\right)=\sum_{J\in\mathbb{Z}^d} x_J\left(\beta\rho_{1,J}^{n}+\rho_{2,J}^{n}\right) \\ +\Delta t\sum_{i=1}^{d}\sum_{J\in\mathbb{Z}^d}\left(\beta {a_1}_{i\ J}^{n}{\rho_1}_J^n+{a_2}_{i\ J}^{n}{\rho_2}_J^n\right). \quad (10)$$

Summing over all the cells in (8), and since ∇W_α and ∇K are odd, one obtains after reindexing:

$$\begin{aligned}\sum_{J\in\mathbb{Z}^d}\beta {a_1}_{i\ J}^{n}{\rho_1}_J^n+{a_2}_{i\ J}^{n}{\rho_2}_J^n &=\sum_{J\in\mathbb{Z}^d}\sum_{L\in\mathbb{Z}^d}\left(\beta {\rho_1}_J^n{\rho_1}_L^n D_i {W_1}_J^L+{\rho_2}_J^n{\rho_2}_L^n D_i {W_2}_J^L\right)\\ &=-\sum_{J\in\mathbb{Z}^d}\sum_{L\in\mathbb{Z}^d}\left(\beta {\rho_1}_J^n{\rho_1}_L^n D_i {W_1}_L^J+{\rho_2}_J^n{\rho_2}_L^n D_i {W_2}_L^J\right)\\ &=0\end{aligned}$$

which yields the conclusion when plugged into (10).

We are now ready to prove the convergence of the scheme (7).

Theorem 2 *Let us assume that W_α, $\alpha\in\{1,2\}$ and K are pointy potentials, and that the following CFL condition holds on the mesh (C_J):*

$$(\omega_{\alpha,\infty}+\kappa_\infty)\sum_{i=1}^{d}\frac{\Delta t}{\Delta x_i}\le 1.$$

Let $\rho_\alpha^{in}\in\mathscr{P}_2(\mathbb{R}^d)$ and $\rho_{\alpha,J}^0$ given by (6) for all $J\in\mathbb{Z}^d$ and define the empirical distribution as

$$\rho_{\alpha,\Delta x}^n=\sum_{J\in\mathbb{Z}^d}\rho_{\alpha,J}^n\delta_{x_J},\quad n\in\mathbb{N},$$

where $((\rho_{\alpha,J}^n)_{J\in\mathbb{Z}^d})_{n\in\mathbb{N}}$ is given by (7).

Then $\rho_{1,\Delta x}$ and $\rho_{2,\Delta x}$ converge weakly in $\mathscr{M}_b([0,T]\times\mathbb{R}^d)$ towards respectively ρ_1 and ρ_2 which are the solutions to (5) as Δx goes to 0.

Proof Let us give the ideas behind this convergence proof, in the unidimensional case (inspired from [7]).

1. Extraction of a convergent subsequence.
   The total variation of $\rho_{\alpha,\Delta x}$ is bounded and we can thus extract a subsequence of $\rho_{\alpha,\Delta x}$ that converges weakly towards $\rho_\alpha\in\mathscr{M}_b([0,T]\times\mathbb{R})$.
2. Modified equations and Taylor expansion.
   We write the modified equation satisfied by $\rho_{\alpha,\Delta x}$ in terms of distributions. Let us consider $\phi\in C_c^\infty([0,T]\times\mathbb{R})$. By using the dual product in sense of distribution $<\cdot,\cdot>$, one has

$$< \partial_t \rho_{\alpha,\Delta x}, \phi > + < \widehat{a}_{\alpha,\Delta x} \rho_{\alpha,\Delta x}, \partial_x \phi >= 0,$$

where $\widehat{a}_{\alpha,\Delta x} = \sum_{n=0}^{N_T} \sum_{J\in\mathbb{Z}} \widehat{a}^n_{\alpha,J} \mathbf{1}_{[t^n,t^{n+1}[}(t)\delta_{x_J}(x)$. Taylor expanding ϕ allows to rewrite this equation in terms of distributions. One then bounds the different terms by using a straightforward adaptation of [7, Lemma 6.2] to this model.

3. Passing to the limit.
   We finaly use [7, Lemma 3.2] to pass to the limit. The limit ρ_α thus satisfies (5). By uniqueness from Theorem 1, ρ_α is the unique solution of (1).

4 Numerical Simulations in 2D

We implemented the scheme in 2 dimensions for a square grid and potentials such as the Newtonian potential $\mathscr{N}(x) = |x|$ (pointy and 0–convex), or $W = 1 - e^{-|x|}$ (pointy and -1–convex). The grid in all the simulations is composed of 50×50 points, with $\Delta t = 0.005$ (according to the CFL condition (9)).

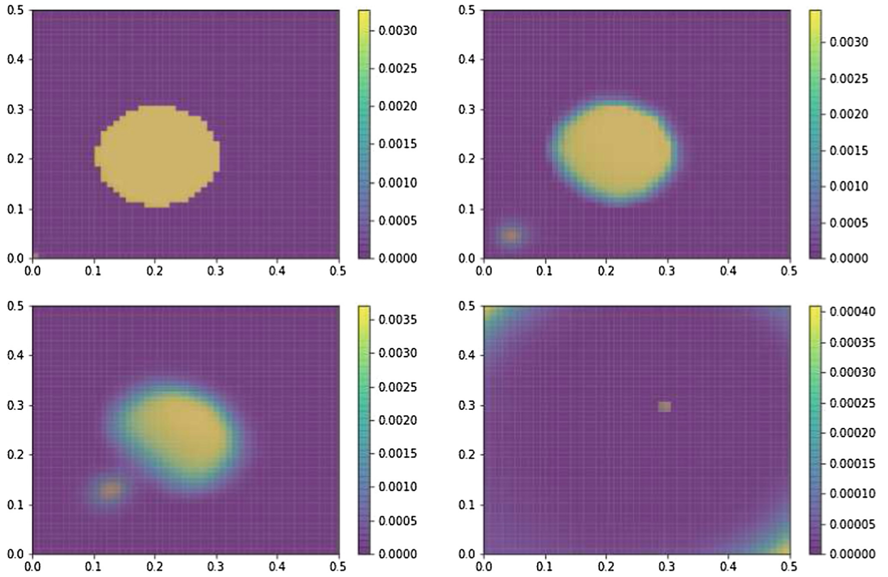


Fig. 1 Test 1. Newtonian potentials $W_1(x) = W_2(x) = 0.1|x|$, $K(x) = |x|$, $\beta = 0.3$. with a single predator at the origin, and an uniform distribution of preys as initial data. Isovalues of $\rho_1 + \rho_2$ at times $t = 0, 0.03, 0.09$ and 7.5

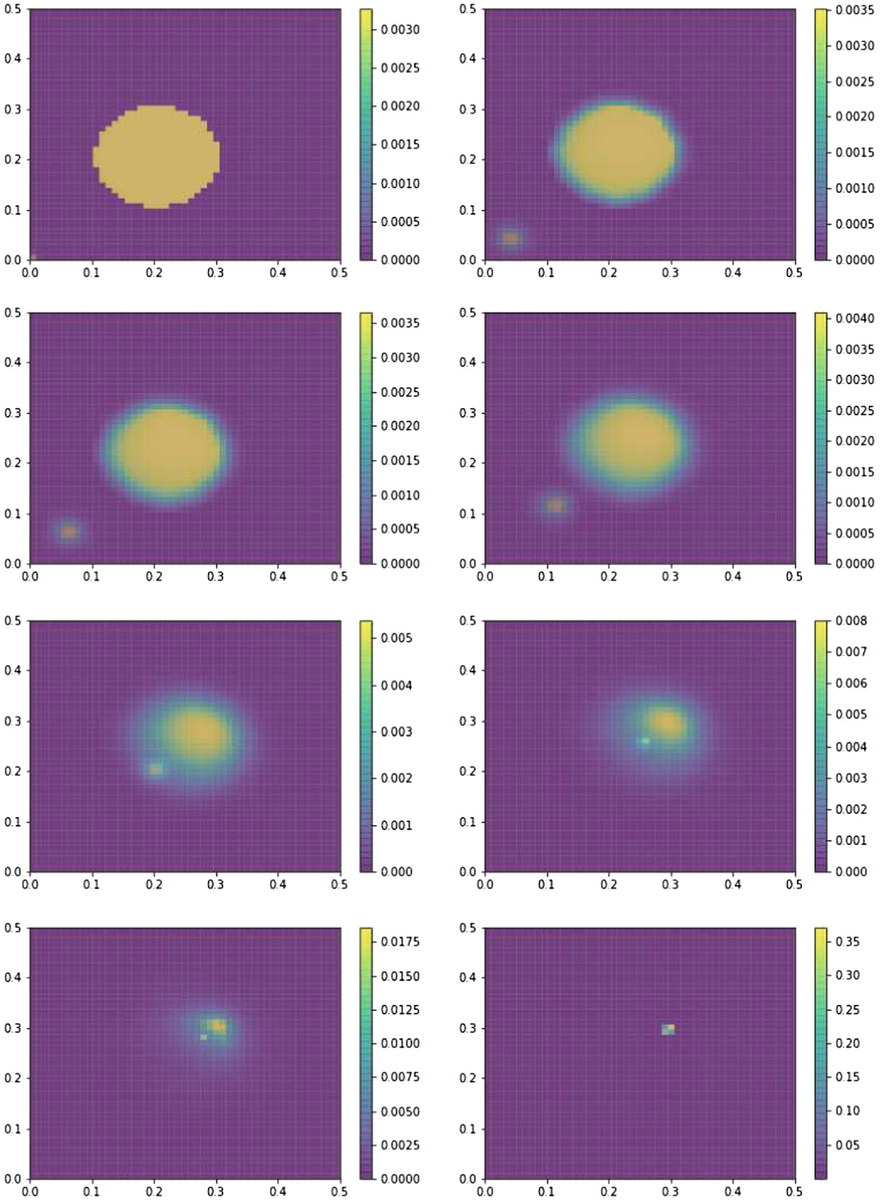


Fig. 2 Test 2. Newtonian potentials $W_1(x) = W_2(x) = 0.1|x|$, "fly-and-regroup" potential $K(x) = 1 - (|x| + 1)e^{-|x|}$, $\beta = 0.3$. with a single predator at the origin, and an uniform distribution of preys as initial data. Isovalues of $\rho_1 + \rho_2$ at times $t = 0, 0.05, 0.1, 0.15, 0.3, 0.6, 1$ and 5

4.1 Test 1. Evading Preys

In Fig. 1, we present simulations made with a Dirac delta as initial data to model a single predator, and a uniform distribution for preys:

$$\rho_1^{in} = \delta_0(x), \qquad \rho_2^{in} = \mathbf{1}_{\mathscr{B}(0.2,0.1)}. \tag{11}$$

We use Newtonian potentials $W_1 = W_2 = 0.1\mathscr{N}$, $K = \mathscr{N}$ for inter and intra-specific interactions, with a mobility $\beta = 0.3$. At the beginning of the simulation, we observe that the predator is getting closer to the preys. When the group of preys is close, the preys create a circular pattern around the predators in order to run away from him.

4.2 Test 2. A More Realistic Potential for Inter-specific Interaction

In [6], the authors introduced a potential K that is more relevant in terms of modeling:

$$K(x) = 1 - (|x| + 1)e^{-|x|}. \tag{12}$$

When the predator is far from the preys, the inter-specific interaction depends very weakly on the distance between preys and predator, and is almost constant. When the predator becomes closer to the preys, they become paralyzed, the potential being the close to 0. We performed simulations with an initial data given by (11) in Fig. 2. We observe a similar behavior than in Fig. 1 in short time, but a convergence toward a single Dirac delta (the predator has gathered all the preys together) in large time.

Acknowledgements TR was partially funded by Labex CEMPI (ANR-11-LABX-0007-01) and ANR Project MoHyCon (ANR-17-CE40-0027-01).

References

1. Bertozzi, A.L., Carrillo, J.A., Laurent, T.: Blow-up in multidimensional aggregation equations with mildly singular interaction kernels. Nonlinearity **22**(3), 683 (2009)
2. Carrillo, J.A., Francesco, M.D., Esposito, A., Fagioli, S., Schmidtchen, M.: Measure solutions to a system of continuity equations driven by Newtonian nonlocal interactions. Discr. Cont. Dyn. Sys. A **40**, 1191 (2020)
3. Carrillo, J.A., James, F., Lagoutière, F., Vauchelet, N.: The filippov characteristic flow for the aggregation equation with mildly singular potentials. J. Diff. Eq. **260**(1), 304–338 (2016)
4. Delarue, F., Lagoutière, F., Vauchelet, N.: Convergence order of upwind type schemes for transport equations with discontinuous coefficients. J. Math. Pures. App. **108**(6), 918–951 (2017)
5. Di Francesco, M., Fagioli, S.: Measure solutions for non-local interaction pdes with two species. Nonlinearity **26**(10), 2777 (2013)

6. Di Francesco, M., Fagioli, S.: A nonlocal swarm model for predators-prey interactions. Math. Mod. Meth. App. Sci. **26**(02), 319–355 (2016)
7. Emako-Kazianou, C., Liao, J., Vauchelet, N.: Synchronising and non-synchronising dynamics for a two-species aggregation model. Discr. Cont. Dyn. Sys. B **22**(6), 2121–2146 (2017)

A Macroscopic Model to Reproduce Self-organization at Bottlenecks

Boris Andreianov and Abraham Sylla

Abstract We propose a model for self-organized traffic flow at bottlenecks that consists of a scalar conservation law with a nonlocal constraint on the flux. The constraint is a function of an organization marker which evolves through an ODE depending on the upstream traffic density and its variations. We prove well-posedness for the problem, construct and analyze a finite volume scheme, perform numerical simulations and discuss the model and related perspectives.

Keywords LWR traffic model · Nonlocal point constraint · Bottleneck · Self-organization · Finite volume scheme

MSC (2010) 35L65 · 90B20 · 65M08 · 65M12

1 Introduction

The LWR framework is the simplest one that can be used to describe macroscopically pedestrian/road traffic in a corridor or on a road. It takes the form

$$\partial_t \rho + \partial_x f(\rho) = 0.$$

Above, $\rho = \rho(x,t) \in [0, R]$ is the density of pedestrians/cars at (x,t). We assume that the flux function $f : [0, R] \to \mathbb{R}$ is Lipschitz continuous and bell-shaped, which are commonly used assumptions in traffic dynamics:

B. Andreianov · A. Sylla (✉)
Institut Denis Poisson, CNRS UMR 7013, Université de Tours,
Université d'Orléans, Parc de Grandmont, 37200 Tours cedex, France
e-mail: Abraham.Sylla@lmpt.univ-tours.fr

B. Andreianov
e-mail: Boris.Andreianov@lmpt.univ-tours.fr

R. Klöfkorn et al. (eds.), *Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples*, Springer Proceedings in Mathematics & Statistics 323,
https://doi.org/10.1007/978-3-030-43651-3_21

$$f(\rho) \geq 0,\ f(0) = f(R) = 0,\ \exists! \overline{\rho} \in (0, R),\ f'(\rho)(\overline{\rho} - \rho) > 0 \text{ for a.e. } \rho \in (0, R). \tag{1}$$

Point constraints were introduced in [16, 17] in the LWR model in order to account for localized in space phenomena that may occur at exits—such as traffic lights or tollgates in the context of road traffic—and which act as obstacles. To do so, one can impose a localized constraint on the flux such as

$$f(\rho)|_{x=0} \leq q(t).$$

One of the typical features of both vehicle and pedestrian flows is self-organization, see [13, 18, 22] for empirical data that put in evidence this phenomenon. Here, we focus on self-organization near exits. We do not intend to model the different mechanisms behind self-organization, but only to reproduce its phenomenology. In [5] the authors attempted to reproduce self-organization with a model based on the LWR-flux constraint framework:

$$\begin{cases} \partial_t \rho + \partial_x f(\rho) = 0 & \mathbb{R} \times (0, T) \\ \rho(x, 0) = \rho_0(x) & x \in \mathbb{R} \\ f(\rho)|_{x=0} \leq p\left(\displaystyle\int_{\mathbb{R}} \rho(x, t)\mu(x)\mathrm{d}x\right) & t \in (0, T). \end{cases} \tag{2}$$

Above, μ is a weight function, supported in a compact neighborhood upstream the exit, used to average the density around the exit and the nonincreasing Lipschitz function $p : [0, R] \to \mathbb{R}^+$ models the exit efficiency. This kind of problems has been tremendously studied in the last decades [2, 6, 14, 16, 19]. In particular, the authors of [4, 5] were able to reproduce the main effects linked to the "capacity drop" that are the Braess paradox and the "Faster Is Slower" effect, but not so much the self-organization. Our first goal is to further advance in this direction. We introduce a model which interpolates between two states of the traffic (organized and disorganized) which we represent by the presence of two levels of constraints and by an organization parameter which evolves through an ODE. This model admits a natural and efficient approximation strategy, relying on a combination of splitting, explicit Euler time integration and of a monotone finite volume scheme for LWR. In passing, we prove well-posedness for our model in Sects. 2–3, but our main interset lies in the Sects. 4–5 where we perform a test to validate and discuss the model.

2 Notion of Solution and Uniqueness

Our starting point is the model (2) proposed by the authors of [2], see also [4, 5]. To go further, we introduce two levels of exit efficiencies $p_{\min} \leq p_{\max}$ (both are required to be Lipschitz continuous nonincreasing functions) and set

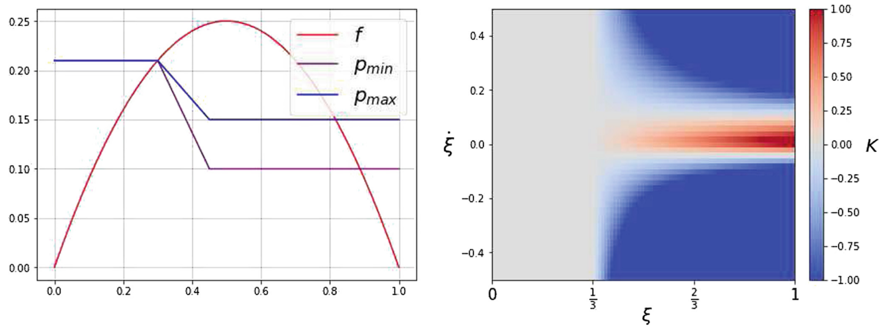


Fig. 1 Typical behavior of exit efficiencies $p_{\min}$, $p_{\max}$ (left) and organization-driving function K in (4) (right)

$$q(t) = (1-\omega(t))p_{\min}(\xi(t)) + \omega(t)p_{\max}(\xi(t)),\ \xi(t) = \int_{\mathbb{R}} \rho(x,t)\mu(x)\mathrm{d}x, \quad (3)$$

where $\omega(t) \in (0, 1)$ is an organization parameter which describes the state of the traffic and evolves through the ODE

$$\dot{\omega}(t) = K\left(\xi(t), \dot{\xi}(t)\right)\omega(t)(1-\omega(t)). \quad (4)$$

Mathematically speaking, we only suppose that $K \in \mathrm{Lip}_{\mathrm{loc}}(\mathbb{R}^2)$. The idea behind phenomenologically relevant choices of K, see Fig. 1(right), is to allow for progressive organization of traffic with time, while keeping the possibility of return to disorganization when sudden and strong variations of the traffic occur; see Sect. 5. For the sake of being definite, in simulations we will choose K under the form

$$K(\xi,\chi) = C\left(\frac{\xi}{\xi_c} - 1\right)^+ \left(1 - \frac{\chi^+}{D_+} - \frac{\chi^-}{D_-}\right), \quad (5)$$

with some positive parameters ξ_c, C, D_+, D_- and the notations $z^+ = \max(z, 0)$, $z^- = |z| - z^+$. This choice will be discussed later. We have the following coupled PDE-ODE system to study:

$$\begin{cases} \partial_t\rho + \partial_x f(\rho) = 0 & \mathbb{R}\times(0,T) \\ f(\rho)|_{x=0} \le q(t) & t\in(0,T), \end{cases} \quad (6)$$

where q is given by (3)–(4). We will denote by Φ the entropy flux associated with the Kružkov entropy $\rho \mapsto |\rho - \kappa|$, for all $\kappa \in [0, R]$, see [23]. Following [6, 14, 16], we give the following definition of solution. Let us underline that the below formulation (i)-(ii)-(iii) is stable with respect to the a.e. convergence of ρ.

Definition 1 A couple (ρ, ω) with $\rho \in \mathrm{L}^\infty(\mathbb{R} \times [0, T]) \cap \mathrm{C}([0, T]; \mathrm{L}^1_{\mathrm{loc}}(\mathbb{R}))$ and $\omega \in \mathrm{W}^{1,\infty}((0, T))$ is called an admissible weak solution to the system (3)–(6) with initial data (ρ_0, ω_0) if
(i) for all non-negative test functions $\varphi \in \mathrm{C}^\infty_{\mathrm{c}}(\mathbb{R} \times [0, T))$ and $\kappa \in [0, R]$, the following entropy inequalities are verified:

$$\int_0^T \int_{\mathbb{R}} |\rho - \kappa| \partial_t \varphi + \Phi(\rho, \kappa) \partial_x \varphi \, \mathrm{d}x\mathrm{d}t + \int_{\mathbb{R}} |\rho_0(x) - \kappa| \varphi(x, 0)\mathrm{d}x + 2 \int_0^T \mathscr{R}(\kappa, q(t))\varphi(0, t)\mathrm{d}t \geq 0, \tag{7}$$

where $\mathscr{R}(\kappa, q(t)) = f(\kappa) - \min\{f(\kappa), q(t)\}$;

(ii) the following weak constraint inequalities are verified:

$$-\int_0^T \int_{\mathbb{R}^+} \rho \partial_t(\varphi\psi) + f(\rho)\partial_x(\varphi\psi) \, \mathrm{d}x\mathrm{d}t \leq \int_0^T q(t)\psi(t)\mathrm{d}t, \;\; \psi \in \mathrm{C}^\infty_{\mathrm{c}}((0, T); \mathbb{R}^+), \;\; \varphi \in \mathrm{C}^\infty_{\mathrm{c}}(\mathbb{R}), \;\; \varphi(0) = 1; \tag{8}$$

(iii) for all $t \in [0, T]$, $\omega(t) = \omega_0 + \displaystyle\int_0^t K\left(\xi(s), \dot{\xi}(s)\right) \omega(s)(1 - \omega(s))\mathrm{d}s$.

Before we prove stability with respect to initial data and uniqueness for admissible weak solutions to the system (3)–(6), let us note that we can directly integrate the ODE (4). This feature is not crucial but it is practical.

Lemma 1 *Fix (ρ, ω) an admissible weak solution to the system (3)–(6). Then for all $t \in [0, T]$,*

$$\omega(t) = \exp(W(t))(1 + \exp(W(t)))^{-1}, \;\; W(t) = \ln(\omega_0) - \ln(1 - \omega_0) + \int_0^t K\left(\xi(s), \dot{\xi}(s)\right) \mathrm{d}s.$$

Theorem 1 *Suppose that f satisfies (1). Fix $\rho^1_0, \rho^2_0 \in \mathrm{L}^1(\mathbb{R}; [0, R])$ and $\omega^1_0, \omega^2_0 \in (0, 1)$. We denote by (ρ^1, ω^1) and (ρ^2, ω^2) two admissible weak solutions to the system (3)–(6) corresponding to the initial data (ρ^1_0, ω^1_0) and (ρ^2_0, ω^2_0), respectively. Then there exist constants A, α, β, γ such that if we note $G(z) = \exp\left(\beta z + \gamma z^2/2\right)$, we have*

$$\text{for a.e. } t \in (0, T), \;\; \|\rho^1(t) - \rho^2(t)\|_{\mathrm{L}^1} \leq \|\rho^1_0 - \rho^2_0\|_{\mathrm{L}^1} G(t) + \alpha|\mathrm{W}^1(0) - \mathrm{W}^2(0)| \int_0^t G(s)\mathrm{d}s \tag{9}$$

and

$$\forall t \in [0, T], \;\; |\omega^1(t) - \omega^2(t)| \leq \left(\frac{|\mathrm{W}^1(0) - \mathrm{W}^2(0)|}{4}\right) + A \int_0^t \left(\alpha|\mathrm{W}^1(0) - \mathrm{W}^2(0)|(t - s) + \|\rho^1_0 - \rho^2_0\|_{\mathrm{L}^1}\right) G(s)\mathrm{d}s, \tag{10}$$

where W^1 and W^2 are defined as in Lemma 1. In particular, the system (3)–(6) admits at most one admissible weak solution.

Proof First, a stability estimate [6, Proposition. 2.10] characteristic of (2) yields Lipschitz continuous dependence $q \mapsto \rho$ for $q \in \mathrm{L}^1(0,T)$ and $\rho \in \mathrm{C}([0,T];\mathrm{L}^1(\mathbb{R}))$. Moreover, the map $\omega \mapsto q$ for $\omega, q \in \mathrm{C}([0,T])$ is obviously Lipschitz. Finally, by exploiting Lemma 1 and the fact that for a.e. $t \in (0,T)$, $\dot{\xi}(t) = \int_{\mathbb{R}} f(\rho)\mu'(x)\mathrm{d}x$, one can obtain Lipschitz dependence $\rho \mapsto \omega$, and Gronwall's lemma concludes.

3 Finite Volume Approximation of the Model

Here, we prove the existence of admissible weak solutions to the system (3)–(6). To do that, we construct and prove the convergence of an explicit Euler in time scheme for the ODE (4) combined with a monotone finite volume scheme for the constraint LWR (6). Fix $\rho_0 \in \mathrm{L}^1(\mathbb{R};[0,R])$ and $\omega_0 \in (0,1)$. For a fixed spatial mesh size Δx and time mesh size Δt, let $x_j = j\Delta x$, $t^n = n\Delta t$. We define the grid cells and $N \in \mathbb{N}$ such that $T \in [N\Delta t, (N+1)\Delta t)$. We write

$$\mathbb{R} \times [0,T] \subset \bigcup_{n=0}^{N} \bigcup_{j\in\mathbb{Z}} \mathscr{P}^n_{j+\frac{1}{2}}, \quad \mathscr{P}^n_{j+\frac{1}{2}} = (x_j, x_{j+1}) \times [t^n, t^{n+1}).$$

Denote by $\left(\rho^0_{j+\frac{1}{2}}\right)_{j\in\mathbb{Z}}$ and $\left(\mu_{j+\frac{1}{2}}\right)_{j\in\mathbb{Z}}$ suitable discretizations of the initial data ρ_0 and of the weight function μ; for instance the mean values on each cell (x_j, x_{j+1}). Initialize with $w^0 = \omega_0$ and $\xi^0 = \sum_{j\in\mathbb{Z}} \rho^0_{j+\frac{1}{2}} \mu_{j+\frac{1}{2}} \Delta x$.

Fix $n \in \{0, \dots, N-1\}$. At each time step, we define a constraint level q^n:

$$q^n = (1 - w^n) p_{\min}(\xi^n) + w^n p_{\max}(\xi^n). \tag{11}$$

We use this value to update the approximate traffic density with the marching formula

$$\forall j \in \mathbb{Z},\ \rho^{n+1}_{j+\frac{1}{2}} = \rho^n_{j+\frac{1}{2}} - \lambda \left(\mathscr{F}^n_{j+1}\left(\rho^n_{j+\frac{1}{2}}, \rho^n_{j+\frac{3}{2}}\right) - \mathscr{F}^n_j\left(\rho^n_{j-\frac{1}{2}}, \rho^n_{j+\frac{1}{2}}\right)\right),\ \lambda = \Delta t/\Delta x, \tag{12}$$

where, given $\mathscr{F}$ a monotone and Lipschitz numerical flux consistent with f, following the recipe of [6, 14], we set

$$\mathscr{F}^n_j(a,b) = \min\left\{\mathscr{F}(a,b), q^n\right\} \text{ if } j = 0, \text{ and } \mathscr{F}^n_j(a,b) = \mathscr{F}(a,b) \text{ if } j \neq 0. \tag{13}$$

Then, setting $\xi^{n+1} = \sum_{j\in\mathbb{Z}} \rho^{n+1}_{j+\frac{1}{2}} \mu_{j+\frac{1}{2}} \Delta x$, we update the organization parameter

$$\chi^{n+1} = \left(\xi^{n+1} - \xi^n\right)/\Delta t, \quad \theta^{n+1} = K\left(\xi^{n+1}, \chi^{n+1}\right) w^n(1-w^n), \quad w^{n+1} = w^n + \theta^{n+1}\Delta t, \tag{14}$$

and finally, define the functions

- $\rho^{\Delta}(x,t) = \rho^n_{j+\frac{1}{2}}$ if $(x,t) \in \mathscr{P}^n_{j+\frac{1}{2}}$, $q^{\Delta}(t) = q^n$ if $t \in [t^n, t^{n+1})$
- $\chi^{\Delta}(t),\ \theta^{\Delta}(t) = \chi^{n+1},\ \theta^{n+1}$ if $t \in [t^n, t^{n+1})$, $\xi^{\Delta}(t) = \xi^0 + \int_0^t \chi^{\Delta}(s)\mathrm{d}s$, $\omega^{\Delta}(t) = w^0 + \int_0^t \theta^{\Delta}(s)\mathrm{d}s$.

Let $\Delta = (\Delta x, \Delta t)$. For the convergence analysis, we will assume that $\Delta \to 0$, with λ verifying the CFL condition

$$\lambda \mathrm{L} \leq 1, \quad \mathrm{L} = \left(\|\partial \mathscr{F} / \partial x\|_{\mathrm{L}^\infty} + \|\partial \mathscr{F} / \partial y\|_{\mathrm{L}^\infty} \right). \tag{15}$$

3.1 Stability and Discrete Entropy Inequalities

Proposition 1 (L^∞ stability) *Given q^n to define the constrained flux in* (13)*, the scheme* (12) *is stable:*

$$\forall n \in \{0, \ldots, N\},\ \forall j \in \mathbb{Z},\ \rho^n_{j+\frac{1}{2}} \in [0, R]. \tag{16}$$

Proof One can follow the argumentation in [24, Proposition. 3.1], or borrow elements from [20, Lemma. 5.1] and [6, Proposition. 4.2].

We now derive discrete entropy inequalities. These inequalities contain terms that will help to pass to the limit in the constrained formulation of the conservation law, as soon as the sequence $(q^{\Delta})_{\Delta}$ of constraints is proved convergent as well. Let us note from now $a \vee b = \max(a,b)$ and $a \wedge b = \min(a,b)$.

Proposition 2 (Discrete entropy inequalities) *The numerical scheme* (11)–(14) *fulfills the following inequalities for all $n \in \{0, \ldots, N-1\}$, $j \in \mathbb{Z}$ and $\kappa \in [0, R]$:*

$$\left(|\rho^{n+1}_{j+\frac{1}{2}} - \kappa| - |\rho^n_{j+\frac{1}{2}} - \kappa|\right)\Delta x + \left(\Phi^n_{j+1} - \Phi^n_j\right)\Delta t \leq \mathscr{R}(\kappa, q^n)\Delta t\ \delta_{j\in\{-1,0\}} + \left(\Phi^n_0 - \overline{\Phi}^n_0\right)\Delta t\ (\delta_{j=-1} - \delta_{j=0}), \tag{17}$$

where

$$\begin{aligned}
\Phi^n_j &= \mathscr{F}(\rho^n_{j-\frac{1}{2}} \vee \kappa, \rho^n_{j+\frac{1}{2}} \vee \kappa) - \mathscr{F}(\rho^n_{j-\frac{1}{2}} \wedge \kappa, \rho^n_{j+\frac{1}{2}} \wedge \kappa) \\
\overline{\Phi}^n_0 &= \min\{\mathscr{F}(\rho^n_{-\frac{1}{2}} \vee \kappa, \rho^n_{\frac{1}{2}} \vee \kappa), q^n\} - \min\{\mathscr{F}(\rho^n_{-\frac{1}{2}} \wedge \kappa, \rho^n_{\frac{1}{2}} \wedge \kappa), q^n\}.
\end{aligned}$$

Proof This is a consequence of the scheme monotonicity. When the constraint does not enter the calculations ie. $j \notin \{-1, 0\}$, the proof follows [20, Lem. 5.4]. When the constraint enters the calculations, the constants $\kappa \in [0, R]$ are no longer stationary solutions of the scheme. Then, calculations make appear the term $\mathscr{R}(\kappa, q^n)$.

Starting from the marching formula (12) and the discrete entropy inequalities (17), we can derive approximate versions of (7) and (8). The proofs can be adapted from the ones of [12, Lem. 4.4] or [24, Propositions. 3.3, 3.4].

Proposition 3 (Approximate entropy/constraint inequalities) *(i) Fix* $\varphi \in \mathrm{C}_{\mathrm{c}}^{\infty}(\mathbb{R} \times [0,T);\mathbb{R}^+)$, $\kappa \in [0,R]$. *Then there exists a constant* $\mathrm{C}_1^{\varphi} = \mathrm{C}_1^{\varphi}(R,T,\mathrm{L})$, *nondecreasing with respect to its arguments, such that*

$$\int_0^T \int_{\mathbb{R}} |\rho^{\Delta} - \kappa|\partial_t\varphi + \Phi^{\Delta}\left(\rho^{\Delta},\kappa\right)\partial_x\varphi \mathrm{d}x\mathrm{d}t + \int_{\mathbb{R}} |\rho_0^{\Delta}(x) - \kappa|\varphi(x,0)\mathrm{d}x + 2\int_0^T \mathscr{R}\left(\kappa, q^{\Delta}(t)\right)\varphi(0,t)\mathrm{d}t \geq -\mathrm{C}_1^{\varphi}(\Delta t + \Delta x). \tag{18}$$

(ii) Fix $\psi \in \mathrm{C}_{\mathrm{c}}^{\infty}((0,T);\mathbb{R}^+)$ *and* $\varphi \in \mathrm{C}_{\mathrm{c}}^{\infty}(\mathbb{R})$ *such that* $\varphi(0) = 1$. *Then there exists a constant* $\mathrm{C}_2^{\varphi,\psi} = \mathrm{C}_2^{\varphi,\psi}(R,T,\mathrm{L},\|f\|_{\mathrm{L}^\infty})$, *nondecreasing with respect to its arguments, such that*

$$-\int_0^T \int_{\mathbb{R}^+} \rho^{\Delta}\partial_t(\varphi\psi) + \mathscr{F}^{\Delta}(\rho^{\Delta})\partial_x(\varphi\psi)\,\mathrm{d}x\mathrm{d}t \leq \int_0^T q^{\Delta}(t)\psi(t)\mathrm{d}t + \mathrm{C}_2^{\varphi,\psi}(\Delta x + \Delta t), \tag{19}$$

where

$$\Phi^{\Delta}\left(\rho^{\Delta},\kappa\right) = \sum_{n=0}^{N-1}\sum_{j\in\mathbb{Z}} \Phi_j^n \mathbb{1}_{\mathscr{P}_{j+\frac{1}{2}}^n}(x,t), \quad \mathscr{F}^{\Delta}(\rho^{\Delta}) = \sum_{n=0}^{N-1}\sum_{j\in\mathbb{Z}} \mathscr{F}(\rho_{j-\frac{1}{2}}^n, \rho_{j+\frac{1}{2}}^n)\mathbb{1}_{\mathscr{P}_{j+\frac{1}{2}}^n}(x,t).$$

The final step is to obtain compactness for the sequences $(\rho^{\Delta})_{\Delta}$ and $(\omega^{\Delta})_{\Delta}$ in order to pass to the limit in (18)–(19).

3.2 Compactness and Convergence

Exploiting the compact embedding of $\mathrm{W}^{1,\infty}(0,T)$ in $\mathrm{C}([0,T])$, we can prove the existence of $\xi, \omega \in \mathrm{C}([0,T])$ such that (up to the extraction of a subsequence) $(\xi^{\Delta})_{\Delta}$ and $(\omega^{\Delta})_{\Delta}$ converge uniformly to ξ and ω, respectively. There are many ways to prove compactness of the sequence $(\rho^{\Delta})_{\Delta}$. For example, one can derive weak BV estimates [6, 20] or use the singular mapping technique [1, 15]. Here, since the conservation law in (6) is invariant under a translation in time, we derive local BV bounds, following [10, Lemma. 4.2].

Proposition 4 *Assume that* $\rho_0 \in \mathrm{BV}(\mathbb{R})$. *Fix* $0 < \varepsilon < X$ *and let* $\Omega(\varepsilon, X)$ *be the open subset* $\Omega(\varepsilon,X) = (-X,-\varepsilon) \cup (\varepsilon, X)$. *There exist two constants* C_3 *and* $\mathrm{C}_4 > 0$ *such that for all* $t \in [0, T - \Delta t)$,

$$\mathrm{TV}(\rho^{\Delta}(t)_{|\Omega(\varepsilon,X)}) \leq \mathrm{TV}(\rho_0) + \frac{\mathrm{C}_3}{\varepsilon}, \quad \int_{\Omega(\varepsilon,X)} |\rho^{\Delta}(x,t+\Delta t) - \rho^{\Delta}(x,t)|\mathrm{d}x \leq \mathrm{C}_4\Delta t.$$

Therefore, up to a subsequence, $(\rho^{\Delta})_{\Delta}$ *converges a.e. on* $\mathbb{R} \times (0,T)$ *to some* $\rho \in \mathrm{L}^{\infty}(\mathbb{R} \times [0,T])$.

Remark 1 At this point, the link between ξ and ρ is established: for a.e. $t \in (0, T)$, $\xi(t) = \displaystyle\int_{\mathbb{R}} \rho(x,t)\mu(x)\mathrm{d}x$.

Theorem 2 *Fix* $\rho_0 \in \mathrm{L}^1(\mathbb{R}; [0, R]) \cap \mathrm{BV}(\mathbb{R})$ *and* $\omega_0 \in (0, 1)$. *Suppose that* f *satisfies* (1) *and that* $\mu \in \mathrm{W}^{2,\infty}(\mathbb{R}^{-*})$. *Then under the CFL condition* (15)*, the scheme* (11)–(14) *converges to an admissible weak solution to the system* (3)–(6).

Proof We show that the couple (ρ, ω) is a solution in the sense of Definition 1. First, let apply inequality (18) with $\varphi \in \mathrm{C}_\mathrm{c}^\infty(\mathbb{R}^* \times [0, T); \mathbb{R}^+)$ and $\kappa \in [0, R]$ to obtain

$$\int_0^T \int_{\mathbb{R}} |\rho^\Delta - \kappa|\partial_t\varphi + \Phi^\Delta(\rho^\Delta, \kappa)\partial_x\varphi \mathrm{d}x\mathrm{d}t + \int_{\mathbb{R}} |\rho_0^\Delta - \kappa|\varphi(x, 0)\mathrm{d}x \geq -\mathrm{C}_1^\varphi(\Delta x + \Delta t).$$

Then when letting $\Delta \to 0$, the a.e. convergence of $(\rho^\Delta)_\Delta$ to ρ ensures that ρ verifies (7) away from the interface. Consequently, $\rho \in \mathrm{C}([0, T]; \mathrm{L}^1_{\mathrm{loc}}(\mathbb{R}^*))$, see [11, Thm. 1.2]. Moreover, since ρ is bounded and $\{x = 0\}$ has a Lebesgue measure 0, $\rho \in \mathrm{C}([0, T]; \mathrm{L}^1_{\mathrm{loc}}(\mathbb{R}))$. It ensures that the equality in Remark 1 actually holds for all $t \in [0, T]$. Moreover, since ρ is an entropy solution in $\mathbb{R}^* \times (0, T)$ to $\partial_t\rho + \partial_x f(\rho) = 0$, ξ defined in Remark 1 is actually in $\mathrm{W}^{1,\infty}(0, T)$ and verifies for a.e. $t \in (0, T)$,

$$\dot{\xi}(t) = \int_{\mathbb{R}} f(\rho)\mu'(x)\mathrm{d}x.$$

(i)–(ii) The uniform convergence of both $(\xi^\Delta)_\Delta$ and $(\omega^\Delta)_\Delta$ ensures the existence of $q \in \mathrm{C}([0, T])$ such that $(q^\Delta)_\Delta$ converges to q a.e. on $(0, T)$. Consequently, for a.e. $t \in (0, T)$,

$$q(t) = (1 - \omega(t))p_{\min}(\xi(t)) + \omega(t)p_{\max}(\xi(t)),$$

and this equality actually holds for all $t \in [0, T]$ by continuity. Then by letting $\Delta \to 0$ in (18)–(19), we obtain that (ρ, ω) verifies the entropy inequalities (7) and the weak constraint inequalities (8).

(iii) An important step towards the assessment of the weak ODE formulation for ω is to show that $(\chi^\Delta)_\Delta$ converges a.e. to $\dot{\xi}$. One way to do that is by using a discrete integration by parts, assuming that $\mu \in \mathrm{W}^{2,\infty}(\mathbb{R}^{-*})$ (cf. [4]).

Corollary 1 *Fix* $\rho_0 \in \mathrm{L}^1(\mathbb{R}; [0, R]) \cap \mathrm{BV}(\mathbb{R})$ *and* $\omega \in (0, 1)$. *Suppose that* f *satisfies* (1) *and that* $\mu \in \mathrm{W}^{2,\infty}(\mathbb{R}^{-*})$. *Then the system* (3)–(6) *admits a unique admissible weak solution.*

Proof Uniqueness comes from Theorem 1, existence comes from Theorem 2, with a constructive proof.

Remark 2 Adopting the formalism proposed in [5], one could also prove well-posedness with fixed point arguments.

4 Numerical Simulations

We report on numerical experiments with the scheme described in Sect. 3. We take the normalized uniformly concave flux $f(\rho) = \rho(1-\rho)$. We choose to use the Godunov flux at the interface ($j = 0$ in (13)) and the Rusanov flux away from the interface ($j \neq 0$ in (13)). The function $x \mapsto 2n(x + 1/n)\mathbb{1}_{[-1/n,0]}(x)$ (with $n = 3$) is issued as weight function. Following [5, Sect. 7], the setup for our simulation is as follows. We consider the domain of computation $[-5, 1]$, the initial data $\rho_0(x) = \mathbb{1}_{[-4,-2]}(x)$, $\omega_0 = 0.2$ and the efficiencies of the exit $p_{\min}$, $p_{\max}$ are represented in Fig. 1(left). For the simulations, we have fixed a locally Lipschitz prefactor K in (4) with behaviour depicted in Fig. 1(right) and parameters $\xi_c = 1/3$, $C = 2/3$, $D_+ = 1/10$ and $D_- = D_+/2$. The phenomenological features encoded in this choice will be addressed in Sect. 5.

We first address the error analysis in the above setup. Introduce the relative error $E_\rho^\Delta = \|\rho^\Delta - \rho^{\Delta/2}\|_{L^1((0,T);L^1(\mathbb{R}))}$. In Table 1, we computed this error for different number of space cells at the final time $T = 17$. We deduce that the order of convergence is approximately 0.852.

Now, let us comment on qualitative features of the simulated traffic flow and provide its interpretation in terms of agents' behaviors. First, as we can see in Fig. 3, the introduction of the organization parameter favors the evacuation time. Figure 2 highlights the fact that the model reproduces some features expected from self-organization. At first, the exit flux increases until it reaches the maximum level of

Table 1 Measured errors at time $T = 17$

Number of cells	640	1280	2560	5120	10240	20480
E_ρ^Δ	1.863×10^{-1}	1.158×10^{-1}	6.507×10^{-2}	3.335×10^{-2}	2.105×10^{-2}	9.501×10^{-3}

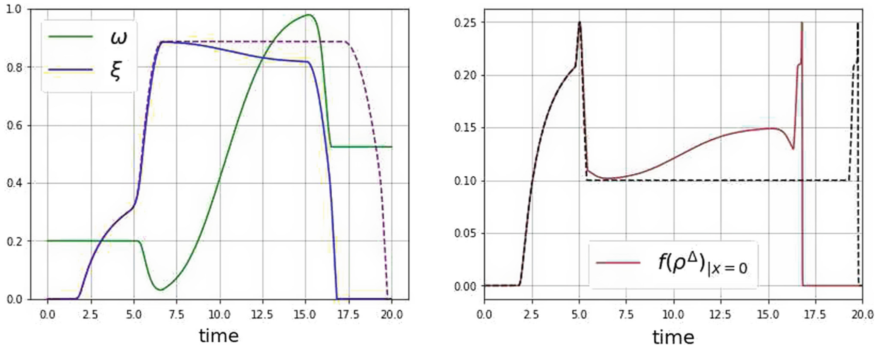


Fig. 2 Left: subjective density ξ and organization marker ω. Right: exit flux $f(\rho)|_{x=0^-}$; dashed lines correspond to the reference solution in absence of self-organization $\omega = 0$ in (3)

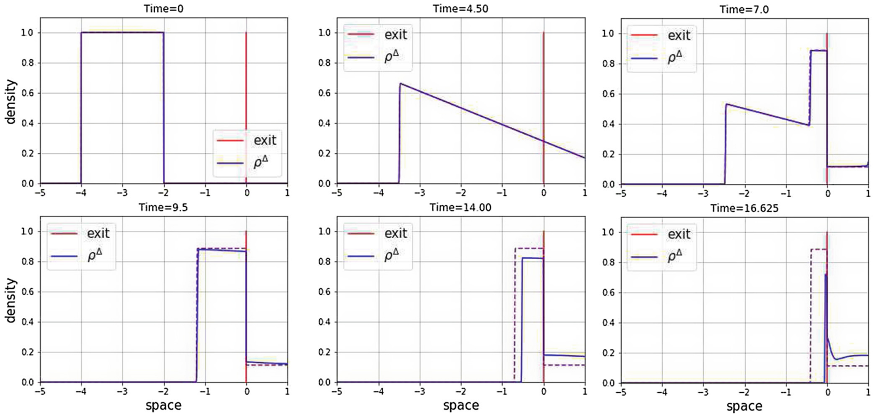


Fig. 3 The numerically computed solution $x \mapsto \rho^\Delta(x, t)$ at different fixed times t; dashed lines correspond to the reference solution in absence of self-organization $\omega = 0$ in (3)

the exit efficiency. As traffic densifies, the exit flux falls down to the lowest value of this efficiency, which reflects rapid disorganization, ie., predominance of agents' individualistic strategies over the rational collective behavior. Then, in the time interval [6, 16], the elevated density upstream has very small variations which leads to the emergence of a coherent collective behavior of the agents. This is witnessed through the increase of both the organization marker and the exit flux. We stress out that without self-organization, the exit flux keeps its minimal value in this time interval. Then a notable phenomenon seems to take place. In the time interval [15.5, 16.3], the jam upstream the exit starts to resorb, and the exit efficiency (which is monitored by the exit flux) slightly falls down while the organization level regresses significantly. In other words, the agents abandon collective strategies in rapidly evolving environments, but this does not affect the traffic dramatically because densities are also strongly decreased.

5 Conclusions and Perspectives

The model we propose here permits a rigorous analysis of well-posedness as well as a robust and simple numerical approximation. It enriches the qualitative behavior of the simple LWR-based models for bottlenecks [2, 5, 16], due to its ability to reproduce a few self-organization features. Let us deeper discuss the model construction, in particular the role of the function K whose behavior is depicted in Fig. 1(right). Its key features are as follows:

- invariance of the organization marker ω in the region of low densities;
- rapid decrease of ω for moderate and particularly for high densities, under strong density variations;
- progressive increase of ω in dense and very dense traffic with small density variations.

The idea behind these features is: rapidly changing traffic conditions, at considerable densities, promote individual behavior and rapidly lead to a somewhat chaotic interactions among agents, thus lowering the exit efficiency; while persistent coercive traffic conditions, such as a jam, help to emerge and promote a collective behavior like formation of well-organized queues, the alternate in the order of passage through the bottleneck, and a higher degree of mutual courtesy among agents; thus the exit efficiency improves accordingly, which enhances the jam evacuation. The form (5) provides a simple example of such behavior, which is confirmed by the simulations of Sect. 4. The parameter ξ_c has the meaning of activation threshold for organization/disorganization of the traffic at bottleneck; D_+, D_- indicate thresholds of transition from cooperative (low variations of ξ) to individualistic (higher ones) dynamics of agents.

One way to improve this model would be to take into account unexpected/rash behavior of certain agents. Let us recall that unlike fluid mechanics models, traffic models deal with a relatively small number of agents. In consequence, we would expect the dynamics to be greatly impacted by the behavior of a few agents. An idea to model such rash behaviors is to introduce a stochastic term in the definition of the prefactor K, for example

$$K(t, \xi, \chi) = C\left(\frac{\xi}{\xi_c} - 1\right)^+ \left(1 - \frac{\chi^+}{D_+} - \frac{\chi^-}{D_-} - X(t)\right),$$

where X is a stochastic process modeling the harmful impact of a random number of mindless agents on the collective dynamics. We plan to study numerically this variant of the model and provide indications concerning the impact of undisciplined agents on the evacuation time.

In the forthcoming work [7] we will take inspiration from second-order macroscopic models of traffic [8, 25] to model self-organization globally on the road; note that bottlenecks can be as well modelled with non-local point constraints within such models, see, e.g., [3]. Mimicking the key elements (3)–(4) of the model we addressed in the present note, we will introduce two fundamental graphs $f_{\min} \le f_{\max}$ to describe the two states of the traffic and make the space-and-time dependent organization parameter act both on the constraint levels (3) and on the fundamental graphs. We will then have to study a variant of nonlocal LWR model, cf. [9, 21] for related mathematical and numerical issues.

References

1. Adimurthi, Jaffré, J., Veerappa Gowda, G.D.: Godunov-type methods for conservation laws with a flux function discontinuous in space. SIAM J. Numer. Anal. **42**(1), 179–208 (2005)
2. Andreianov A, Donadello C, Rosini MD (2014) Crowd dynamics and conservation laws with nonlocal constraints and capacity drop. Math. Models Methods Appl. Sci. 24:2685–2722
3. Andreianov A, Donadello C, Rosini MD (2016) A second-order model for vehicular traffics with local point constraints on the flow. Math. Models Methods Appl. Sci. 26(4):751–802

4. Andreianov, A., Donadello, C., Razafison, U., Rosini, M.D.: Qualitative behaviour and numerical approximation of solutions to conservation laws with non-local point constraints on the flux and modeling of crowd dynamics at the bottlenecks. ESAIM: M2AN **50**, 1269–1287 (2016)
5. Andreianov A, Donadello C, Razafison U, Rosini MD (2018) Analysis and approximation of one-dimensional scalar conservation laws with general point constraints on the flux. J. Math. Pures et Appl. 116:309–346
6. Andreianov A, Goatin P, Seguin N (2010) Finite volume schemes for locally constrained conservation laws. Numer. Math 115(4):609–645
7. Andreianov, B., Sylla, A.: A hybrid LWR model to reproduce self-organization of traffic. In preparation
8. Aw A, Rascle M (2000) Resurrection of "second order" models of traffic flow. SIAM J. Appl. Math 60(3):916–938
9. Blandin S, Goatin P (2016) Well-posedness of a conservation law with non-local flux arising in traffic flow modeling. Numer. Math 132(2):217–241
10. Bürger R, García A, Karlsen KH, Towers JD (2008) A family of numerical schemes for kinematic flows with discontinuous flux. J. Eng. Math. 60:387–425
11. Cancès C, Gallouët T (2011) On the time continuity of entropy solutions. J. Evol. Equ. 11(1):43–55
12. Cancès C, Seguin N (2012) Error estimate for Godunov approximation of locally constrained conservation laws. SIAM J. Numer. Anal. 50(6):3036–3060
13. Cepolina EM (2009) Phased evacuation: an optimisation model which takes into account the capacity drop phenomenon in pedestrian flows. Fire Saf. J. 44(4):532–544
14. Chalons, C., Goatin, P., Seguin, N.: General constrained conservation laws. Application to pedestrian flow modeling. Netw Heterogen. Media **8**(2), 433–463 (2013)
15. Coclite GM, Risebro NH (2005) Conservation laws with time dependent discontinuous coefficients. SIAM J. Math. Anal. 36(4):1293–1309
16. Colombo RN, Goatin P (2007) A well posed conservation law with a variable unilateral constraint. J. Differ. Equ. 234(2):654–675
17. Colombo RN, Rosini MD (2005) Pedestrian flows and non-classical shocks. Math. Methods Appl. Sci. 28(13):1553–1567
18. Cristiani, E., Piccoli, B., Tosin, A.: How can macroscopic models reveal self-organization in traffic flow? In: 51st IEEE Conference on Decision and Control, pp. 6989–6994 (2012)
19. Delle Monache ML, Goatin P (2014) Scalar conservation laws with moving constraints arising in traffic flow modeling: an existence result. J. Differ. Equ. 257(11):4015–4029
20. Eymard, R., Gallouët, T., Herbin, R.: Finite volume methods. In Handbook of Numerical Analysis, North-Holland, Amsterdam, pp. 713–1020 (2000)
21. Goatin P, Scialanga S (2016) Well-posedness and finite volume approximations of the LWR traffic flow model with non-local velocity. Netw. Heterogen. Media 11(1):107–121
22. Kerner BS (1998) Experimental features of self-organization in traffic flow. Phys. Rev. Lett. 81(17):3797–3800
23. Kružkov SN (1970) First order quasilinear equations with several independent variables. Mat. Sb. 10(2):217–243
24. Sylla, A.: Influence of a slow moving vehicle on traffic: well-posedness for a mildly non-local model (2019). https://hal.archives-ouvertes.fr/hal-02391844
25. Zhang HM (2002) A non-equilibrium traffic model devoid of gas-like behavior. Transport. Res. Part B 36:275–290

A Three-Dimensional Hybrid High-Order Method for Magnetostatics

Florent Chave, Daniele A. Di Pietro, and Simon Lemaire

Abstract We introduce a three-dimensional Hybrid High-Order method for magnetostatic problems. The proposed method is easy to implement, supports general polyhedral meshes, and allows for arbitrary orders of approximation.

Keywords Hybrid High-Order methods · Polyhedral meshes · Nonconforming methods · Magnetostatics

MSC (2010) 65N08 · 65N12 · 65N30

1 Introduction

Let $\Omega \subset \mathbb{R}^3$ denote an open, bounded and connected polyhedral domain, with boundary $\partial\Omega$ and unit outward normal $\mathbf{n}$. We assume that Ω is topologically trivial, and that $\partial\Omega$ is connected. For any $X \subset \overline{\Omega}$, we denote by $(\cdot, \cdot)_X$ and $||\cdot||_X$ the usual inner product and norm on $L^2(X; \mathbb{R}^l)$, $l \in \{1, 2, 3\}$. The standard magnetostatic problem consists in finding the magnetic field $\mathbf{u} : \Omega \to \mathbb{R}^3$ such that

F. Chave (✉) · S. Lemaire
Inria, Univ. Lille, CNRS, UMR 8524 – Laboratoire Paul Painlevé,
F-59000 Lille, France
e-mail: florent.chave@inria.fr

S. Lemaire
e-mail: simon.lemaire@inria.fr

D. A. Di Pietro
IMAG, Univ Montpellier, CNRS,
Montpellier, France
e-mail: daniele.di-pietro@umontpellier.fr

R. Klöfkorn et al. (eds.), *Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples*, Springer Proceedings in Mathematics & Statistics 323,
https://doi.org/10.1007/978-3-030-43651-3_22

$$\mathbf{curl}\,\mathbf{u} = \mathbf{f} \quad \text{in } \Omega, \tag{1a}$$

$$\operatorname{div}\mathbf{u} = 0 \quad \text{in } \Omega, \tag{1b}$$

$$\mathbf{n}\times\mathbf{u}\times\mathbf{n} = \mathbf{0} \quad \text{on } \partial\Omega, \tag{1c}$$

where $\mathbf{H}(\mathrm{div};\Omega) \ni \mathbf{f} : \Omega \to \mathbb{R}^3$ denotes the current density and is such that $\operatorname{div}\mathbf{f} = 0$ in Ω and $\mathbf{f}\cdot\mathbf{n} = 0$ on $\partial\Omega$. We supplement Problem (1) with another unknown, namely the potential $p : \Omega \to \mathbb{R}$, that satisfies $p = 0$ in Ω. From now on, Problem (1) refers to this augmented problem with unknowns $(\mathbf{u}, p)$. The starting point of our discretization is the following equivalent weak formulation of Problem (1), originally introduced in [8, Eq. (58)]: Find $(\mathbf{u}, p) \in \mathbf{X_0} \times Y_0$ such that

$$a(\mathbf{u},\mathbf{v}) + b(\mathbf{v}, p) = (\mathbf{f}, \mathbf{curl}\,\mathbf{v})_\Omega \qquad \forall \mathbf{v} \in \mathbf{X_0}, \tag{2a}$$

$$-b(\mathbf{u}, q) + c(p, q) = 0 \qquad \forall q \in Y_0, \tag{2b}$$

where

$$\mathbf{X_0} := \{\mathbf{v} \in \mathbf{H}(\mathbf{curl};\Omega) \ : \ \mathbf{n}\times\mathbf{v}\times\mathbf{n} = \mathbf{0} \text{ on } \partial\Omega\}, \qquad Y_0 := H^1_0(\Omega),$$

and the bilinear forms $a : \mathbf{H}(\mathbf{curl};\Omega) \times \mathbf{H}(\mathbf{curl};\Omega) \to \mathbb{R}$, $b : \mathbf{H}(\mathbf{curl};\Omega) \times H^1(\Omega) \to \mathbb{R}$, and $c : H^1(\Omega) \times H^1(\Omega) \to \mathbb{R}$ are given by

$$a(\mathbf{w},\mathbf{v}) := (\mathbf{curl}\,\mathbf{w}, \mathbf{curl}\,\mathbf{v})_\Omega, \qquad b(\mathbf{w}, q) := (\mathbf{w}, \nabla q)_\Omega, \qquad c(r, q) := (r, q)_\Omega. \tag{3}$$

Testing (2a) with $\mathbf{v} = \nabla p \in \mathbf{X_0}$, it is inferred that $p = 0$ in Ω. The well-posedness of Problem (2) is then a consequence of the coercivity of a on the subspace of $\mathbf{X_0}$ given by $\{\mathbf{w} \in \mathbf{X_0} \ : \ b(\mathbf{w}, q) = 0 \ \ \forall q \in Y_0\} = \{\mathbf{w} \in \mathbf{X_0} \ : \ \operatorname{div}\mathbf{w} = 0\}$ which, in turn, follows from the first Weber inequality (see, e.g., [1, Theorem 3.4.3]).

Various discretization methods have been studied in the literature to approximate the Maxwell equations. We can, in particular, cite the seminal work of [9] on simplicial elements. On more general element shapes, one can mention the Discontinuous Galerkin method of [11], the Hybridizable Discontinuous Galerkin (HDG) methods of [10] and [3], or the Virtual Element method of [12].

In this paper, we devise an easy-to-implement Hybrid High-Order (HHO) method to solve Problem (1). HHO methods have been originally introduced in [6, 7]. Their connections with HDG methods have been later discussed in [4] in the context of scalar variable diffusion problems. The method we introduce here shares some similarities with the HDG method of [3]. It indeed hinges, as in [3], on face unknowns for the magnetic field belonging to a subtle subspace of $\mathbb{P}^{k+1}(F;\mathbb{R}^2)$. However, there are two main differences between our method and the one in [3]. First, taking advantage of the fact that Problem (1) is actually first-order, we do not (locally) reconstruct a discrete **curl** operator. We hence (i) can consider a smaller local set of face unknowns, and (ii) we do not have to solve a local problem on each mesh cell (which may become, for a sequential implementation, rather costly in 3D, especially for large polynomial degrees). Second, and as opposed to [3] in which the bilinear form c is not introduced (therein, p may be nonzero and the authors are also interested in its approximation, which is not our case), we consider the formulation (2) of Problem (1). At the discrete level, it enables to improve the stability of the method without jeopardizing the approximation of $\mathbf{u}$.

The rest of the paper is organized as follows. In Sect. 2 we describe the discrete setting and our HHO discretization. In Sect. 3 we state the discrete problem and discuss its well-posedness. Finally, in Sect. 4, we numerically validate the proposed method.

2 Hybrid High-Order Discretization

2.1 Discrete Setting

We consider sequences of refined meshes that are admissible in the sense of [5, Definition 1.9]. Each mesh $\mathscr{T}_h$ in the sequence is a finite collection $\{T\}$ of nonempty, disjoint, open polyhedra that are assumed to be star-shaped with respect to some interior point. There holds $\overline{\Omega} = \bigcup_{T\in\mathscr{T}_h} \overline{T}$ with $h = \max_{T\in\mathscr{T}_h} h_T$, where h_T denotes the diameter of the cell T. For all $T \in \mathscr{T}_h$, the boundary of T is decomposed into planar faces collected in the set $\mathscr{F}_T$. For admissible mesh sequences, $\operatorname{card}(\mathscr{F}_T)$ is bounded uniformly in h. Interfaces are collected in the set $\mathscr{F}_h^{\mathrm{i}}$, boundary faces in the set $\mathscr{F}_h^{\mathrm{b}}$, and we define $\mathscr{F}_h := \mathscr{F}_h^{\mathrm{i}} \cup \mathscr{F}_h^{\mathrm{b}}$. For all $T \in \mathscr{T}_h$ and all $F \in \mathscr{F}_T$, the diameter of F is denoted h_F and the unit normal to F pointing outward T is denoted $\mathbf{n}_{TF}$. For admissible mesh sequences, h_F is uniformly comparable to h_T.

2.2 Discrete Unknowns

Let an arbitrary polynomial degree $k \geq 0$ be given. For $X \in \{F, T\}$ and, respectively, $d \in \{2, 3\}$, and for $l \in \{1, 2, 3\}$, we denote by $\mathbb{P}^k(X; \mathbb{R}^l)$ the vector space of d-variate, l-valued polynomial functions on X of total degree at most k. When $l = 1$, we simply write $\mathbb{P}^k(X)$. The global sets of discrete unknowns for the magnetic field and the potential are given by

$$\underline{\mathbf{X}}_h^{k+1} := \left\{ \underline{\mathbf{v}}_h = \big((\mathbf{v}_T)_{T\in\mathscr{T}_h}, (\mathbf{v}_F)_{F\in\mathscr{F}_h}\big) \ : \ \begin{array}{ll} \mathbf{v}_T \in \mathbb{P}^{k+1}(T; \mathbb{R}^3) & \forall T \in \mathscr{T}_h \\ \mathbf{v}_F \in \nabla_\tau \, \mathbb{P}^{k+2}(F) & \forall F \in \mathscr{F}_h \end{array} \right\},$$

$$\underline{\mathrm{Y}}_h^{k+1} := \left\{ \underline{\mathrm{q}}_h = \big((\mathrm{q}_T)_{T\in\mathscr{T}_h}, (\mathrm{q}_F)_{F\in\mathscr{F}_h}\big) \ : \ \begin{array}{ll} \mathrm{q}_T \in \mathbb{P}^k(T) & \forall T \in \mathscr{T}_h \\ \mathrm{q}_F \in \mathbb{P}^{k+1}(F) & \forall F \in \mathscr{F}_h \end{array} \right\},$$

where, for all $F \in \mathscr{F}_h$, $\nabla_\tau \, \mathbb{P}^{k+2}(F)$ denotes the space of (tangential) gradients of polynomials of degree $k + 2$ on F. For all $\underline{\mathbf{v}}_h \in \underline{\mathbf{X}}_h^{k+1}$, $\mathbf{v}_h$ (not underlined) denotes the function in the broken space $\mathbb{P}^{k+1}(\mathscr{T}_h; \mathbb{R}^3)$ such that $\mathbf{v}_{h|T} := \mathbf{v}_T$ for all $T \in \mathscr{T}_h$.

Remark 1 In [3], the authors consider face unknowns $\mathbf{v}_F$ in the larger space

$$\mathbb{P}^k(F;\mathbb{R}^2) \oplus \nabla_\tau \widetilde{\mathbb{P}}^{k+2}(F),$$

where $\widetilde{\mathbb{P}}^{k+2}(F)$ is the space of homogeneous polynomials of degree $k+2$ on F.

We define the interpolators $\underline{\mathbf{I}}_{\mathbf{X},h}^{k+1} : H^1(\Omega;\mathbb{R}^3) \to \underline{\mathbf{X}}_h^{k+1}$ and $\underline{\mathrm{I}}_{\mathrm{Y},h}^{k+1} : H^1(\Omega) \to \underline{\mathrm{Y}}_h^{k+1}$ such that, for any $\mathbf{v} \in H^1(\Omega;\mathbb{R}^3)$ and $q \in H^1(\Omega)$,

$$\underline{\mathbf{I}}_{\mathbf{X},h}^{k+1}\mathbf{v} := \left(\left(\pi_T^{k+1}(\mathbf{v}_{|T})\right)_{T\in\mathscr{T}_h}, \left(\pi_F^{k+1,\nabla}\gamma_\tau(\mathbf{v}_{|F})\right)_{F\in\mathscr{F}_h}\right), \tag{4a}$$

$$\underline{\mathrm{I}}_{\mathrm{Y},h}^{k+1}q := \left(\left(\pi_T^{k}(q_{|T})\right)_{T\in\mathscr{T}_h}, \left(\pi_F^{k+1}(q_{|F})\right)_{F\in\mathscr{F}_h}\right), \tag{4b}$$

where (i) $\gamma_\tau(\mathbf{v}_{|F}) \in L^2(F;\mathbb{R}^2)$ denotes the tangential trace of $\mathbf{v} \in H^1(\Omega;\mathbb{R}^3)$ on F, (ii) for $X \in \{F,T\}$ and $q \in \mathbb{N}$, π_X^q denotes, for $l \in \{1,2,3\}$, the $L^2(X;\mathbb{R}^l)$-orthogonal projector onto $\mathbb{P}^q(X;\mathbb{R}^l)$, and (iii) $\pi_F^{k+1,\nabla}$ denotes the $L^2(F;\mathbb{R}^2)$-orthogonal projector onto $\nabla_\tau \mathbb{P}^{k+2}(F)$. In what follows, we denote by π_h^q the global L^2-orthogonal projector such that, for all $T \in \mathscr{T}_h$, $\pi_{h|T}^q := \pi_T^q$.

We finally introduce the following global sets of discrete unknowns, that enforce the zero Dirichlet boundary conditions:

$$\underline{\mathbf{X}}_{h,\mathbf{0}}^{k+1} := \left\{\underline{\mathbf{v}}_h \in \underline{\mathbf{X}}_h^{k+1} \ : \ \mathbf{v}_F \equiv \mathbf{0}\ \forall F \in \mathscr{F}_h^{\mathrm{b}}\right\},$$

$$\underline{\mathrm{Y}}_{h,0}^{k+1} := \left\{\underline{\mathrm{q}}_h \in \underline{\mathrm{Y}}_h^{k+1} \ : \ \mathrm{q}_F \equiv 0\ \forall F \in \mathscr{F}_h^{\mathrm{b}}\right\}.$$

2.3 Discrete Bilinear Forms

The discrete counterpart of the bilinear form a defined in (3) is the bilinear form $\mathrm{a}_h : \underline{\mathbf{X}}_h^{k+1} \times \underline{\mathbf{X}}_h^{k+1} \to \mathbb{R}$ given by

$$\mathrm{a}_h(\underline{\mathbf{w}}_h, \underline{\mathbf{v}}_h) := (\mathbf{curl}_h\, \mathbf{w}_h, \mathbf{curl}_h\, \mathbf{v}_h)_\Omega + \mathrm{s}_h(\underline{\mathbf{w}}_h, \underline{\mathbf{v}}_h), \tag{5}$$

where $\mathbf{curl}_h$ denotes the broken $\mathbf{curl}$ operator on $\mathscr{T}_h$ and $\mathrm{s}_h : \underline{\mathbf{X}}_h^{k+1} \times \underline{\mathbf{X}}_h^{k+1} \to \mathbb{R}$ is the stabilization bilinear form such that

$$\mathrm{s}_h(\underline{\mathbf{w}}_h, \underline{\mathbf{v}}_h) := \sum_{T\in\mathscr{T}_h}\sum_{F\in\mathscr{F}_T} h_F^{-1}\left(\pi_F^{k+1,\nabla}(\mathbf{w}_F - \gamma_\tau(\mathbf{w}_{T|F})), \pi_F^{k+1,\nabla}(\mathbf{v}_F - \gamma_\tau(\mathbf{v}_{T|F}))\right)_F.$$

On the other hand, the discrete coupling bilinear form $\mathrm{b}_h : \underline{\mathbf{X}}_h^{k+1} \times \underline{\mathrm{Y}}_h^{k+1} \to \mathbb{R}$ is given by

$$\mathrm{b}_h(\underline{\mathbf{w}}_h, \underline{\mathrm{q}}_h) := \sum_{T\in\mathscr{T}_h}\left(-(\mathrm{q}_T, \operatorname{div}\mathbf{w}_T)_T + \sum_{F\in\mathscr{F}_T}\left(\mathrm{q}_F, \mathbf{w}_{T|F}\cdot\mathbf{n}_{TF}\right)_F\right). \tag{6}$$

From the definitions (6) and (4b) of, respectively, b_h and $\underline{\mathrm{I}}_{\mathrm{Y},h}^{k+1}$, one can easily prove the following commutation property: For any $q \in H^1(\Omega)$,

$$\mathrm{b}_h(\underline{\mathbf{w}}_h, \underline{\mathrm{I}}_{\mathrm{Y},h}^{k+1} q) = (\mathbf{w}_h, \nabla q)_\Omega \qquad \text{for all } \underline{\mathbf{w}}_h \in \underline{\mathbf{X}}_h^{k+1}. \tag{7}$$

Finally, the discrete counterpart of the bilinear form c is the bilinear form $\mathrm{c}_h : \underline{\mathrm{Y}}_h^{k+1} \times \underline{\mathrm{Y}}_h^{k+1} \to \mathbb{R}$ given by

$$\mathrm{c}_h(\underline{\mathrm{r}}_h, \underline{\mathrm{q}}_h) := \sum_{T\in\mathscr{T}_h}\left((\mathrm{r}_T, \mathrm{q}_T)_T + \sum_{F\in\mathscr{F}_T} h_F(\mathrm{r}_F, \mathrm{q}_F)_F\right).$$

One can easily see that $\mathrm{c}_h(\cdot,\cdot)^{1/2}$ defines a norm on $\underline{\mathrm{Y}}_h^{k+1}$.

3 Discrete Problem

Our HHO discretization of Problem (2) reads: Find $(\underline{\mathbf{u}}_h, \underline{\mathrm{p}}_h) \in \underline{\mathbf{X}}_{h,\mathbf{0}}^{k+1} \times \underline{\mathrm{Y}}_{h,0}^{k+1}$ such that

$$\mathrm{a}_h(\underline{\mathbf{u}}_h, \underline{\mathbf{v}}_h) + \mathrm{b}_h(\underline{\mathbf{v}}_h, \underline{\mathrm{p}}_h) = (\mathbf{f}, \mathbf{curl}_h\, \mathbf{v}_h)_\Omega \qquad \forall \underline{\mathbf{v}}_h \in \underline{\mathbf{X}}_{h,\mathbf{0}}^{k+1}, \tag{8a}$$

$$-\mathrm{b}_h(\underline{\mathbf{u}}_h, \underline{\mathrm{q}}_h) + \mathrm{c}_h(\underline{\mathrm{p}}_h, \underline{\mathrm{q}}_h) = 0 \qquad \forall \underline{\mathrm{q}}_h \in \underline{\mathrm{Y}}_{h,0}^{k+1}. \tag{8b}$$

Some remarks are in order.

Remark 2 (Well-posedness) At the discrete level, $\underline{\mathrm{p}}_h$ is a priori nonzero. The well-posedness of Problem (8) hinges on the following discrete Weber inequality, whose proof will be given in the forthcoming article [2]: For all $(\underline{\mathbf{w}}_h, \underline{\mathrm{r}}_h) \in \underline{\mathbf{X}}_{h,\mathbf{0}}^{k+1} \times \underline{\mathrm{Y}}_{h,0}^{k+1}$ such that

$$-\mathrm{b}_h(\underline{\mathbf{w}}_h, \underline{\mathrm{q}}_h) + \mathrm{c}_h(\underline{\mathrm{r}}_h, \underline{\mathrm{q}}_h) = 0 \qquad \forall \underline{\mathrm{q}}_h \in \underline{\mathrm{Y}}_{h,0}^{k+1}, \tag{9}$$

it holds

$$\|\mathbf{w}_h\|_\Omega^2 \lesssim \mathrm{a}_h(\underline{\mathbf{w}}_h, \underline{\mathbf{w}}_h) + \mathrm{c}_h(\underline{\mathrm{r}}_h, \underline{\mathrm{r}}_h). \tag{10}$$

Note that the commutation property (7) is instrumental to prove (10). Remark, as well, that the discrete solution $(\underline{\mathbf{u}}_h, \underline{\mathrm{p}}_h)$ to Problem (8) satisfies (9). The inequality (10) implies that $|(\underline{\mathbf{w}}_h, \underline{\mathrm{r}}_h)|_{\mathrm{e},h}^2 := \mathrm{a}_h(\underline{\mathbf{w}}_h, \underline{\mathbf{w}}_h) + \mathrm{c}_h(\underline{\mathrm{r}}_h, \underline{\mathrm{r}}_h)$ defines a norm on the subspace of $\underline{\mathbf{X}}_{h,\mathbf{0}}^{k+1} \times \underline{\mathrm{Y}}_{h,0}^{k+1}$ given by (9). As a consequence, the bilinear form of Problem (8), that is

Table 1 Dimensions of the local sets of face unknowns for the potential (first column) and the magnetic field (second column), for both our method (left) and the one of [3] (right)

k	$\dim\left[\mathbb{P}^{k+1}(F)\right]$	$\dim\left[\nabla_\tau\,\mathbb{P}^{k+2}(F)\right]$	$\dim\left[\mathbb{P}^{k+1}(F)\right]$	$\dim\left[\mathbb{P}^{k}(F;\mathbb{R}^2)\oplus\nabla_\tau\,\widetilde{\mathbb{P}}^{k+2}(F)\right]$
0	3	5	3	5
1	6	9	6	10
2	10	14	10	17
3	15	20	15	26

$$\mathscr{A}_h\big((\underline{\mathbf{w}}_h, \underline{\mathrm{r}}_h), (\underline{\mathbf{v}}_h, \underline{\mathrm{q}}_h)\big) := \mathrm{a}_h(\underline{\mathbf{w}}_h, \underline{\mathbf{v}}_h) + \mathrm{b}_h(\underline{\mathbf{v}}_h, \underline{\mathrm{r}}_h) - \mathrm{b}_h(\underline{\mathbf{w}}_h, \underline{\mathrm{q}}_h) + \mathrm{c}_h(\underline{\mathrm{r}}_h, \underline{\mathrm{q}}_h),$$

is coercive on the latter subspace. This is not true if c_h is only a semi-norm on $\underline{\mathrm{Y}}_{h,0}^{k+1}$, as is the case in [3].

Remark 3 (Algebraic aspects) We point out that all the element unknowns can be locally eliminated, resulting in a global system written in terms of face unknowns only. In Table 1 we collect the dimensions, for several values of k, of the local sets of face unknowns for both the potential and the magnetic field, and we provide a comparison with [3].

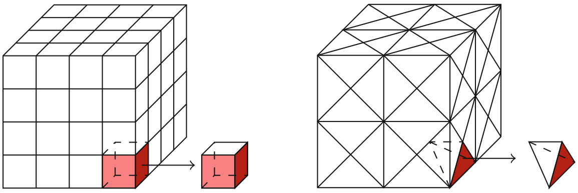
Remark 4 (Convergence rates) For smooth enough solutions, the error in discrete energy-norm $|\cdot|_{\mathrm{e},h}$ is expected to be of order $k+1$, whereas an order $k+2$ is expected for the L^2-error on the magnetic field. Details will be given in [2]. Recall that we are not interested here in the approximaton of $p=0$, but only in that of $\mathbf{u}$.

4 Numerical Experiments

We let Ω be the unit cube, and we consider the following smooth solution:

$$\mathbf{u}(x_1, x_2, x_3) := \begin{pmatrix} \sin(\pi x_2)\sin(\pi x_3) \\ \sin(\pi x_1)\sin(\pi x_3) \\ \sin(\pi x_1)\sin(\pi x_2) \end{pmatrix}. \tag{11}$$

One can easily verify that $\mathbf{u}$ defined by (11) satisfies (1b) and the boundary condition (1c). The expression of the source term $\mathbf{f}$ is inferred from (1a). The numerical experiments are performed on two mesh families, a cubic one and a regular tetrahedral one, as shown on Fig. 1. Element unknowns are locally eliminated, and the resulting (condensed) global linear system is solved using the SparseLU direct solver



(a) Cubic (b) Regular tetrahedral

Fig. 1 Mesh families for the numerical tests

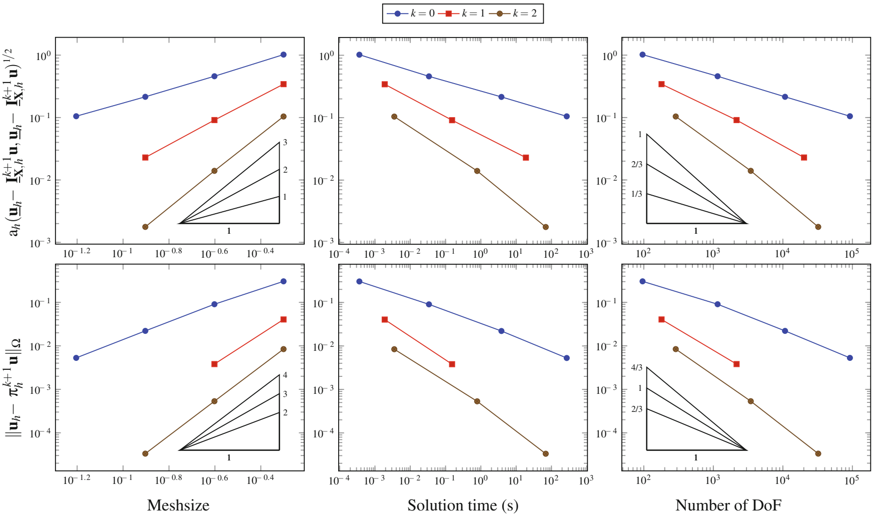


Fig. 2 Errors versus h (left column), solution time (middle column), and number of DoF (right column) on cubic meshes

of the Eigen library, on an Intel Xeon E5-2680 v4 2.4 GHz with 128 Go of RAM. We display on Figs. 2 and 3 the relative errors as functions of, respectively, the meshsize, the solution time in seconds, i.e. the time needed to solve the (condensed) global linear system, and the number of (interface) degrees of freedom (DoF). For both mesh families, the observed convergence orders are, as expected, (i) $k + 1$ for the error $a_h(\underline{\mathbf{u}}_h - \underline{\mathbf{I}}_{\mathbf{X},h}^{k+1}\mathbf{u}, \underline{\mathbf{u}}_h - \underline{\mathbf{I}}_{\mathbf{X},h}^{k+1}\mathbf{u})^{1/2}$, and (ii) $k + 2$ for the error $\|\mathbf{u}_h - \pi_h^{k+1}\mathbf{u}\|_\Omega$. Figures 2 and 3 also clearly exemplify the fact that, whenever the solution is smooth enough (at least locally), if one wants to increase the accuracy, then raising the polynomial degree is computationally much more efficient than refining the mesh.

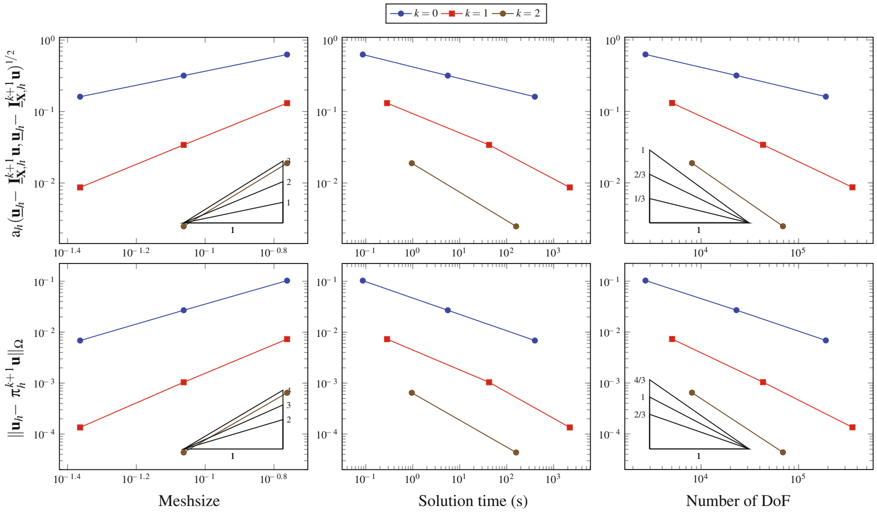


Fig. 3 Errors versus h (left column), solution time (middle column), and number of DoF (right column) on regular tetrahedral meshes

Acknowledgements The authors thank Lorenzo Botti (University of Bergamo) for giving them access to his 3D C++ code SpaFEDTe (https://github.com/SpaFEDTe/spafedte.github.com).

References

1. Assous, F., Ciarlet Jr., P., Labrunie, S.: Mathematical foundations of computational electromagnetism, Applied Mathematical Sciences, vol. 198. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-70842-3
2. Chave, F., Di Pietro, D.A., Lemaire, S.: A discrete Weber inequality on three-dimensional hybrid spaces with application to the HHO approximation of magnetostatics. In preparation
3. Chen, H., Qiu, W., Shi, K., Solano, M.: A superconvergent HDG method for the Maxwell equations. J. Sci. Comput. **70**(3), 1010–1029 (2017). https://doi.org/10.1007/s10915-016-0272-z
4. Cockburn, B., Di Pietro, D.A., Ern, A.: Bridging the hybrid high-order and hybridizable discontinuous galerkin methods. ESAIM: Math. Model. Numer. Anal. **50**(3), 635–650 (2016). https://doi.org/10.1051/m2an/2015051
5. Di Pietro, D.A., Droniou, J.: The Hybrid High-Order method for polytopal meshes. No. 19 in Modeling, Simulation and Applications. Springer International Publishing, New York (2020). https://doi.org/10.1007/978-3-030-37203-3
6. Di Pietro, D.A., Ern, A.: A Hybrid High-Order locking-free method for linear elasticity on general meshes. Comput. Methods Appl. Mech. Eng. **283**, 1–21 (2015). https://doi.org/10.1016/j.cma.2014.09.009
7. Di Pietro, D.A., Ern, A., Lemaire, S.: An arbitrary-order and compact-stencil discretization of diffusion on general meshes based on local reconstruction operators. Comput. Methods Appl. Math. **14**(4), 461–472 (2014). https://doi.org/10.1515/cmam-2014-0018
8. Kikuchi, F.: Mixed formulations for finite element analysis of magnetostatic and electrostatic problems. Japan J. Appl. Math. **6**(2), 209–221 (1989). https://doi.org/10.1007/BF03167879

9. Nédélec, J.C.: Mixed finite elements in $\mathbb{R}^3$. Numer. Math. **35**(3), 315–341 (1980). https://doi.org/10.1007/BF01396415
10. Nguyen, N.C., Peraire, J., Cockburn, B.: Hybridizable discontinuous Galerkin methods for the time-harmonic Maxwell's equations. J. Comput. Phys. **230**(19), 7151–7175 (2011). https://doi.org/10.1016/j.jcp.2011.05.018
11. Perugia, I., Schötzau, D., Monk, P.: Stabilized interior penalty methods for the time-harmonic Maxwell equations. Comput. Methods Appl. Mech. Engrg. **191**(41–42), 4675–4697 (2002). https://doi.org/10.1016/S0045-7825(02)00399-7
12. Beirão da Veiga, L., Brezzi, F., Dassi, F., Marini, L.D., Russo, A.: A family of three-dimensional virtual elements with applications to magnetostatics. SIAM J. Numer. Anal. **56**(5), 2940–2962 (2018). https://doi.org/10.1137/18M1169886

Hyperbolic Conservation Laws with Stochastic Discontinuous Flux Functions

Lukas Brencher and Andrea Barth

Abstract Hyperbolic conservation laws are utilized to describe a variety of real-world applications, which require the consideration of the influence of uncertain parameters on the solution to the problem. To extend these models, one is often interested in including discontinuities in the state space to the flux function of the conservation law. This paper studies the solution of a stochastic nonlinear hyperbolic partial differential equation (PDE), whose flux function contains random spatial discontinuities. The first part of the paper defines the corresponding stochastic adapted entropy solution and required properties for existence and uniqueness are addressed. The second part contains the numerical simulation of the nonlinear hyperbolic problem as well as the estimation of the expectation of the problem via the multilevel Monte Carlo method.

Keywords Stochastic conservation laws · Discontinuous flux function · Stochastic entropy solution · Jump-advection coefficient · Uncertainty quantification · Finite volume method

MSC (2010) 35L65 · 35R05 · 35R60 · 60H15 · 65M08 · 68U20

1 Stochastic Scalar Conservation Laws with Discontinuous Flux Function

In many applications, e.g., two-phase flows in porous media [1], vehicular traffic flows [2] or sedimentation [6], one is interested in including stochastic discontinuities to the state space of the flux function. In this paper, we consider the nonlinear hyperbolic PDE with a stochastic flux function (for $T \in (0, +\infty)$):

L. Brencher (✉) · A. Barth
University of Stuttgart, Allmandring 5b, 70569 Stuttgart, Germany
e-mail: lukas.brencher@mathematik.uni-stuttgart.de

A. Barth
e-mail: andrea.barth@mathematik.uni-stuttgart.de

R. Klöfkorn et al. (eds.), *Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples*, Springer Proceedings in Mathematics & Statistics 323,
https://doi.org/10.1007/978-3-030-43651-3_23

$$u_t + F(\omega, x, u)_x = 0 \qquad \text{on } \Omega \times \mathbb{R} \times (0, T)\ , \tag{1}$$

with $u(\omega, \cdot, 0) = u_0(\cdot)$ on $\mathbb{R}$. Here, the flux function $F : \mathbb{R} \to \mathbb{R}$ is assumed to depend discontinuously on the space variable and to be nonlinear. During the last decades, (deterministic) conservation laws with discontinuous flux functions have received a lot of attention. When the flux function is sufficiently smooth, i.e, $F(x, u)$ is locally Lipschitz in u and globally Lipschitz in x, one can use the Kružkov [13] entropy inequality

$$\begin{aligned} \partial_t |u - m| + \partial_x \left(\operatorname{sgn}(u - m)\left(F(x, u) - F(x, m)\right)\right) \\ + \operatorname{sgn}(u - m)\, \partial_x F(x, m) \le 0 \end{aligned} \tag{2}$$

to find a weak entropy solution to problem (1). However, when F depends discontinuously on the space variable x, the last term of (2) is not well defined. During the last decades, different criteria were proposed for the right solution concept of entropy solutions for discontinuous flux functions.

Towers [17] defined a notion of entropy solutions, which is based on the idea of Klingenberg and Risebro [12], who consider a wave entropy condition to select a weak entropy solution. The theory in [17] only considers flux functions $F(x, u) = a(x) f(u)$, where f has the form $f(u) = u(1 - u)$. This theory was extended in [16] to non-separated variables and can also be stated for multidimensional conservation laws [16].

A different theory, which we will adapt to in this paper, was introduced by Baiti and Jenssen [4] and extended by [3]. The main idea is to rewrite the Kruzkov entropy condition to allow spatial discontinuities in the flux function. This is achieved by considering the solutions of the stationary problem as adapted entropies instead of the usual Kruzkov entropies. Thus, this framework allows the number of discontinuities to be infinite, as no interface condition needs to be imposed.

In the context of (smooth) stochastic conservation laws, Holden and Risebro [10] and Kim [11] introduced the notion of stochastic weak entropy solutions. Mishra and Schwab [15] extended this theory to stochastic systems of conservation laws.

2 Stochastic Adapted Entropy Solutions

In this section, we give an existence and uniqueness result for the solution to problem (1). Therefore, let $(\Omega, \mathscr{A}, \mathbb{P})$ be a complete probability space and consider the nonlinear hyperbolic PDE with a discontinuous random field (for $T \in (0, +\infty)$):

$$\begin{aligned} u_t + (a(\omega, x) f(u))_x &= 0 \qquad \text{on } \Omega \times \mathbb{R} \times (0, T) \\ u(\omega, \cdot, 0) &= u_0(\cdot) \quad \text{on } \mathbb{R}. \end{aligned} \tag{3}$$

Here, the initial condition u_0 satisfies $u_0 \in L^\infty(\mathbb{R})$ and the flux function $f : \mathbb{R} \to \mathbb{R}$ is assumed to be nonlinear, in particular we consider the Burgers' flux. The random field $a : \Omega \times \mathbb{R} \to \mathbb{R}$ is assumed to depend discontinuously on the spatial variable. Throughout this section, we make the following assumptions on the flux function $f : \mathbb{R} \to \mathbb{R}$ and on the discontinuous random field $a : \Omega \times \mathbb{R} \to \mathbb{R}$:

(A.1) For almost all $\omega \in \Omega$, the random field $a(\omega, \cdot)$ is continuous at all points $x \in (\mathbb{R} \setminus \mathcal{N}(\omega))$, where $\mathcal{N}$ is a closed set of measure zero, that might depend on $\omega \in \Omega$.

(A.2) For almost all $\omega \in \Omega$, the random field admits the following estimates

$$a_-(\omega) := \inf_{x\in\mathbb{R}} a(\omega, x) > 0 \qquad \text{and} \qquad a_+(\omega) := \|a(\omega, \cdot)\|_{L^\infty(\mathbb{R})} < +\infty.$$

(F.1) $f(u)$ is continuous.

(F.2) There exists a constant $u_m \in \mathbb{R}$ such that f is a locally Lipschitz one-to-one function from $(-\infty, u_m]$ and $[u_m, \infty)$ to $[0, \infty)$ (or $(-\infty, 0]$) with $f(u_m) = 0$ with common Lipschitz constant L_I for all $u \in I$, where I is any bounded interval in $\mathbb{R}$.

Alternatively, instead of Assumption (F.2), one may assume

(F.2') f is a locally Lipschitz one-to-one function from $\mathbb{R}$ to $\mathbb{R}$ with common Lipschitz constant L_I for all $u \in I$, where I is any bounded interval in $\mathbb{R}$.

If the aforementioned assumptions (F.1)–(F.2') and (A.1)–(A.2) are satisfied, for any constant $\alpha \in [0, \infty)$ (or $(-\infty, 0]$), there exist $\mathbb{P}$-a.s. two steady-state solutions $m_\alpha^\pm$, with $m_\alpha^+ : \mathbb{R} \to [0, \infty)$ and $m_\alpha^- : \mathbb{R} \to (-\infty, 0]$ of (3), such that

$$F(\omega, x, m_\alpha^\pm(x)) := a(\omega, x) f(m_\alpha^\pm(x)) = \alpha \qquad \text{for a.e. } x \in \mathbb{R}. \tag{4}$$

If instead of assumption (F.2') assumption (F.2) is satisfied, we have $m_\alpha^+ = m_\alpha^-$. With these assumptions, we are now able to define the notion of stochastic adapted entropy solutions.

Definition 1 (*Stochastic adapted entropy solutions*) Let $T > 0$ be given. We say that an $L^\infty(\mathbb{R} \times [0, T]) \cap C^0([0, T], L^1_{loc}(\mathbb{R}))$-valued random variable u is a stochastic adapted entropy solution of problem (3) provided that, for $\alpha \in [0, \infty)$ (or $(-\infty, 0]$) and the corresponding two stochastic steady state solutions $m_\alpha^\pm(\omega, x)$ of (3), the following inequality holds in the sense of distributions:

$$\begin{aligned}&\partial_t |u(\omega, x, t) - m_\alpha^\pm(\omega, x)| + \\ &\partial_x \Big[\operatorname{sgn}\big(u(\omega, x, t) - m_\alpha^\pm(\omega, x)\big) \big(F(\omega, x, u(\omega, x, t)) - F(\omega, x, m_\alpha^\pm(\omega, x))\big) \Big] \leq 0.\end{aligned} \tag{5}$$

Theorem 1 (Uniqueness) *Let Assumptions (A.1)–(A.2) together with assumptions (F.1)–(F.2) or (F.1)–(F.2') be satisfied and set* $T > 0$.

Further, let $u, v \in L^\infty(\mathbb{R} \times [0,T]) \cap C^0([0,T], L^1_{loc}(\mathbb{R}))$ *be two stochastic adapted entropy solutions with initial data* $u_0, v_0 \in L^\infty(\mathbb{R})$*. Then, for a.e.* $t \in [0,T]$*, we have*

$$\int_a^b |u(\omega,x,t) - v(\omega,x,t)|dx \le \int_{a-M(\omega)t}^{b+M(\omega)t} |u_0(x) - v_0(x)|dx, \tag{6}$$

for almost all $\omega \in \Omega$*.*

Proof Let $\omega \in \Omega$ be fixed. By hypothesis, the flux function $F(\omega,x,u) = a(\omega,x) f(u)$ satisfies $\mathbb{P}$-a.s. the assumptions for the uniqueness of a solution to deterministic scalar conservation laws with discontinuous flux functions via adapted entropies, see [3, Theorem 4.1].

Theorem 2 (Existence of a stochastic adapted entropy solution) *Let Assumptions (A.1)–(A.2) together with assumptions (F.1)–(F.2) or (F.1)–(F.2') be satisfied for almost all* $\omega \in \Omega$ *and let* $u_0 \in L^\infty(\mathbb{R})$ *with* $u_0(x) \ge 0$*. Then for* $T > 0$ *and almost all* $\omega \in \Omega$ *there exists a stochastic adapted entropy solution to problem (3).*

Proof Let $\omega \in \Omega$ be fixed. The existence of a stochastic adapted entropy solution is then proved identically as for the deterministic case [7, Theorem 3.1].

3 Discontinuous Random Field

The stochastic coefficient in problem (3) should model heterogeneities and/or fractures in a medium. Therefore, we utilize and adapt the random jump coefficient a, which was introduced (in a slightly modified way) in [5] for an elliptic diffusion problem. It consists of a (spatial) Gaussian random field with additive discontinuities on random submanifolds of a domain $\mathscr{D} \subset \mathbb{R}$.

Definition 2 (*Jump-advection coefficient*) We consider a random field with the additive form

$$a(\omega,x) := \overline{a}(x) + \phi(W_{\mathscr{D}}(\omega,x)) + P(\omega,x),$$

where $\overline{a} \in C(\mathbb{R}; \mathbb{R}_{\ge 0})$ is a deterministic, uniformly bounded mean function and $\phi \in C^1(\mathbb{R}; \mathbb{R}_{>0})$. For a (zero-mean) Gaussian random field $W \in L^2(\Omega; L^2(\mathbb{R}))$ associated to a non-negative, symmetric trace class (covariance) operator $Q : L^2(\mathbb{R}) \to L^2(\mathbb{R})$, the random field $W_{\mathscr{D}} \in L^2(\Omega; L^2(\mathbb{R}))$ is defined as

$$W_{\mathscr{D}}(\omega,x) = \begin{cases} W(\omega,x), & x \in \mathscr{D} \\ \min(W(\omega,x), \sup\limits_{x\in\mathscr{D}} W(\omega,x)), & x \in \mathbb{R} \setminus \mathscr{D}. \end{cases}$$

For the random discontinuities we define $\mathscr{T} : \Omega \to \mathscr{B}(\mathscr{D}),\ \omega \mapsto \{\mathscr{T}_1, \ldots, \mathscr{T}_\tau\}$ as a random partition of $\mathscr{D}$, i.e., the $\mathscr{T}_i$ are disjoint open subsets of $\mathscr{D}$ with $\overline{\mathscr{D}} = \bigcup_{i=1}^{\tau} \overline{\mathscr{T}_i}$.

The number of elements in $\mathscr{T}$ is a random variable $\tau : \Omega \to \mathbb{N}$ on $(\Omega, \mathscr{A}, \mathbb{P})$. For $\mathscr{D}_l$ and $\mathscr{D}_r$ being the left and right boundary of $\mathscr{D}$, respectively, we define $\mathscr{T}_0 := (-\infty, \mathscr{D}_l)$ and $\mathscr{T}_{\tau+1} := (\mathscr{D}_r, +\infty)$. The jump heights of the random field are then given by a sequence $(P_i, i \in \mathbb{N}_0)$ of random variables on $(\Omega, \mathscr{A}, \mathbb{P})$ with arbitrary non-negative distribution(s), which is independent of τ (but not necessarily i.i.d.). Further, we have

$$P : \Omega \times \mathscr{D} \to \mathbb{R}_{\geq 0}, \quad (\omega, x) \mapsto \sum_{i=0}^{\tau+1} \mathbf{1}_{\mathscr{T}_i}(x) P_i(\omega).$$

4 Numerical Experiments

In this section, we present numerical simulations of the stochastic Burgers' equation, i.e., problem (3) with $f(u) = \frac{u^2}{2}$. Therefore, we only consider a subset $\mathscr{D} \subset \mathbb{R}$. Further, we introduce an approximation of the random field a and discretize the problem (3) by a Godunov Finite Volume method.

4.1 Approximation of the Random Field

For the continuous part of the random field introduced in Sect. 3, we set $\bar{a} \equiv 0$ and $\phi(w) = \exp(w)$, where we assume that the Gaussian field W is characterized by the *Matérn covariance operator* $Q_M : L^2(\mathscr{D}) \to L^2(\mathscr{D})$:

$$[Q_M \varphi](y) := \int_{\mathscr{D}} \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu} \frac{|x-y|}{\rho} \right)^{\nu} K_\nu \left(\sqrt{2\nu} \frac{|x-y|}{\rho} \right) \varphi(x) dx, \qquad (7)$$

for $\varphi \in L^2(\mathscr{D})$, where $\nu > 0$ denotes the smoothness parameter, $\sigma^2 > 0$ is the variance and with correlation length $\rho > 0$. Throughout our experiments, we set $\sigma^2 = 0.1$ and $\rho = 0.1$. Further, Γ is the Gamma function and K_ν is the modified Bessel function of second kind with ν degrees of freedom. The Gaussian field W admits the Karhunen-Loève expansion and thus is approximated via the truncated decomposition

$$W(\omega, x) \approx \tilde{W}(\omega, x) := \sum_{i=1}^{K} \sqrt{\eta_i} e_i(x) w(\omega) \quad x \in \mathscr{D},\ K \in \mathbb{N},\ \omega \in \Omega, \qquad (8)$$

where $((\eta_i, e_i), i \in \mathbb{N})$ denotes the spectral basis of Q_M. We approximate the eigenbasis of Q_M via Nyström's method [18].

The partition $\mathscr{T}$ is generated by $\tau \sim \text{Poi}(5) + 2$ elements, resulting almost surely in at least one discontinuity of the random field. Let the jump positions $(\chi_i, i \in \mathbb{N})$

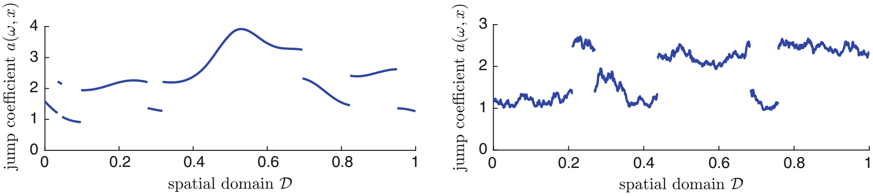


Fig. 1 Two realizations of the random field $a(\omega, x)$ for varying smoothness parameter of the covariance operator. Left: $\nu = \infty$, right: $\nu = \frac{1}{2}$

be an i.i.d. sequence of $\mathcal{U}(\mathcal{D})$-random variables, which are independent of τ. The corresponding jump heights are given by

$$P_i \sim \mathcal{U}\left(\left[\frac{3}{4} + (-1)^i \frac{1}{2}, \frac{5}{4} + (-1)^i \frac{1}{2}\right]\right) = \begin{cases} \mathcal{U}\left(\left[\frac{1}{4}, \frac{3}{4}\right]\right) & i \text{ odd}, \\ \mathcal{U}\left(\left[\frac{5}{4}, \frac{7}{4}\right]\right) & i \text{ even}. \end{cases} \tag{9}$$

Figure 1 shows two different realizations of the random field $a(\omega, x)$ for varying smoothness parameter ν.

4.2 Finite Volume Discretization

For the pathwise spatial discretization, we employ a classical Finite Volume discretization in conservative form, which is given by

$$u_j^{m+1} := u_j^m - \frac{\Delta t}{\Delta x}\left(g_{j+\frac{1}{2}} - g_{j-\frac{1}{2}}\right), \qquad u_j^0 := \frac{1}{\Delta x}\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} u_0(x)dx, \tag{10}$$

for $j \in \mathbb{Z}$. In the presented numerical experiments, the Godunov flux is used. For an extensive discussion on Finite Volume methods and conservation laws, we refer to the books of Dafermos and LeVeque [8, 14]. For the temporal discretization, we employ the Backward Euler time-stepping scheme with an equidistant time step size $\Delta t > 0$.

4.3 Multilevel Monte Carlo Estimation

We aim to approximate the stochastic moments of the solution via the multilevel Monte Carlo method. Therefore, let $(u_l, l \in \mathbb{N})$ be a sequence of discretizations converging to the exact solution. For a given discretization level $l \in \mathbb{N}$ we denote by E_{M_l} the Monte Carlo estimator with $M_l \in \mathbb{N}$ samples. The multilevel Monte Carlo estimator of $\mathbb{E}(u_L)$ is then defined as

$$E^L(u_L) := E_{M_0}(u_0) + \sum_{l=1}^{L} E_{M_l}(u_l - u_{l-1}). \tag{11}$$

Here, $M_0 > \cdots > M_L$ are the number of computed samples on each level. For a detailed discussion on the multilevel Monte Carlo method, we refer to [9].

4.4 Numerical Experiments

For the numerical simulations, we consider the domain $\mathscr{D} = (0, 1)$ with $T = 1$, the initial condition $u_0 = 0.3 \sin(\pi x)$ and apply the numerical methods describe in the previous subsections to problem (3) with the Burgers' flux. The level-dependent spatial step sizes in the multilevel Monte Carlo computation are set to $\Delta x_l = 2^{-l} \Delta x_0$.

In Fig. 2 we illustrate the solution of two realizations together with the underlying random field. The same realizations of the random field are also shown in Fig. 1. Here, we see the influence of the discontinuities in the random field to the creation of shocks and we note that regularity of the random coefficient, i.e., the smoothness of the covariance operator, directly influences the regularity of the solution.

In Fig. 3 we show the multilevel Monte Carlo estimator $E^L(u_L)$ for $L = 8$. In contrast to the realizations presented in Fig. 2, the smoothness of the covariance operator does not seem to have a major impact on the estimated solution.

In Fig. 4, we present results on the convergence behaviour of the multilevel Monte Carlo estimation for a smooth and rough covariance operator, i.e., $\nu = \infty$ and $\nu = 0.5$, respectively. Therefore, we consider the error $\mathscr{E} := \mathbb{E}(\|\mathbb{E}(u) - E^L(u_L)\|_{L^1(\mathscr{D})})$, where we approximate $\mathbb{E}(u)$ by a finer reference solution $E^{L_{\text{ref}}}(u_{L_{\text{ref}}})$ and the outer expectation by a standard Monte Carlo estimate. The observed convergence rate is in both cases approximately 0.8.

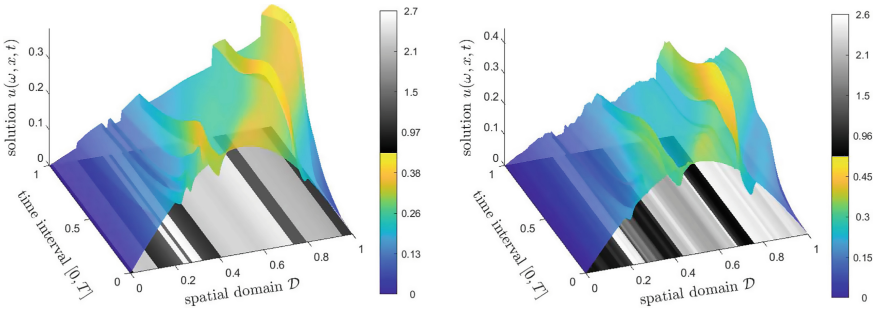


Fig. 2 Two realizations of the solution to problem (3) with the Burgers' flux. Left: solution with the underlying smooth random field ($\nu = \infty$). Right: solution with the underlying rough random field ($\nu = 0.5$)

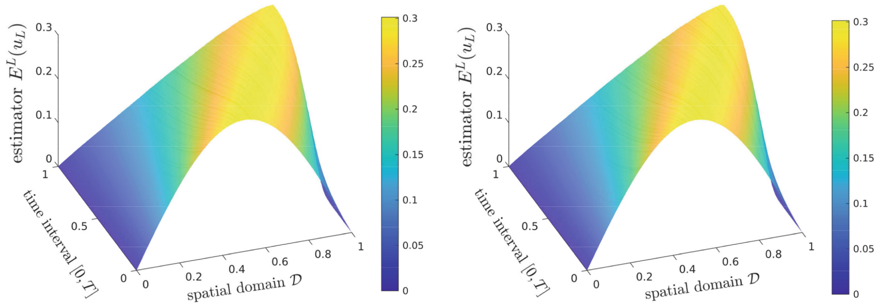


Fig. 3 Multilevel Monte Carlo estimate of the solution to problem (3) with the Burgers' flux. Left: Estimate for smooth covariance operator ($\nu = \infty$). Right: Estimate for rough covariance operator ($\nu = 0.5$)

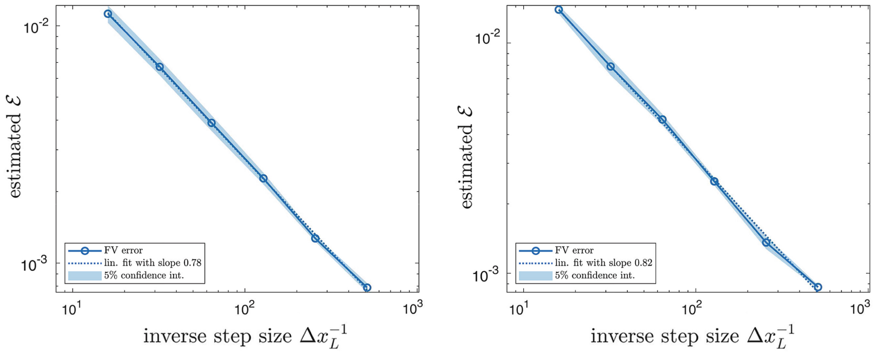


Fig. 4 Convergence of the multilevel Monte Carlo method for different number of levels. Left: Convergence for smooth covariance operator ($\nu = \infty$). Right: Convergence for rough covariance operator ($\nu = 0.5$)

Acknowledgements Funded by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy—EXC 2075—390740016.

References

1. Andreianov, B., Brenner, K., Cancès, C.: Approximating the vanishing capillarity limit of two-phase flow in multi-dimensional heterogeneous porous medium. ZAMM-J. Appl. Math. Mech./Z. für Angew. Math. Und Mech. **94**(7–8), 655–667 (2014)
2. Andreianov, B., Goatin, P., Seguin, N.: Finite volume schemes for locally constrained conservation laws. Numer. Math. **115**(4), 609–645 (2010)
3. Audusse, E., Perthame, B.: Uniqueness for scalar conservation laws with discontinuous flux via adapted entropies. Proc. R. Soc. Edinb. Sect. A: Math. **135**(2), 253–265 (2005)
4. Baiti, P., Jenssen, H.K.: Well-posedness for a class of 2×2 conservation laws with l^∞ data. J. Differ. Equ. **140**(1), 161–185 (1997)

5. Barth, A., Stein, A.: A study of elliptic partial differential equations with jump diffusion coefficients. SIAM/ASA J. Uncertain. Quantif. **6**(4), 1707–1743 (2018)
6. Bürger, R., Karlsen, K.H., Risebro, N.H., Towers, J.D.: Well-posedness in Bv t and convergence of a difference scheme for continuous sedimentation in ideal clarifier-thickener units. Numer. Math. **97**(1), 25–65 (2004)
7. Chen, G.Q., Even, N., Klingenberg, C.: Hyperbolic conservation laws with discontinuous fluxes and hydrodynamic limit for particle systems. J. Differ. Equ. **245**(11), 3095–3126 (2008)
8. Dafermos, C.M., Mathématicien, G., Mathematician G (2000) Hyperbolic conservation laws in continuum physics, vol 325. Springer
9. Giles, M.B.: Multilevel monte carlo methods. Acta Numer. **24**, 259–328 (2015)
10. Holden, H., Risebro, N.H.: Conservation laws with a random source. Appl. Math. Optim. **36**(2), 229–241 (1997)
11. Kim, J.U.: On a stochastic scalar conservation law. Indiana Univ. Math. J. pp. 227–256 (2003)
12. Klingenberg, C., Risebro, N.H.: Convex conservation laws with discontinuous coefficients. Existence, uniqueness and asymptotic behavior. Commun. Part. Differ. Equ. **20**(11–12), 1959–1990 (1995)
13. Kružkov, S.N.: First order quasilinear equations in several independent variables. Math. USSR-Sb. **10**(2), 217 (1970)
14. LeVeque, R.J., et al.: Finite Volume Methods for Hyperbolic Problems, vol. 31. Cambridge university press (2002)
15. Mishra, S., Schwab, C., Šukys, J.: Multi-level monte carlo finite volume methods for nonlinear systems of conservation laws in multi-dimensions. J. Comput. Phys. **231**(8), 3365–3388 (2012)
16. Panov, E.Y.: Existence and strong pre-compactness properties for entropy solutions of a first-order quasilinear equation with discontinuous flux. Arch. Rat. Mech. Anal. **195**(2), 643–673 (2010)
17. Towers, J.D.: Convergence of a difference scheme for conservation laws with a discontinuous flux. SIAM J. Numer. Anal. **38**(2), 681–698 (2000)
18. Williams, C.K., Rasmussen, C.E.: Gaussian Processes for Machine Learning, vol 2. MIT press Cambridge, MA (2006)

Convergence of a Finite-Volume Scheme for a Heat Equation with a Multiplicative Stochastic Force

Caroline Bauzet and Flore Nabet

Abstract We present here the discretization by a finite-volume scheme of a heat equation perturbed by a multiplicative noise of Itô type and under homogeneous Neumann boundary conditions. The idea is to adapt well-known methods in the deterministic case for the approximation of parabolic problems to our stochastic PDE. In this paper, we try to highlight difficulties brought by the stochastic perturbation in the adaptation of these deterministic tools.

Keywords Stochastic heat equation · Itô integral · Multiplicative noise · Itô formula · Predictable process · Finite volume method

MSC (2010) 35K05 · 60H15 · 65M08

1 Introduction

In this section, we present the stochastic heat equation we are studying. After deriving assumptions on the data, we explain the goal of the paper and give the definition of weak solution we are looking for. More precisely, we are interested in the following stochastic heat equation set in $(0, T) \times \Omega \times \Theta$,

$$\partial_t \left(u(t, x, \omega) - \int_0^t \lambda u(s, x, \omega) dW(s) \right) - \Delta u(t, x, \omega) = 0, \tag{1}$$

where Ω is an open bounded polygonal subset of $\mathbb{R}^2$, $T > 0$, $W = \{W_t, \mathscr{F}_t; 0 \leqslant t \leqslant T\}$ is a standard adapted one-dimensional continuous Brownian motion defined on

C. Bauzet (✉)
Aix Marseille University, CNRS, Centrale Marseille, LMA, 13013 Marseille, France
e-mail: caroline.bauzet@univ-amu.fr

F. Nabet
CMAP, Ecole Polytechnique, CNRS, I.P. Paris, 91128 Palaiseau, France
e-mail: flore.nabet@polytechnique.edu

R. Klöfkorn et al. (eds.), *Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples*, Springer Proceedings in Mathematics & Statistics 323,
https://doi.org/10.1007/978-3-030-43651-3_24

the classical Wiener space $(\Theta, \mathscr{F}, \mathbb{P})$ and $\lambda \in \mathbb{R}$. An initial condition is given by a deterministic function $u_0 \in L^2(\Omega)$:

$$u(0, x, \omega) = u_0(x),\ x \in \Omega,\ \omega \in \Theta, \tag{2}$$

and we consider homogeneous Neumann boundary condition:

$$\nabla u(t, x, \omega) \cdot \mathbf{n}(x) = 0,\ t \in (0, T),\ x \in \partial\Omega,\ \omega \in \Theta, \tag{3}$$

where $\mathbf{n}$ denotes the unit vector to $\partial\Omega$, outward to Ω. In order to make the lecture more fluent, we omit in the sequel the random variable ω. Note that the present study can be easily adapted to the case where Ω is a subset of $\mathbb{R}^3$ but for the sake of readability, we restrict the presentation to the 2-dimensional case.

In this paper, the stochastic integral $\int_0^{\cdot} \lambda u dW$ is understood in the sense of Itô, so that due to its non-anticipative construction with simple processes, we must consider an explicit time-discretization of this object. Note that the unknown u appears in the stochastic integral, thus the noise is said to be multiplicative, otherwise it is additive.

The numerical analysis of heat equation (and generally second-order parabolic equations) under stochastic perturbation, random source term or random coefficients has been the subject of several studies by the way of finite-element approximations (see [7] for a thorough presentation of the state of the art on this subject). The aim of the present paper is to expose tools of stochastic calculus used to adapt known methods in the deterministic case to get the convergence of a suitable finite-volume scheme for the approximation of problem (1)–(3). This work stands for an introductive study in order to apprehend more complex problems such as the finite-volume approximation of stochastic nonlinear degenerate parabolic equations, having in mind the deterministic case treated by [6].

In what follows, we will show the convergence of a suitable numerical scheme through a stochastic process u, weak solution of (1)–(3) in the following sense:

Definition 1 A predictable process u with values in $L^2(\Omega)$ is a weak solution of (1)–(3) if $u \in L^2\left((0, T) \times \Theta; H^1(\Omega)\right) \cap L^\infty\left((0, T); L^2(\Omega \times \Theta)\right)$ and if it satisfies $\mathbb{P}$-a.s in Θ and for any $\psi \in \mathscr{A}_T = \{\varphi \in C^\infty\left(\mathbb{R} \times \mathbb{R}^2\right) : \varphi(T, .) = 0\}$ the variational formulation

$$\begin{aligned}&\int_0^T \int_\Omega u(t, x)\partial_t\psi(t, x)dxdt - \int_0^T \int_\Omega \nabla u(t, x) \cdot \nabla\psi(t, x)dxdt + \int_\Omega u_0(x)\psi(0, x)dx \\ &= \int_0^T \int_\Omega \int_0^t \lambda u(s, x)dW(s)\partial_t\psi(t, x)dxdt. \end{aligned} \tag{4}$$

Remark 1 The predictability property of u with values in $L^2(\Omega)$ is a condition of measurability of the solution u with respect to the filtration $\mathscr{F} = (\mathscr{F}_t)_{0\leqslant t\leqslant T}$, which represents the history of the Brownian motion up to time T. It is required since we consider a multiplicative noise. More precisely, it means that u belongs to $L^2\left((0, T) \times \Theta, \mathscr{P}_T, dt \otimes \mathbb{P}; L^2(\Omega)\right)$ where $\mathscr{P}_T$ denotes the predictable σ-field generated by (see [8, p. 27])

$$\{X : (0,T)\times\Theta \to \mathbb{R} : \ X \text{ is left-continuous and } \forall t \in [0,T],\ X_t \text{ is } \mathscr{F}_t\text{-measurable}\}.$$

We denote by $\mathscr{N}_W^2(0,T;L^2(\Omega))$ the space of predictable processes with values in $L^2(\Omega)$. Endowed with the norm $||X||^2 = \int_0^T\int_\Theta ||X||^2_{L^2(\Omega)}d\mathbb{P}dt$, it is a Hilbert space.

Remark 2 An application of Itô derivation formula (see [3, Theorem 4.17, p. 105]) allows us to remark that the right-hand side of (4) can also be written in the following manner

$$\int_0^T\int_\Omega\int_0^t \lambda u(s,x)dW(s)\partial_t\psi(t,x)dxdt = -\int_\Omega\int_0^T \lambda u(t,x)\psi(t,x)dW(t)dx.$$

2 Meshes, Scheme and Discrete Norms

We will use a classical two-point flux approximation scheme with an admissible mesh as in [5]. Firstly, we remind for convenience this definition adapted to our subset Ω of $\mathbb{R}^2$ and give some notations. Secondly, we present the finite-volume scheme used to approximate the weak solution u of (1)–(3). Thirdly, we introduce discrete $L^2(\Omega)$-norm and $H^1(\Omega)$-seminorm used for the stability results exposed in the next section.

An admissible mesh $\mathscr{T}$ is given by a family of disjoint open polygonal subsets of Ω, called "control volumes" and denoted by K such that:

- $\overline{\Omega} = \cup_{K\in\mathscr{T}}\overline{K}$;
- if $K, L \in \mathscr{T}$, $K \neq L$, then $\mathring{K}\cap\mathring{L} = \emptyset$;
- if $K, L \in \mathscr{T}$, $K \neq L$, either the 1-dimensional Lebesgue measure of $\overline{K}\cap\overline{L}$ is 0 or $\overline{K}\cap\overline{L}$ is the edge σ of the mesh separating the control volumes K and L;
- at each $K \in \mathscr{T}$, we associate a point $x_K \in K$, called the center of K, such that if K, L are two neighbouring control volumes, the edge $\sigma = K|L$ which separates K and L is orthogonal to the straight line going through x_K and x_L.

Once an admissible finite-volume mesh $\mathscr{T}$ of Ω is fixed, we will use in the sequel the following notations.

Notations

- $E[.]$ denotes the expectation, *i.e* the integral over Θ with respect to the probability measure $\mathbb{P}$.
- $\mathscr{E}$ is the set of the edges of the mesh $\mathscr{T}$ and $\mathscr{E}_{int} = \{\sigma \in \mathscr{E} : \sigma \not\subset \partial\Omega\}$ the set of interior edges.
- For any $K \in \mathscr{T}$, $\mathscr{E}_K$ is the set of the edges of the control volume K and m_K the Lebesgue measure of K.
- For any $\sigma = K|L$, m_σ is the length of σ and $d_{K|L}$ the distance between the centers x_k and x_L.
- h =size($\mathscr{T}$) = sup{diam(K), $K \in \mathscr{T}$}, the mesh size.

In order to compute an approximation of u on $[0, T]$, we take $N \in \mathbb{N}^*$ and consider a fixed time step $\Delta t = \frac{T}{N} \in \mathbb{R}_+^*$. We first define the set $\{u_K^0, K \in \mathscr{T}\}$ by the discretization of the initial condition using its mean value over the control volume K,

$$u_K^0 = \frac{1}{m_K} \int_K u_0(x)dx. \tag{5}$$

The equations satisfied by the discrete unknowns denoted by $u_K^n, n \in \{0, \dots, N-1\}$, $K \in \mathscr{T}$ are given by the following explicit scheme

$$\frac{m_K}{\Delta t}(u_K^{n+1} - u_K^n) + \sum_{\sigma \in \mathscr{E}_K \cap \mathscr{E}_{int}} \frac{m_\sigma}{d_{K|L}}(u_K^n - u_L^n) = \frac{m_K}{\Delta t} \lambda u_K^n (W^{n+1} - W^n), \tag{6}$$

where $W^n = W(n\Delta t), \forall n \in \{0, \dots, N\}$. Note that since the Brownian motion W is standard, thus $W(0) = 0$. The random dependence of the discrete unknowns u_K^n $n \in \{1, \dots, N\}$, comes from the randomness of the increments $W^{n+1} - W^n$, again for convenience, we omit the random variable ω and write u_K^n instead of $u_K^n(\omega)$. We define the piecewise constant approximate solution $(u_{\mathscr{T}}^{\Delta t})$ on $(0, T) \times \Omega \times \Theta$ from the discrete unknowns u_K^n by

$$u_{\mathscr{T}}^{\Delta t}(t, x, \omega) = u_{\mathscr{T}}^n(x, \omega) = u_K^n(\omega) = u_K^n, \quad t \in [n\Delta t, (n+1)\Delta t),\ x \in K,\ \omega \in \Theta, \tag{7}$$

where $(u_{\mathscr{T}}^n)_{\mathscr{T}}$ defined on $\Omega \times \Theta$ is the sequence of the approximate solution at time $t^n = n\Delta t$ for $n \in \{0, ..., N\}$.

Remark 3 Let us mention that using properties of the Brownian motion W, for all $K \in \mathscr{T}$ and all $n \in \{0, \dots, N\}$, u_K^n is $\mathscr{F}_{n\Delta t}$-measurable. Thus, $u_{\mathscr{T}}^{\Delta t}$ is predictable with values in $L^2(\Omega)$ as an elementary process adapted to the filtration $(\mathscr{F}_t)_{0 \leqslant t \leqslant T}$.

We then define for any $n \in \{0, .., N\}$ the following discrete $L^2(\Omega)$-norm $||.||_{L^2(\Omega)}$ and $H^1(\Omega)$ seminorm $|.|_{1,\mathscr{T}}^2$ for the approximate sequence $(u_{\mathscr{T}}^n)_{\mathscr{T}}$, $\mathbb{P}$-a.s in Θ

$$||u_{\mathscr{T}}^n||_{L^2(\Omega)}^2 = \sum_{K \in \mathscr{T}} m_K |u_K^n|^2 \text{ and } |u_{\mathscr{T}}^n|_{1,\mathscr{T}}^2 = \sum_{\sigma = K|L \in \mathscr{E}_{int}} \frac{m_\sigma}{d_{K|L}} |u_K^n - u_L^n|^2.$$

3 Convergence of the Scheme

In this section, we propose a study of the approximate sequence $(u_{\mathscr{T}}^{\Delta t})$. After the derivation of boundedness estimates for $(u_{\mathscr{T}}^{\Delta t})$ independent of the discretization parameters Δt and h, we propose to show the convergence of $(u_{\mathscr{T}}^{\Delta t})$ towards a weak solution u of (1)–(3) in the sense of Definition 1.

Proposition 1 *Let $T > 0$, $\mathscr{T}$ be an admissible mesh, $N \in \mathbb{N}^*$ and $\Delta t = \frac{T}{N} \in \mathbb{R}_+^*$. Assume that the condition*

$$\frac{\Delta t}{m_K}\sum_{\sigma\in\mathscr{E}_K\cap\mathscr{E}_{int}}\frac{m_\sigma}{d_{K|L}}\leqslant\frac{1}{2},\ \forall K\in\mathscr{T} \tag{8}$$

is satisfied. Then there exists a constant $C>0$ only depending on T,Ω,λ and u_0 such that

$$\sup_{n\leqslant N}E\left[||u^n_{\mathscr{T}}||^2_{L^2(\Omega)}\right]+\sum_{n=0}^{N-1}\Delta t E\left[|u^n_{\mathscr{T}}|^2_{1,\mathscr{T}}\right]\leqslant C.$$

Proof We will principally use here properties of the Brownian motion W for the control of its discrete increments $W^{n+1}-W^n$ (see [3, p. 87]). For any $n\in\{0,\dots,N-1\}$, note that since $E[W^{n+1}-W^n]=0$, one gets that for any $\mathscr{F}_{n\Delta t}$-measurable random variable $X:\Theta\to\mathbb{R}$, $E[(W^{n+1}-W^n)X]=E[W^{n+1}-W^n]E[X]=0$. Thus, by multiplying (6) by u^n_K and taking the expectation, since $a(b-a)=\frac{1}{2}(b^2-a^2-(a-b)^2)$ for any $a,b\in\mathbb{R}$, one gets

$$\frac{m_K}{2\Delta t}E\left[|u^{n+1}_K|^2-|u^n_K|^2-|u^{n+1}_K-u^n_K|^2\right]+\sum_{\sigma\in\mathscr{E}_K\cap\mathscr{E}_{int}}\frac{m_\sigma}{d_{K|L}}E\left[(u^n_K-u^n_L)u^n_K\right]=0.$$

Moreover thanks to (6), by noting that $E[(W^{n+1}-W^n)^2]=\Delta t$, one has

$$E\left[|u^{n+1}_K-u^n_K|^2\right]=\Delta t E\left[|\lambda u^n_K|^2\right]+E\left[\left(\frac{\Delta t}{m_K}\sum_{\sigma\in\mathscr{E}_K\cap\mathscr{E}_{int}}\frac{m_\sigma}{d_{K|L}}(u^n_K-u^n_L)\right)^2\right],$$

and one arrives at

$$\begin{aligned}&\frac{m_K}{2}E\left[|u^{n+1}_K|^2-|u^n_K|^2\right]+\Delta t\sum_{\sigma\in\mathscr{E}_K\cap\mathscr{E}_{int}}\frac{m_\sigma}{d_{K|L}}E\left[(u^n_K-u^n_L)u^n_K\right]\\&\leqslant\Delta t\frac{m_K}{2}\lambda^2E\left[|u^n_K|^2\right]+\frac{\Delta t}{2}\left(\frac{\Delta t}{m_K}\sum_{\sigma\in\mathscr{E}_K\cap\mathscr{E}_{int}}\frac{m_\sigma}{d_{K|L}}\right)\left(\sum_{\sigma\in\mathscr{E}_K\cap\mathscr{E}_{int}}\frac{m_\sigma}{d_{K|L}}E\left[|u^n_K-u^n_L|^2\right]\right).\end{aligned}$$

Summing over $K\in\mathscr{T}$ and using the condition (8), thanks to classical reorderings of the summations, we obtain

$$\begin{aligned}&\sum_{K\in\mathscr{T}}m_KE\left[|u^{n+1}_K|^2\right]+2\Delta t\sum_{\sigma=K|L\in\mathscr{E}_{int}}\frac{m_\sigma}{d_{K|L}}E\left[|u^n_K-u^n_L|^2\right]\\&\qquad\leqslant(1+\Delta t\lambda^2)\sum_{K\in\mathscr{T}}m_KE\left[|u^n_K|^2\right]+\frac{1}{2}\Delta t\sum_{K\in\mathscr{T}}\sum_{\sigma\in\mathscr{E}_K\cap\mathscr{E}_{int}}\frac{m_\sigma}{d_{K|L}}E\left[|u^n_K-u^n_L|^2\right]\end{aligned}$$

which leads to

$$E\left[\|u^{n+1}_{\mathscr{T}}\|^2_{L^2(\Omega)}\right]+\Delta t E\left[|u^n_{\mathscr{T}}|^2_{1,\mathscr{T}}\right]\leqslant(1+\Delta t\lambda^2)E\left[\|u^n_{\mathscr{T}}\|^2_{L^2(\Omega)}\right]. \tag{9}$$

Summing (9) over $n \in \{0, \cdots, m\}$, the discrete Gronwall lemma gives the expected $L^\infty(0, T; L^2(\Omega \times \Theta))$ bound,

$$E\left[\|u^m_{\mathcal{T}}\|^2_{L^2(\Omega)}\right] \leqslant (1 + \Delta t\lambda^2)e^{\lambda^2 T}||u_0||^2_{L^2(\Omega)}, \quad \forall m \in 0, \ldots, N.$$

Summing (9) over $n \in \{0, \ldots, N-1\}$ we obtain the $L^2\left((0, T) \times \Theta; H^1(\Omega)\right)$ bound,

$$\sum_{n=0}^{N-1} \Delta t E\left[|u^n_{\mathcal{T}}|^2_{1,\mathcal{T}}\right] \leqslant \left(1 + T\lambda^2(1 + \Delta t\lambda^2)e^{\lambda^2 T}\right)||u_0||^2_{L^2(\Omega)}.$$

Remark 4 Let us mention that the stochastic perturbation does not impact the condition (8) on the time and space discretization parameter to get a stability result on the finite-volume approximation $(u^{\Delta t}_{\mathcal{T}})$. Indeed, the condition is the same as in the deterministic case (which corresponds to $\lambda = 0$).

Remark 5 Note that Proposition 1 also holds for a more general stochastic noise taking the form $\int_0^\cdot g(u)dW$ with $g : \mathbb{R} \to \mathbb{R}$ Lipschitz-continuous. It also implies boundedness of the sequence $(g(u^{\Delta t}_{\mathcal{T}}))$ and so the existence of a weak limit g_u in $\mathcal{N}^2_W((0, T); L^2(\Omega))$ for a subsequence of $(g(u^{\Delta t}_{\mathcal{T}}))$. When g is not a linear function, the convergence result stated in Theorem 1 below requires a compactness tool compatible with the random variable in order to affirm that $g_u = g(u)$. This extension will be carried out in a future work.

Theorem 1 *For $m \in \mathbb{N}$, let $\mathcal{T}_m$ be an admissible mesh, $N_m \in \mathbb{N}^*$ and $\Delta t_m = \frac{T}{N_m}$ satisfying the condition* (8)*. Let $(u^{\Delta t_m}_{\mathcal{T}_m})$ be given by* (5)*–*(7) *with $\mathcal{T} = \mathcal{T}_m$ and $N = N_m$. Then there exists a subsequence of $(u^{\Delta t_m}_{\mathcal{T}_m})$, still denoted $(u^{\Delta t_m}_{\mathcal{T}_m})$, which converges weakly in $L^2((0, T) \times \Omega \times \Theta)$ towards a weak solution u of* (1)*–*(3) *in the sense of Definition 1.*

Proof We will only give here the idea of the proof to handle the stochastic term and refer to [5, Proof of Theorem 18.1 p. 858] for the deterministic contributions. Let $m \in \mathbb{N}$, A be a $\mathbb{P}$-measurable set and $\psi \in C^\infty(\mathbb{R} \times \mathbb{R}^2)$ such that $\psi(T, .) = 0$ and $\nabla\psi \cdot \mathbf{n} = 0$ on $(0, T) \times \partial\Omega$. Since Ω is a polygonal subset of $\mathbb{R}^2$, the set of such functions ψ is dense in the set $\mathcal{A}_T$ for the $L^2(0, T; H^1(\Omega))$-norm (see [4]). For the sake of simplicity we shall use the notations $\mathcal{T} = \mathcal{T}_m$, $h = size(\mathcal{T}_m)$ and $\Delta t = \Delta t_m$. We define the piecewise constant in space function $\psi_{\mathcal{T}}$ on $(0, T) \times \Omega$ by

$$\psi_{\mathcal{T}}(t, x) = \frac{1}{m_{B_K}} \int_{B_K} \psi(t, y)dy, \quad \forall x \in K, t \in (0, T),$$

where $B_K \subset K$ is a ball centered at x_K and m_{B_K} denotes its Lebesgue measure. We multiply (6) by $\Delta t 1_A \psi_{\mathcal{T}}(n\Delta t, x_K)$, sum the result over $n \in \{0, \ldots, N-1\}$ and $K \in \mathcal{T}$, to get after taking the expectation: $T_{1,m} + T_{2,m} = T_{3,m}$, where

$$T_{1,m} = E\left[1_A \sum_{n=0}^{N-1} \sum_{K\in\mathscr{T}} m_K(u_K^{n+1} - u_K^n)\frac{1}{m_{B_K}}\int_{B_K} \psi(n\Delta t, x)dx\right]$$
$$T_{2,m} = E\left[1_A \sum_{n=0}^{N-1} \Delta t \sum_{K\in\mathscr{T}} \sum_{\sigma\in\mathscr{E}_K\cap\mathscr{E}_{int}} \frac{m_\sigma}{d_{K|L}} \times \frac{(u_K^n - u_L^n)}{m_{B_K}}\int_{B_K} \psi(n\Delta t, x)dx\right]$$
$$T_{3,m} = E\left[1_A \sum_{n=0}^{N-1} \sum_{K\in\mathscr{T}} \lambda m_K u_K^n(W^{n+1} - W^n)\frac{1}{m_{B_K}}\int_{B_K} \psi(n\Delta t, x)dx\right].$$

Proposition 1 allows us to extract firstly a subsequence of $(u_{\mathscr{T}}^{\Delta t})$, still denoted $(u_{\mathscr{T}}^{\Delta t})$, which converges as $m \to +\infty$ to an element u in $L^\infty((0, T); L^2(\Omega \times \Theta))$ for the weak-$\star$ topology. Secondly, since $(u_{\mathscr{T}}^{\Delta t})$ is bounded in $\mathscr{N}_W^2(0, T; L^2(\Omega))$, one can affirm that u is predictable with values in $L^2(\Omega)$ (see Remark 1) and that the previous convergence also holds (up to a subsequence) for the weak topology in $\mathscr{N}_W^2(0, T; L^2(\Omega))$. Now, following [1] (see Remark 6), [2, 5], one shows that

$$T_{1,m} \xrightarrow[m\to+\infty]{} -E\left[1_A \int_0^T \int_\Omega u(t, x)\partial_t\psi(t, x)dxdt\right] - E\left[1_A \int_\Omega u_0(x)\psi(x, 0)dx\right]$$
$$T_{2,m} \xrightarrow[m\to+\infty]{} -E\left[1_A \int_0^T \int_\Omega u(t, x)\Delta\psi(t, x)dxdt\right].$$

By adapting the classical result of the two-point flux approximation scheme, one shows that $u \in L^2((0, T) \times \Theta; H^1(\Omega))$ by introducing a definition of discrete gradient for $(u_{\mathscr{T}}^{\Delta t})$. Since the stochastic integral $I_T : X \mapsto \int_0^T X(t, x, \omega)dW(t)$ is linear and continuous from $\mathscr{N}_W^2(0, T; L^2(\Omega))$ to $L^2(\Omega \times \Theta)$, it is particularly weakly continuous and so the regularity of ψ gives

$$E\left[1_A \int_\Omega \int_0^T u_{\mathscr{T}}^{\Delta t}(t, x)\psi_{\mathscr{T}}(t, x)dW(t)dx\right] \xrightarrow[m\to+\infty]{} E\left[1_A \int_\Omega \int_0^T u(t, x)\psi(t, x)dW(t)dx\right].$$

Using successively Cauchy-Schwarz inequality on $\Omega \times \Theta$, Itô isometry and Proposition 1, one shows by following [2] in the hyperbolic setting that (using the notation $|\Omega|$ for the area of Ω)

$$\left|T_{3,m} - E\left[1_A \int_\Omega \int_0^T \lambda u_{\mathscr{T}}^{\Delta t}(t, x)\psi_{\mathscr{T}}(t, x)dW(t)dx\right]\right|$$
$$\leqslant \sqrt{|\Omega|} \sum_{n=0}^{N-1} \left(\sum_{K\in\mathscr{T}} E\left[\int_K \left|\int_{n\Delta t}^{(n+1)\Delta t} \lambda u_K^n \left(\psi_{\mathscr{T}}(n\Delta t, x) - \psi_{\mathscr{T}}(t, x)\right) dW(t)\right|^2 dx\right]\right)^{\frac{1}{2}}$$
$$= \sqrt{|\Omega|} \sum_{n=0}^{N-1} \left(\sum_{K\in\mathscr{T}} E\left[\int_{n\Delta t}^{(n+1)\Delta t} \int_K \left|\lambda u_K^n(\psi_{\mathscr{T}}(n\Delta t, x) - \psi_{\mathscr{T}}(t, x))\right|^2 dxdt\right]\right)^{\frac{1}{2}}$$

$$\leqslant |\lambda| T\sqrt{|\Omega|}||\partial_t\psi||_\infty \sup_{n\leqslant N} E\left[||u^n_{\mathscr{T}}||^2_{L^2(\Omega)}\right]\sqrt{\Delta t} \xrightarrow[m\to+\infty]{} 0.$$

Thus $T_{3,m} \xrightarrow[m\to+\infty]{} E\left[1_A\int_\Omega\int_0^T \lambda u(t,x)\psi(t,x)dW(t)dx\right]$. Finally, for any $\psi \in \mathscr{A}_T$ and any $\mathbb{P}$-measurable set A, one gets

$$E\left[1_A\int_0^T\int_\Omega u(t,x)\partial_t\psi(t,x)dxdt\right] + E\left[1_A\int_\Omega u_0(x)\psi(0,x)dx\right]$$
$$-E\left[1_A\int_0^T\int_\Omega \nabla u(t,x)\cdot\nabla\psi(t,x)dxdt\right] = -E\left[1_A\int_\Omega\int_0^T \lambda u(t,x)\psi(t,x)dW(t)dx\right],$$

and the result holds using Remark 2. □

Remark 6 In the deterministic case, Theorem 1 is classically proved by multiplying (6) by $\Delta t\psi(n\Delta t, x_K)$ where x_K is the center of the control volume K. Here, the application of Itô isometry gives us a coefficient $\sqrt{\Delta t}$ which is not sufficient to compensate the summation over $n \in \{0, \cdots, N-1\}$. Indeed, in this case there exists $\tilde{C}_\psi > 0$ which only depends on ψ such that

$$\sqrt{|\Omega|}\sum_{n=0}^{N-1}\left(\sum_{K\in\mathscr{T}} E\left[\int_{n\Delta t}^{(n+1)\Delta t}\int_K \left|\lambda u^n_K(\psi(n\Delta t, x_K) - \psi(t,x))\right|^2 dxdt\right]\right)^{\frac{1}{2}}$$
$$\leqslant \sqrt{\Delta t}\sqrt{|\Omega|}|\lambda|\left(\sup_{n\leqslant N} E\left[||u^n_{\mathscr{T}}||^2_{L^2(\Omega)}\right]\right)^{\frac{1}{2}} \tilde{C}_\psi \sum_{n=0}^{N-1}(\Delta t + h),$$

but this last term does not converge to 0 when m tends to $+\infty$ under reasonable assumptions over h and Δt. Choosing here the mean-value on B_K allows us to show both the convergence of the terms $T_{3,m}$ and $T_{2,m}$ (using for $T_{2,m}$ similar arguments as in the deterministic framework, see [1, Proposition 3.5]).

References

1. Andreianov, B., Boyer, F., Hubert, F.: Discrete duality finite volume schemes for Leray-Lions-type elliptic problems on general 2D meshes. Numer. Methods Part. Differ. Equ. **23**(1), 145–195 (2007)
2. Bauzet, C., Charrier, J., Gallouët, T.: Convergence of monotone finite volume schemes for hyperbolic scalar conservation laws with multiplicative noise. Stoch. Part. Differ. Equ. Anal. Comput. **4**, 150–223 (2016)
3. Da Prato, G., Zabczyk, J.: Stochastic equations in infinite dimensions. In: Encyclopedia of Mathematics and its Applications, vol. 44. Cambridge University Press, Cambridge (1992)
4. Droniou, J.: A density result in Sobolev spaces. J. Math. Pures Appl. (9) **81**(7), 697–714 (2002)

5. Eymard, R., Gallouët, T., Herbin, R.: Finite volume methods. In: Handbook of Numerical Analysis, vol. VII. Handb. Numer. Anal. **VII**, 713–1020. North-Holland, Amsterdam (2000)
6. Eymard, R., Gallouët, T., Herbin, R., Michel, A.: Convergence of a finite volume scheme for nonlinear degenerate parabolic equations. Numer. Math. **92**(1), 41–82 (2002)
7. Martel, S.: Theoretical and numerical analysis of invariant measures of viscous stochastic scalar conservation laws. Ph.D. thesis, Université Paris-Est (2019)
8. Prévôt, C., Röckner, M.: A Concise Course on Stochastic Partial Differential Equations. Lecture Notes in Mathematics, vol. 1905. Springer, Berlin (2007)

A New Gradient Scheme of a Time Fractional Fokker–Planck Equation with Time Independent Forcing and Its Convergence Analysis

Abdallah Bradji

Abstract We apply the GDM (Gradient Discretization Method) developed recently in [5, 6] to approximate the time fractional Fokker–Planck equation with time independent forcing in any space dimension. Using [5] which dealt with GDM for linear advection problems, we develop a new fully discrete implicit GS (Gradient Scheme) for the stated model. We prove new discrete a priori estimates which yield estimates on the discrete solution in $L^\infty(L^2)$ and $L^2(H^1)$ discrete norms. Thanks to these discrete a priori estimates, we prove new error estimates in the discrete norms of $L^\infty(L^2)$ and $L^2(H^1)$. The main ingredients in the proof of these error estimates are the use of the stated discrete a priori estimates and a comparison with some well chosen auxiliary schemes. These auxiliary schemes are approximations of convective-diffusive elliptic problems in each time level. We state without proof the convergence analysis of these auxiliary schemes. Such proof uses some adaptations of the [6, Proof of Theorem 2.28] dealt with GDM for the case of elliptic diffusion problems. These results hold for all the schemes within the framework of GDM. This work can be viewed as an extension to our recent one [2].

Keywords Fokker–Planck equation · Time fractional · GDM · Fully discrete implicit GS · Convergence

MSC 2010 65M08 · 65M12 · 65M15

1 Problem to Be Solved and Motivation

The time fractional Fokker–Planck equation can be written as follows, see [7, 8]:

A. Bradji (✉)
LMA Laboratory, University of Annaba, Annaba, Algeria
e-mail: abdallah.bradji@gmail.com; abdallah.bradji@etu.univ-amu.fr
URL: https://www.i2m.univ-amu.fr/perso/abdallah.bradji/

R. Klöfkorn et al. (eds.), *Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples*, Springer Proceedings in Mathematics & Statistics 323,
https://doi.org/10.1007/978-3-030-43651-3_25

$$\begin{aligned}&\partial_t u(\boldsymbol{x},t) - \nabla\cdot\left({}_{\mathrm{RL}}\partial_t^{1-\alpha}\kappa_\alpha\nabla u(\boldsymbol{x},t) - \mathbf{F}(\boldsymbol{x},t)_{\mathrm{RL}}\partial_t^{1-\alpha}u(\boldsymbol{x},t)\right)\\&= g(\boldsymbol{x},t), \qquad (\boldsymbol{x},t)\in\Omega\times(0,T),\end{aligned} \tag{1}$$

where Ω is an open polyhedral bounded subset in $\mathbb{R}^d$ with $d\in\mathbb{N}\setminus\{0\}$, $T>0$, $0<\alpha<1, \kappa_\alpha>0, \mathbf{F}$, and g are given functions. Here the operator ${}_{\mathrm{RL}}\partial_t^{1-\alpha}u(t)$ is the Riemann–Liouville derivative defined by $\partial_t\left(\partial_t^{-\alpha}u(t)\right)$ with $\partial_t^{-\alpha}u(t)$ is the fractional integral operator:

$$\partial_t^{-\alpha}u(t) = \frac{1}{\Gamma(\alpha)}\int_0^t (t-s)^{\alpha-1}u(s)ds. \tag{2}$$

Initial condition is given by

$$u(\boldsymbol{x},0)=0, \qquad \boldsymbol{x}\in\Omega. \tag{3}$$

Homogeneous Dirichlet boundary conditions are given by

$$u(\boldsymbol{x},t)=0, \qquad (\boldsymbol{x},t)\in\partial\Omega\times(0,T). \tag{4}$$

For the sake of simplicity and clarity of the present contribution, we assume that the driving force $\mathbf{F}$ is independent of time. Equation (1) can be written then as:

$$\partial_t u - {}_{\mathrm{RL}}\partial_t^{1-\alpha}\nabla\cdot(\kappa_\alpha\nabla u - \mathbf{F}u) = g.$$

By acting the operator $\partial_t^{\alpha-1}$ on the both sides of the last equation yields (see [7])

$$\partial_t^\alpha u - \nabla\cdot(\kappa_\alpha\nabla u - \mathbf{F}u) = f, \tag{5}$$

where $f=\partial_t^{\alpha-1}g$ and $\partial_t^\alpha u$ is the Caputo derivative of order α given by

$$\partial_t^\alpha u(t) = \frac{1}{\Gamma(1-\alpha)}\int_0^t (t-s)^{-\alpha}u_t(s)ds. \tag{6}$$

The case of time dependent driving force $\mathbf{F}$ will be addressed in a future work. We will assume in addition that $\mathbf{F}\in W^{1,\infty}(\Omega)^d$ and $\operatorname{div}\mathbf{F}(\boldsymbol{x})\geq 0$, for a.e. $\boldsymbol{x}\in\Omega$.

The Fokker–Planck equations describe for instance the time evolution of the probability density function of the position and the velocity of a particle, see [4, 8]. When α tends to one and the diffusion coefficient is constant, we get the standard Fokker–Planck equation. Several numerical methods are devoted to approximate Fokker–Planck equations. We quote among them [4, 7] and [8] which dealt respectively with finite element and finite difference methods. In this contribution, we apply the GDM to approximate the time fractional Fokker–Planck equation with time independent forcing in any space dimension. GDM is a general framework for the discretization and numerical analysis which has been originally designed for elliptic and parabolic partial differential equations, see [6]. Such framework includes conforming and nonconforming finite element, mixed finite element, hybrid mixed mimetic family, some Multi-Point Flux approximation finite volume schemes, and some discontinuous Galerkin schemes. It has been used successfully to approximate other types of differential equations, e.g. nonlinear variational inequalities in [1] and fractional differential equations in [3]. In the present work, we establish a new GS for the considered model. Such scheme is inspired by a GS which has been developed recently in [5] for linear advection problems. We develop some new discrete a priori estimates which serve to prove new error estimates in the discrete norms of $L^2(H^1)$ and $L^\infty(L^2)$ when the exact solution is smooth. These error estimates are proved thanks to a comparison with well chosen auxiliary schemes. We expect that these discrete a priori estimates can also be used to prove the convergence of the family of the approximate solutions towards a unique solution of a weak formulation. This will be addressed thoroughly in a future work. In fact, almost of the existing works on the numerical methods for the considered problem, see [7, 8], focus only on $L^\infty(L^2)$–error estimate. In [4], there are estimates in some energy–norms $\|\cdot\|_{H^{\mu/2}}$, where $\mu \in (1,2)$ for finite element methods but in one space dimension.

2 Space, Time Discretizations, and Preliminaries

Definition 1 (*Definition of an approximate gradient discretization, cf.* [6]) Let Ω be an open domain of $\mathbb{R}^d$, where $d \in \mathbb{N} \setminus \{0\}$. An approximate gradient discretization $\mathscr{D}$ is defined by $\mathscr{D} = (\mathscr{X}_{\mathscr{D},0}, \Pi_{\mathscr{D}}, \nabla_{\mathscr{D}})$, where
1. The set of discrete unknowns $\mathscr{X}_{\mathscr{D},0}$ is a finite dimensional vector space on $\mathbb{R}$.
2. The linear mapping $\Pi_{\mathscr{D}} : \mathscr{X}_{\mathscr{D},0} \to L^2(\Omega)$ is the reconstruction of the approximate function.
3. The gradient reconstruction $\nabla_{\mathscr{D}} : \mathscr{X}_{\mathscr{D},0} \to L^2(\Omega)^d$ is a linear mapping which reconstructs, from an element of $\mathscr{X}_{\mathscr{D},0}$, a "gradient" (vector-valued function) over Ω. The gradient reconstruction must be chosen such that $\|\nabla_{\mathscr{D}} \cdot \|_{L^2(\Omega)^d}$ is a norm on $\mathscr{X}_{\mathscr{D},0}$.

In order to analyse the convergence of schemes, we consider the following parameters related to the approximate gradient discretization $\mathscr{D}$ given in Definition 1.

- The **coercivity** of the discretization is measured via the constant $C_{\mathscr{D}}$ given by $C_{\mathscr{D}} = \max_{v \in \mathscr{X}_{\mathscr{D},0} \setminus \{0\}} \dfrac{\|\Pi_{\mathscr{D}} v\|_{L^2(\Omega)}}{\|\nabla_{\mathscr{D}} v\|_{L^2(\Omega)^d}}$. This yields the following Poincaré inequality:

$$\|\Pi_{\mathscr{D}} v\|_{L^2(\Omega)} \le C_{\mathscr{D}} \|\nabla_{\mathscr{D}} v\|_{L^2(\Omega)^d}, \quad \forall v \in \mathscr{X}_{\mathscr{D},0}. \tag{7}$$

- The **strong consistency** of the discretization is measured through the interpolation error function $S_{\mathscr{D}} : H_0^1(\Omega) \to [0, +\infty)$ defined by, for all $\varphi \in H_0^1(\Omega)$

$$S_{\mathscr{D}}(\varphi) = \min_{v \in \mathscr{X}_{\mathscr{D},0}} \left(\|\Pi_{\mathscr{D}} v - \varphi\|^2_{L^2(\Omega)} + \|\nabla_{\mathscr{D}} v - \nabla\varphi\|^2_{L^2(\Omega)^d} \right)^{\frac{1}{2}}.$$

- The **dual consistency** of the discretization is measured through the conformity error function $W_{\mathscr{D}} : H_{\mathrm{div}}(\Omega) \to [0, +\infty)$ defined by, for all $\varphi \in H_{\mathrm{div}}(\Omega)$

$$W_{\mathscr{D}}(\varphi) = \max_{u \in \mathscr{X}_{\mathscr{D},0}\setminus\{0\}} \frac{1}{\|\nabla_{\mathscr{D}} u\|_{L^2(\Omega)^d}} \left| \int_{\Omega} \left(\nabla_{\mathscr{D}} u(\boldsymbol{x}) \cdot \varphi(\boldsymbol{x}) + \Pi_{\mathscr{D}} u(\boldsymbol{x}) \mathrm{div}\varphi(\boldsymbol{x})\right) d\boldsymbol{x} \right|.$$

The discretization of $[0, T]$ is performed with a constant time step $k = \dfrac{T}{N+1}$, where $N \in \mathbb{N}^\star$, and we shall denote $t_n = nk$, for $n \in [\![0, N+1]\!]$. For a discrete function $(v^n)_{n=0}^{N+1}$, we denote by $\left(\partial^1 v^n\right)_{n=1}^{N+1}$ the discrete temporal derivative given by $\partial^1 v^n = \dfrac{v^n - v^{n-1}}{k}$. We also denote $\partial^0 v^n = v^n$.

Throughout this paper, the letter C stands for a positive constant which is independent of the parameters of the space and time discretizations.

For any $n \in [\![0, N]\!]$, we use the consistent approximation of $\partial_t^\alpha u(t_{n+1})$ which is defined as a linear combination of the discrete time derivatives $\{\partial^1 u(t_{j+1}),\ j \in [\![0, n]\!]\}$ and it is given by (see [3] and references therein):

$$\partial_t^\alpha u(t_{n+1}) = \sum_{j=0}^{n} k\lambda_j^{n+1} \partial^1 u(t_{j+1}) + \mathbb{T}_1^{n+1}(u), \tag{8}$$

where

$$\lambda_j^{n+1} = \frac{(n-j+1)^{1-\alpha} - (n-j)^{1-\alpha}}{k^\alpha \Gamma(2-\alpha)} \quad \text{and} \quad |\mathbb{T}_1^{n+1}(u)| \le Ck^{2-\alpha}\|u\|_{\mathscr{C}^2([0,T])}. \tag{9}$$

The following lemma gives some properties of the coefficients λ_j^{n+1} given by (9).

Lemma 1 (Properties of the coefficients λ_j^{n+1}, cf. [2]) *For any $n \in [\![0, N]\!]$ and for any $j \in [\![0, n]\!]$, let λ_j^{n+1} be defined by (9). The following properties hold:*

$$\frac{k^{-\alpha}}{\Gamma(2-\alpha)} = \lambda_n^{n+1} > \cdots > \lambda_0^{n+1} \ge \lambda_0 = \frac{T^{-\alpha}}{\Gamma(1-\alpha)} \quad \text{and} \quad \sum_{j=0}^{n} k\lambda_j^{n+1} \le \frac{T^{1-\alpha}}{\Gamma(2-\alpha)}.$$

3 First Main Result: Formulation of a New GS for (5) with (3)–(4)

Taking $t = t_{n+1}$ in (5) and using (8) yield

$$\sum_{j=0}^{n} k\lambda_j^{n+1}\partial^1 u(t_{j+1}) - \nabla\cdot(\kappa_\alpha \nabla u(t_{n+1}) - \mathbf{F}u(t_{n+1})) = f(t_{n+1}) - \mathbb{T}_1^{n+1}(u). \quad (10)$$

We set now a new GS for (5) with (3) and (4) based on the approximation (10).

Definition 2 (*Formulation of a GS for* (5) *with* (3)–(4))
Let Ω be a polyhedral open bounded subset of $\mathbb{R}^d$, where $d \in \mathbb{N}\setminus\{0\}$, and $\partial\Omega = \overline{\Omega}\setminus\Omega$ its boundary. Let $\mathscr{D} = (\mathscr{X}_{\mathscr{D},0}, \Pi_{\mathscr{D}}, \nabla_{\mathscr{D}})$ be a gradient discretization in the sense of Definition 1. For any $n \in [\![0, N]\!]$ and for any $j \in [\![0, n]\!]$, let λ_j^{n+1} be defined by (9). We define the following GS as approximation for (5) with (3)–(4): For any $n \in [\![0, N]\!]$, find $u_{\mathscr{D}}^{n+1} \in \mathscr{X}_{\mathscr{D},0}$ such that, for all $v \in \mathscr{X}_{\mathscr{D},0}$

$$\begin{aligned}&\sum_{j=0}^{n} k\lambda_j^{n+1}\left(\partial^1 \Pi_{\mathscr{D}} u_{\mathscr{D}}^{j+1}, \Pi_{\mathscr{D}} v\right)_{L^2(\Omega)} + \kappa_\alpha\left(\nabla_{\mathscr{D}} u_{\mathscr{D}}^{n+1}, \nabla_{\mathscr{D}} v\right)_{L^2(\Omega)^d}\\&+\frac{1}{2}\left(\mathbf{F}\cdot\nabla_{\mathscr{D}} u_{\mathscr{D}}^{n+1}, \Pi_{\mathscr{D}} v\right)_{L^2(\Omega)} - \frac{1}{2}\left(\mathbf{F}\Pi_{\mathscr{D}} u_{\mathscr{D}}^{n+1}, \nabla_{\mathscr{D}} v\right)_{L^2(\Omega)^d}\\&+\frac{1}{2}\left(\mathrm{div}(\mathbf{F})\Pi_{\mathscr{D}} u_{\mathscr{D}}^{n+1}, \Pi_{\mathscr{D}} v\right)_{L^2(\Omega)} = (f(t_{n+1}), \Pi_{\mathscr{D}} v)_{L^2(\Omega)}, \quad (11)\end{aligned}$$

where $u_{\mathscr{D}}^0 = 0$.

4 Second Main Results: New a Priori Estimate and Error Estimate

This section is devoted to analyse the convergence of the GS (11) of Definition 2.

Theorem 1 (New error estimate for GS (11)) *Let $\alpha \in (0, 1)$ be given. Let Ω be a polyhedral open bounded subset of $\mathbb{R}^d$, where $d \in \mathbb{N}\setminus\{0\}$, and $\partial\Omega = \overline{\Omega}\setminus\Omega$ its boundary. Assume that the solution of (5) with (3) and (4) satisfies $u \in \mathscr{C}^2([0, T]; \mathscr{C}^2(\overline{\Omega}))$. Let $k = \dfrac{T}{N+1}$, where $N \in \mathbb{N}\setminus\{0\}$. We shall denote $t_n = nk$, for $n \in [\![0, N+1]\!]$. For any $n \in [\![0, N]\!]$ and for any $j \in [\![0, n]\!]$, let λ_j^{n+1} be defined by (9). Let $\mathscr{D} = (\mathscr{X}_{\mathscr{D},0}, \Pi_{\mathscr{D}}, \nabla_{\mathscr{D}})$ be a gradient discretization in the sense of Definition 1. Then, there exists a unique solution $\left(u_{\mathscr{D}}^n\right)_{n=0}^{N+1} \in \mathscr{X}_{\mathscr{D},0}^{N+2}$ for the GS (11) of Definition 2 and the following $L^\infty(L^2)$ and $L^2(H_0^1)$–error estimates hold:*

$$\max_{n=0}^{N+1} \|\Pi_{\mathscr{D}} u_{\mathscr{D}}^n - u(t_n)\|_{L^2(\Omega)} + \left(\sum_{n=0}^{N} k \|\nabla_{\mathscr{D}} u_{\mathscr{D}}^n - \nabla u(t_n)\|_{L^2(\Omega)^d}^2\right)^{\frac{1}{2}}$$
$$\leq C(1 + C_{\mathscr{D}}) \left(\mathbb{E}_{\mathscr{D}}^k(u) + k^{2-\alpha} \|u\|_{\mathscr{C}^2([0,T];\, L^2(\Omega))}\right), \tag{12}$$

where for any function $u \in \mathscr{C}([0, T];\ H^2(\Omega))$, $\mathbb{E}_{\mathscr{D}}^k(u)$ is an upper bound of the error estimates in discrete norms of $W^{1,\infty}(L^2)$ and $L^\infty(H_0^1)$–norms for the auxiliary GS (21) below, see Lemma 3, and it is given by

$$\mathbb{E}_{\mathscr{D}}^k(u) = \max_{j\in\{0,1\}} \max_{n\in[\![j,N+1]\!]} \mathbb{E}_{\mathscr{D}}(\partial^j u(t_n)) \tag{13}$$

and, for any $\overline{u} \in H^2(\Omega)$

$$\mathbb{E}_{\mathscr{D}}(\overline{u}) = (1 + C_{\mathscr{D}}) \left(W_{\mathscr{D}}(\nabla\overline{u}) + W_{\mathscr{D}}(\boldsymbol{F}\overline{u})\right) + (1 + C_{\mathscr{D}} + C_{\mathscr{D}}^2) S_{\mathscr{D}}(\overline{u}). \tag{14}$$

To prove Theorem 1, we use the following a priori estimates:

Lemma 2 (New discrete a priori estimates) *Under the same hypotheses of Theorem 1, assume that there exists $\left(\eta_{\mathscr{D}}^n\right)_{n=0}^{N+1} \in \mathscr{X}_{\mathscr{D},0}^{N+2}$ such that $\eta_{\mathscr{D}}^0 = 0$ and for all $n \in [\![0, N]\!]$ and for all $v \in \mathscr{X}_{\mathscr{D},0}$*

$$\sum_{j=0}^{n} k\lambda_j^{n+1} \left(\partial^1 \Pi_{\mathscr{D}} \eta_{\mathscr{D}}^{j+1}, \Pi_{\mathscr{D}} v\right)_{L^2(\Omega)} + \kappa_\alpha \left(\nabla_{\mathscr{D}} \eta_{\mathscr{D}}^{n+1}, \nabla_{\mathscr{D}} v\right)_{L^2(\Omega)^d}$$
$$+ \frac{1}{2} \left(\boldsymbol{F} \cdot \nabla_{\mathscr{D}} \eta_{\mathscr{D}}^{n+1}, \Pi_{\mathscr{D}} v\right)_{L^2(\Omega)} - \frac{1}{2} \left(\boldsymbol{F} \cdot \Pi_{\mathscr{D}} \eta_{\mathscr{D}}^{n+1}, \nabla_{\mathscr{D}} v\right)_{L^2(\Omega)^d}$$
$$+ \frac{1}{2} \left(\mathrm{div}(\boldsymbol{F}) \Pi_{\mathscr{D}} \eta_{\mathscr{D}}^{n+1}, \Pi_{\mathscr{D}} v\right)_{L^2(\Omega)} = \left(\mathscr{S}^{n+1}, \Pi_{\mathscr{D}} v\right)_{L^2(\Omega)}, \tag{15}$$

where $\mathscr{S}^{n+1} \in L^2(\Omega)$, for all $n \in [\![0, N]\!]$, is given. Then, the following $L^\infty(L^2)$ and $L^2(H_0^1)$–estimates hold:

$$\max_{n=0}^{N+1} \|\Pi_{\mathscr{D}} \eta_{\mathscr{D}}^n\|_{L^2(\Omega)} + \left(\sum_{n=0}^{N+1} k \|\nabla_{\mathscr{D}} \eta_{\mathscr{D}}^n\|_{L^2(\Omega)^d}^2\right)^{\frac{1}{2}} \leq C(1 + C_{\mathscr{D}}) \mathscr{S}, \tag{16}$$

where $\mathscr{S} = \max_{n=0}^{N} \|\mathscr{S}^{n+1}\|_{L^2(\Omega)}$.

Proof Sketch for Lemma 2

1. **Proof of the $L^2(H_0^1)$–estimate stated in** (16). Taking $v = \eta_{\mathscr{D}}^{n+1}$ in (15) and re-ordering the sum yield

$$\lambda_n^{n+1}\|\Pi_{\mathscr{D}}\eta_{\mathscr{D}}^{n+1}\|^2_{L^2(\Omega)} + \kappa_\alpha\|\nabla_{\mathscr{D}}\eta_{\mathscr{D}}^{n+1}\|^2_{L^2(\Omega)^d} \le \left(\mathscr{S}^{n+1},\ \Pi_{\mathscr{D}}\eta_{\mathscr{D}}^{n+1}\right)_{L^2(\Omega)}$$
$$+ \sum_{j=1}^{n}\left(\lambda_j^{n+1} - \lambda_{j-1}^{n+1}\right)\left(\Pi_{\mathscr{D}}\eta_{\mathscr{D}}^{j}, \Pi_{\mathscr{D}}\eta_{\mathscr{D}}^{n+1}\right)_{L^2(\Omega)}. \tag{17}$$

Using inequality $xy \le x^2/2 + y^2/2$, and the facts that $\lambda_j^{n+1} - \lambda_{j-1}^{n+1} > 0$ (this stems from Lemma 1) and $\lambda_{j-1}^{n+1} = \lambda_j^{n+2}$, inequality (17) implies that

$$\frac{1}{2}\sum_{j=1}^{n+1}\lambda_j^{n+2}\|\Pi_{\mathscr{D}}\eta_{\mathscr{D}}^{j}\|^2_{L^2(\Omega)} - \frac{1}{2}\sum_{j=1}^{n}\lambda_j^{n+1}\|\Pi_{\mathscr{D}}\eta_{\mathscr{D}}^{j}\|^2_{L^2(\Omega)} + \kappa_\alpha\|\nabla_{\mathscr{D}}\eta_{\mathscr{D}}^{n+1}\|^2_{L^2(\Omega)^d}$$
$$\le \left(\mathscr{S}^{n+1},\ \Pi_{\mathscr{D}}\eta_{\mathscr{D}}^{n+1}\right)_{L^2(\Omega)}. \tag{18}$$

Multiplying both sides of (18) by $2k$ and summing the result over $n \in [\![0, N]\!]$ lead to

$$\sum_{j=1}^{N+1} k\lambda_j^{N+2}\|\Pi_{\mathscr{D}}\eta^{j}\|^2_{L^2(\Omega)} + 2\kappa_\alpha\sum_{n=0}^{N} k\|\nabla_{\mathscr{D}}\eta_{\mathscr{D}}^{n+1}\|^2_{L^2(\Omega)^d}$$
$$\le 2k\sum_{n=0}^{N}\left(\mathscr{S}^{n+1},\ \Pi_{\mathscr{D}}\eta_{\mathscr{D}}^{n+1}\right)_{L^2(\Omega)}. \tag{19}$$

Using $xy \le x^2/2 + y^2/2$ and Lemma 1, inequality (19) implies the $L^2(H_0^1)$–estimate stated in (16).

2. **Proof of the $L^\infty(L^2)$–estimate stated in** (16). Using inequality $xy \le x^2/2 + y^2/2$ twice, the Poincaré inequality (7), and Lemma 1, (17) yields

$$\lambda_n^{n+1}\|\Pi_{\mathscr{D}}\eta_{\mathscr{D}}^{n+1}\|^2_{L^2(\Omega)} \le \frac{(C_{\mathscr{D}}\mathscr{S})^2}{\kappa_\alpha} + \sum_{j=1}^{n}\left(\lambda_j^{n+1} - \lambda_{j-1}^{n+1}\right)\|\Pi_{\mathscr{D}}\eta_{\mathscr{D}}^{j}\|^2_{L^2(\Omega)}. \tag{20}$$

Using a mathematical induction on n together with Lemma 1, inequality (20) yields
$\|\Pi_{\mathscr{D}}\eta_{\mathscr{D}}^{n+1}\|^2_{L^2(\Omega)} \le \dfrac{(C_{\mathscr{D}}\mathscr{S})^2}{\lambda_0\kappa_\alpha}$. This gives the $L^\infty(L^2)$–estimate stated in (16). □

To prove Theorem 1, we also need to use the following auxiliary GSs: For any $n \in [\![0, N+1]\!]$, find $\bar{u}_{\mathscr{D}}^n \in \mathscr{X}_{\mathscr{D},0}$ such that, for all $v \in \mathscr{X}_{\mathscr{D},0}$

$$\kappa_\alpha\left(\nabla_{\mathscr{D}}\bar{u}_{\mathscr{D}}^n, \nabla_{\mathscr{D}}v\right)_{L^2(\Omega)^d} + \frac{1}{2}\left(\mathbf{F}\cdot\nabla_{\mathscr{D}}\bar{u}_{\mathscr{D}}^n, \Pi_{\mathscr{D}}v\right)_{L^2(\Omega)} - \frac{1}{2}\left(\mathbf{F}\Pi_{\mathscr{D}}\bar{u}_{\mathscr{D}}^n, \nabla_{\mathscr{D}}v\right)_{L^2(\Omega)^d}$$
$$+ \frac{1}{2}\left(\mathrm{div}(\mathbf{F})\Pi_{\mathscr{D}}\bar{u}_{\mathscr{D}}^n, \Pi_{\mathscr{D}}v\right)_{L^2(\Omega)} = \left(-\kappa_\alpha\Delta u(t_n) + \nabla\cdot(\mathbf{F}u)(t_n), \Pi_{\mathscr{D}}v\right)_{L^2(\Omega)}. \tag{21}$$

For any $n \in [\![0, N+1]\!]$, the scheme (21) is an approximation of a convective-diffusive elliptic problem. The following lemma gives some convergences results for (21):

Lemma 3 (Convergence results for the GSs (21)) *We consider the same hypotheses of Theorem 1. Then, for any $n \in [\![0, N+1]\!]$, the GS (21) has a unique solution and the following error estimates hold:*

$$\max_{n=0}^{n=N+1} \|\nabla u(t_n) - \nabla_{\mathscr{D}} \bar{u}_{\mathscr{D}}^n\|_{L^2(\Omega)} + \max_{j=0}^{j=1} \max_{n=j}^{n=N+1} \|\partial^1 (u(t_n) - \Pi_{\mathscr{D}} \bar{u}_{\mathscr{D}}^n)\|_{L^2(\Omega)} \leq C\mathbb{E}_{\mathscr{D}}^k(u).$$

Proof We will address this in a future paper. It is useful to note that it is possible to provide an explicit order for $\mathbb{E}_{\mathscr{D}}^k(u)$ under some conditions, see [3, Remark 3].

Proof Sketch for Theorem 1

1. **Existence and uniqueness for GS** (11). They can be justified using the facts that (11) yields a linear system whose matrix is square and $\|\nabla_{\mathscr{D}} \cdot \|_{L^2(\Omega)^d}$ is a norm.

2. **Proof of estimates** (12). We follow the following steps:

   - **First step**: comparison between the solution of (21) and the solution of problem (5) with (3) and (4). Thanks to the uniqueness stated in Lemma 3 together with (3), we have $\bar{u}_{\mathscr{D}}^0 = 0$. A comparison between the GS (21) and (5) with (3)–(4) is given in Lemma 3.
   - **Second step**: comparison between the solution of (11) and the auxiliary scheme (21). We set $\eta_{\mathscr{D}}^n = u_{\mathscr{D}}^n - \bar{u}_{\mathscr{D}}^n$. We have then $\eta_{\mathscr{D}}^0 = 0$. Writing now (21) in the level $n+1$, subtracting the result from (11), subtracting $\sum_{j=0}^n k\lambda_j^{n+1} \left(\partial^1 \bar{u}_{\mathscr{D}}^{j+1}, v\right)_{L^2(\Omega)}$ from the both sides of the resulting equation, and using (10) we find that $\left(\eta_{\mathscr{D}}^n\right)_{n=0}^{N+1}$ is satisfying the hypothesis (15) with

$$\mathscr{S}^{n+1} = \sum_{j=0}^{n} k\lambda_j^{n+1} \partial^1 \left(u(t_{j+1}) - \bar{u}_{\mathscr{D}}^{j+1}\right) + \mathbb{T}_1^{n+1}(u).$$

   Applying then the discrete *a priori estimate* (16) and gathering this with Lemma 1, Lemma 3, and (9) to get

$$\left(\sum_{n=0}^{N+1} k\|\nabla_{\mathscr{D}} \eta_{\mathscr{D}}^n\|_{L^2(\Omega)^d}^2\right)^{\frac{1}{2}} + \max_{n=0}^{N+1} \|\Pi_{\mathscr{D}} \eta_{\mathscr{D}}^n\|_{L^2(\Omega)}$$
$$\leq C(1 + C_{\mathscr{D}})(k^{2-\alpha}\|u\|_{\mathscr{C}^2([0,T];\ \mathscr{C}^2(\overline{\Omega}))} + \mathbb{E}_{\mathscr{D}}^k(u)).$$

Gathering this estimate with the fact that $u(t_n) - u_{\mathscr{D}}^n = u(t_n) - \bar{u}_{\mathscr{D}}^n - \eta_{\mathscr{D}}^n$ and Lemma 3 implies the desired estimates (12). This completes the proof of Theorem 1. □

5 Conclusion and Perspectives

We applied GDM to approximate a fractional Fokker–Planck equation with time independent forcing in any space dimension. We established a new fully discrete implicit GS. To analyse the convergence of this scheme, we first proved new discrete a priori estimates. Thanks to these discrete a priori estimates, we proved new error estimates in the discrete norms of $L^2(H^1)$ and $L^\infty(L^2)$. This contribution is an initiation for a future work in which we shall deal with GDM for time and space fractional Fokker–Planck equations with time dependent forcing.

References

1. Alnashri, Y., Droniou, J.: A gradient discretization method to analyze numerical schemes for nonlinear variational inequalities, application to the seepage problem. SIAM J. Numer. Anal. **56**(4), 2375–2405 (2018)
2. Bradji, A.: A new analysis for the convergence of the gradient discretization method for multidimensional time fractional diffusion and diffusion-wave equations. Comput. Math. Appl. **79**(2), 500–520 (2020)
3. Bradji, A.: Notes on the convergence order of gradient schemes for time fractional differential equations. C. R. Math. Acad. Sci. Paris **356**(4), 439–448 (2018)
4. Deng, W.: Finite element method for the space and time fractional Fokker-Planck equation. SIAM J. Numer. Anal. **47**(1), 204–226 (2008/2009)
5. Droniou, J., Eymard, R., Gallouët, T., Herbin, R.: The Gradient discretisation method for linear advection problems. Comput. Methods Appl. Math. https://doi.org/10.1515/cmam-2019-0060. Accessed 17 Oct 2019
6. Droniou, J., Eymard, R., Gallouët, T., Guichard, C., Herbin, R.: The Gradient Discretisation Method. Mathématiques et Applications, vol. 82. Springer Nature Switzerland AG, Switzerland (2018)
7. Le, K.N., McLean, W., Mustapha, K.: A semidiscrete finite element approximation of a time-fractional Fokker-Planck equation with nonsmooth initial data. SIAM J. Sci. Comput. **40**(6), A3831–A3852 (2018)
8. Pinto, L., Sousa, E.: Numerical solution of a time-space fractional Fokker Planck equation with variable force field and diffusion. Commun. Nonlinear Sci. Numer. Simul. **50**, 211–228 (2017)

The Gradient Discretisation Method for Two-Phase Discrete Fracture Matrix Models in Deformable Porous Media

F. Bonaldi, Konstantin Brenner, J. Droniou, and R. Masson

Abstract We consider a two-phase Darcy flow in a fractured porous medium consisting in a matrix flow coupled with a tangential flow in the fractures, described as a network of planar surfaces. This flow model is also coupled with the mechanical deformation of the matrix assuming that the fractures are open and filled by the fluids, as well as small deformations and a linear elastic constitutive law. The model is discretized using the gradient discretization method (Droniou et al. in Mathematics & applications. Springer, 2018, [1]), which covers a large class of conforming and non conforming discretizations. This framework allows a generic convergence analysis of the coupled model using a combination of discrete functional tools. Here, we describe the model together with its numerical discretisation, and we state the convergence result, whose proof will be detailed in a forthcoming paper. This is, to our knowledge, the first convergence result for this type of models taking into account two-phase flows and the nonlinear poro-mechanical coupling. Previous related works consider a linear approximation obtained for a single phase flow by freezing the fracture conductivity (Girault et al. in Math Models Methods Appl Sci 25:4, 2015, [2]).

Keywords Poromechanics · Discrete fracture matrix models · Two-phase darcy flows · Gradient discretization · Convergence analysis

MSC (2010) 65M12 · 76S05 · 74B10

F. Bonaldi (✉) · K. Brenner · R. Masson
Université Côte d'Azur, Inria, CNRS, Laboratoire J.A. Dieudonné, team Coffee,
Nice, France
e-mail: francesco.bonaldi@univ-cotedazur.fr

K. Brenner
e-mail: konstantin.brenner@univ-cotedazur.fr

R. Masson
e-mail: roland.masson@univ-cotedazur.fr

J. Droniou
School of Mathematics, Monash University, Melbourne, Victoria 3800, Australia
e-mail: jerome.droniou@monash.edu

R. Klöfkorn et al. (eds.), *Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples*, Springer Proceedings in Mathematics & Statistics 323,
https://doi.org/10.1007/978-3-030-43651-3_26

1 Continuous Model

We consider a bounded polytopal domain Ω of $\mathbb{R}^d$, $d \in \{2, 3\}$, partitioned into a fracture domain Γ and a matrix domain $\Omega \backslash \overline{\Gamma}$. The network of fractures is $\Gamma = \bigcup_{i \in I} \Gamma_i$, where each Γ_i is planar and has therefore two sides denoted by ± in the matrix domain, with unit normal vectors $\mathbf{n}^{\pm}$ oriented outward to the sides ± (Fig. 1). We denote by γ the trace operator on Γ for functions in $H^1(\Omega)$ and by $[\![\cdot]\!]$ the normal trace jump operator on Γ for functions in $H_{\text{div}}(\Omega \backslash \overline{\Gamma})$. We denote by ∇_τ the tangential gradient and by div_τ the tangential divergence on the fracture network Γ. The symmetric gradient operator $\mathbb{e}$ is defined such that $\mathbb{e}(\mathbf{v}) = \frac{1}{2}(\nabla \mathbf{v} +^t (\nabla \mathbf{v}))$ for a given vector field $\mathbf{v}$.

Let us fix a continuous function $d_0 : \Gamma \to (0, +\infty)$ with zero limits at $\partial\Gamma \setminus (\partial\Gamma \cap \partial\Omega)$ (i.e. the tips of Γ) and stricly positive limits at $\partial\Gamma \cap \partial\Omega$. Let us introduce the following function spaces: $\mathbf{U}_0 = \{\bar{\mathbf{v}} \in (H^1(\Omega \backslash \overline{\Gamma}))^d \mid \gamma_{\partial\Omega} \bar{\mathbf{v}} = 0 \text{ on } \partial\Omega\}$ for the displacement vector, and $V_0 = \{\bar{v} \in H_0^1(\Omega) \mid \gamma \bar{v} \in H_{d_0}^1(\Gamma)\}$ for each phase pressure, where the space $H_{d_0}^1(\Gamma)$ is made of functions v_Γ in $L^2(\Gamma)$, such that $d_0^{3/2} \nabla_\tau v_\Gamma$ is in $L^2(\Gamma)$, whose traces are continuous at fracture intersections $\partial\Gamma_i \cap \partial\Gamma_j$ and vanish on the boundary $\partial\Gamma \cap \partial\Omega$. The matrix and fracture rock types are denoted by the indices rt $= m$ and rt $= f$, respectively, and the non-wetting and wetting phases by the superscripts $\alpha =$ nw and $\alpha =$ w, respectively.

The PDEs model reads: find the phase pressures $\bar{p}^\alpha$, $\alpha \in \{\text{nw}, \text{w}\}$, and the displacement vector field $\bar{\mathbf{u}}$, such that $\bar{p}_c = \bar{p}^{\text{nw}} - \bar{p}^{\text{w}}$ and, for $\alpha \in \{\text{nw}, \text{w}\}$,

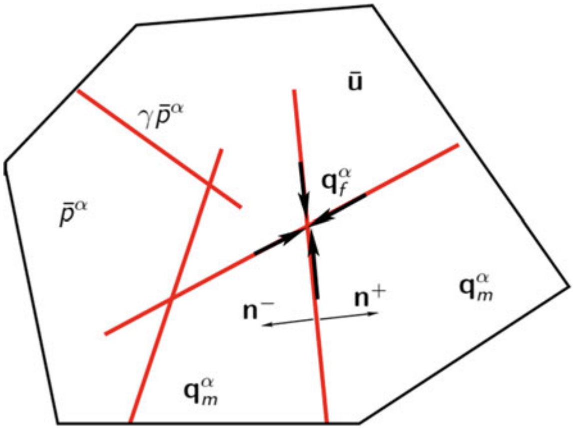


Fig. 1 Example of a 2D domain Ω with its fracture network Γ, the unit normal vectors $\mathbf{n}^{\pm}$ at Γ, the phase pressures $\bar{p}^\alpha$ in the matrix and $\gamma\bar{p}^\alpha$ in the fracture network, the displacement vector field $\bar{\mathbf{u}}$, the matrix Darcy velocities $\mathbf{q}_m^\alpha$ and the fracture tangential Darcy velocities $\mathbf{q}_f^\alpha$ integrated along the fracture

$$\begin{cases} \partial_t\left(\bar{\phi}_m S_m^\alpha(\bar{p}_c)\right) + \operatorname{div}\left(\mathbf{q}_m^\alpha\right) = h_m^\alpha & \text{on } (0,T)\times\Omega\setminus\overline{\Gamma}, \\ \mathbf{q}_m^\alpha = -\eta_m^\alpha(S_m^\alpha(\bar{p}_c))\mathbb{K}_m\nabla\bar{p}^\alpha & \text{on } (0,T)\times\Omega\setminus\overline{\Gamma}, \\ \partial_t\left(\bar{d}_f S_f^\alpha(\gamma\bar{p}_c)\right) + \operatorname{div}_\tau(\mathbf{q}_f^\alpha) - [\![\mathbf{q}_m^\alpha]\!] = h_f^\alpha & \text{on } (0,T)\times\Gamma, \\ \mathbf{q}_f^\alpha = -\eta_f^\alpha(S_f^\alpha(\gamma\bar{p}_c))(\frac{1}{12}\bar{d}_f^3)\nabla_\tau\gamma\bar{p}^\alpha & \text{on } (0,T)\times\Gamma, \\ -\operatorname{div}\left(\sigma(\bar{\mathbf{u}}) - b\,\bar{p}_m^E\mathbb{I}\right) = \mathbf{f} & \text{on } (0,T)\times\Omega\setminus\overline{\Gamma} \\ \sigma(\bar{\mathbf{u}}) = 2\mu\,\epsilon(\bar{\mathbf{u}}) + \lambda\operatorname{div}(\bar{\mathbf{u}})\,\mathbb{I} & \text{on } (0,T)\times\Omega\setminus\overline{\Gamma}, \end{cases} \tag{1}$$

with

$$\begin{cases} \partial_t\bar{\phi}_m = b\operatorname{div}\partial_t\bar{\mathbf{u}} + \frac{1}{M}\partial_t\bar{p}_m^E & \text{on } (0,T)\times\Omega\setminus\overline{\Gamma}, \\ (\sigma(\bar{\mathbf{u}}) - b\,\bar{p}_m^E\mathbb{I})\mathbf{n}^\pm = -\bar{p}_f^E\mathbf{n}^\pm & \text{on } (0,T)\times\Gamma, \\ \bar{d}_f = -[\![\bar{\mathbf{u}}]\!] & \text{on } (0,T)\times\Gamma, \end{cases}$$

and the initial conditions

$$\bar{p}^\alpha|_{t=0} = \bar{p}_0^\alpha, \quad \bar{\phi}_m|_{t=0} = \bar{\phi}_m^0.$$

Here, the equivalent pressures p_m^E and p_f^E are defined, following [3], by

$$\bar{p}_m^E = \sum_{\alpha\in\{\mathrm{nw},\mathrm{w}\}} \bar{p}^\alpha\ S_m^\alpha(\bar{p}_c) - U_m(\bar{p}_c), \quad \bar{p}_f^E = \sum_{\alpha\in\{\mathrm{nw},\mathrm{w}\}} \gamma\bar{p}^\alpha\ S_f^\alpha(\gamma\bar{p}_c) - U_f(\gamma\bar{p}_c),$$

where $U_{\mathrm{rt}}(\bar{p}_c) = \int_0^{\bar{p}_c} q\left(S_{\mathrm{rt}}^{\mathrm{nw}}\right)'(q)dq$ is the capillary energy density function for each rock type rt $\in \{m, f\}$. This is a key choice to obtain the energy estimates which are the starting point for the convergence analysis.

We make the following main assumptions on the data:

- For each phase $\alpha \in \{\mathrm{nw}, \mathrm{w}\}$ and rock type rt $\in \{m, f\}$, the mobility function $\eta_{\mathrm{rt}}^\alpha$ is continuous non-decreasing and there exist $0 < \eta_{\mathrm{rt,min}}^\alpha \le \eta_{\mathrm{rt,max}}^\alpha < +\infty$ such that $\eta_{\mathrm{rt,min}}^\alpha \le \eta_{\mathrm{rt}}^\alpha(s) \le \eta_{\mathrm{rt,max}}^\alpha$ for all $s \in [0, 1]$.
- For each rock type rt $\in \{m, f\}$, $S_{\mathrm{rt}}^{\mathrm{nw}}$ is a non-decreasing Lipschitz continuous function with values in [0, 1], and $S_{\mathrm{rt}}^{\mathrm{w}} = 1 - S_{\mathrm{rt}}^{\mathrm{nw}}$.
- $b \in [0, 1]$ is the Biot coefficient, $M > 0$ is the Biot modulus, and $\lambda > 0$, $\mu > 0$ are the Lamé coefficients. These coefficients are assumed to be constant for simplicity.
- There exist $0 < \phi_{m,min}^0 \le \phi_{m,max}^0 < 1$ such that $\phi_{m,min}^0 \le \bar{\phi}_m^0(\mathbf{x}) \le \phi_{m,max}^0$ for a.e. $\mathbf{x} \in \Omega$.
- The initial fracture aperture satisfies $\bar{d}_f^0(\mathbf{x}) \ge d_0(\mathbf{x})$ for a.e. $\mathbf{x} \in \Gamma$.
- The permeability tensor $\mathbb{K}_m$ is symmetric and uniformly elliptic on Ω.

Definition 1 *(Weak solution of the model)* A weak solution of the model for $\mathbf{f} \in L^2(\Omega)^d$, $h_m^\alpha \in L^2((0,T)\times\Omega)$, and $h_f^\alpha \in L^2((0,T)\times\Gamma)$, is given by $\bar{p}^\alpha \in L^2(0,T;V_0)$, $\alpha \in \{\mathrm{nw}, \mathrm{w}\}$, and $\bar{\mathbf{u}} \in L^\infty(0,T;\mathbf{U}_0)$, such that for any $\alpha \in \{\mathrm{nw}, \mathrm{w}\}$, $\bar{d}_f^{3/2}\nabla_\tau\gamma\bar{p}^\alpha \in L^2((0,T)\times\Gamma))^d$ and, for all $\bar{\varphi}^\alpha \in C_c^\infty([0,T)\times\Omega)$ and all smooth functions $\bar{\mathbf{v}} : [0, T] \times (\Omega \setminus \overline{\Gamma}) \to \mathbb{R}^d$ vanishing on $\partial\Omega$ and having finite limits on each side of Γ,

$$\begin{aligned}
&\int_0^T \int_\Omega \Big(-\bar{\phi}_m S_m^\alpha(\bar{p}_c)\partial_t \bar{\varphi}^\alpha + \eta_m^\alpha(S_m^\alpha(\bar{p}_c))\mathbb{K}_m \nabla \bar{p}^\alpha \cdot \nabla \bar{\varphi}^\alpha \Big) d\mathbf{x} dt \\
&+ \iint_{0\,\Gamma}^{T} \Big(-\bar{d}_f S_f^\alpha(\gamma \bar{p}_c)\partial_t \gamma \bar{\varphi}^\alpha + \eta_f^\alpha(S_f^\alpha(\gamma \bar{p}_c)) \frac{\bar{d}_f^{\,3}}{12} \nabla_\tau \gamma \bar{p}^\alpha \cdot \nabla_\tau \gamma \bar{\varphi}^\alpha \Big) d\sigma(\mathbf{x}) dt \\
&- \int_\Omega \bar{\phi}_m^0 S_m^\alpha(\bar{p}_c^0)\bar{\varphi}^\alpha(0,\cdot) d\mathbf{x} - \int_\Gamma \bar{d}_f^0 S_f^\alpha(\gamma \bar{p}_c^0)\gamma \bar{\varphi}^\alpha(0,\cdot) d\sigma(\mathbf{x}) \\
&= \int_0^T \int_\Omega h_m^\alpha \bar{\varphi}^\alpha d\mathbf{x} dt + \int_0^T \int_\Gamma h_f^\alpha \gamma \bar{\varphi}^\alpha d\sigma(\mathbf{x}) dt,
\end{aligned} \tag{2}$$

$$\begin{aligned}
&\int_0^T \int_\Omega \Big(\sigma(\bar{\mathbf{u}}) : \mathbb{\varepsilon}(\bar{\mathbf{v}}) - b \bar{p}_m^E \mathrm{div}(\bar{\mathbf{v}}) \Big) d\mathbf{x} dt + \int_0^T \int_\Gamma \bar{p}_f^E \, [\![\bar{\mathbf{v}}]\!] d\sigma(\mathbf{x}) dt \\
&= \int_0^T \int_\Omega \mathbf{f} \cdot \bar{\mathbf{v}} d\mathbf{x} dt,
\end{aligned} \tag{3}$$

with $\bar{p}_c = \bar{p}^{\mathrm{nw}} - \bar{p}^{\mathrm{w}}$, $\bar{d}_f = -[\![\bar{\mathbf{u}}]\!]$, $\bar{\phi}_m - \bar{\phi}_m^0 = b\,\mathrm{div}(\bar{\mathbf{u}} - \bar{\mathbf{u}}^0) + \dfrac{1}{M}(\bar{p}_m^E - \bar{p}_m^{E,0})$, $\bar{d}_f^0 = -[\![\bar{\mathbf{u}}^0]\!]$, where $\bar{\mathbf{u}}^0$ is the solution of (3) without the time integral and using the initial equivalent pressures $\bar{p}_m^{E,0}$ and $\bar{p}_f^{E,0}$ obtained from the initial pressures $\bar{p}_0^\alpha \in V_0 \cap L^\infty(\Omega)$, $\gamma \bar{p}_0^\alpha \in L^\infty(\Gamma)$, $\alpha \in \{\mathrm{nw}, \mathrm{w}\}$.

Remark 1 *(Regularity of the displacement field)* Notice that $\bar{\mathbf{u}} \in L^\infty(0, T; \mathbf{U}_0)$ implies $\bar{d}_f = -[\![\bar{\mathbf{u}}]\!] \in L^\infty(0, T; L^4(\Gamma))$. All the integrals above are thus well-defined.

2 The Gradient Scheme

The gradient discretization for the mechanics is defined by the vector space of d.o.f. $X^0_{\mathcal{D}_\mathbf{u}}$ and

- a symmetric gradient operator $\mathbb{\varepsilon}_{\mathcal{D}_\mathbf{u}} : X^0_{\mathcal{D}_\mathbf{u}} \to L^2(\Omega, \mathcal{S}_d(\mathbb{R}))$,
- a displacement function reconstruction operator $\Pi_{\mathcal{D}_\mathbf{u}} : X^0_{\mathcal{D}_\mathbf{u}} \to L^2(\Omega)^d$,
- a normal jump function reconstruction operator $[\![\cdot]\!]_{\mathcal{D}_\mathbf{u}} : X^0_{\mathcal{D}_\mathbf{u}} \to L^4(\Gamma)$,

where $\mathscr{S}_d(\mathbb{R})$ is the vector space of real symmetric matrices of size d. Let us define the divergence operator $\mathrm{div}_{\mathscr{D}_\mathbf{u}}(\cdot) = \mathrm{Trace}(\mathbb{\varepsilon}_{\mathscr{D}_\mathbf{u}}(\cdot))$, the stress tensor operator

$$\mathbb{\sigma}_{\mathscr{D}_\mathbf{u}}(\mathbf{v}) = 2\mu\mathbb{\varepsilon}_{\mathscr{D}_\mathbf{u}}(\mathbf{v}) + \lambda\mathrm{div}_{\mathscr{D}_\mathbf{u}}(\mathbf{v})\mathbb{I},$$

and the fracture width $d_{f,\mathscr{D}_\mathbf{u}} = -[\![\mathbf{u}]\!]_{\mathscr{D}_\mathbf{u}}$. It is assumed that $\|\mathbf{v}\|_{\mathscr{D}_\mathbf{u}} = \|\mathbb{\varepsilon}_{\mathscr{D}_\mathbf{u}}(\mathbf{v})\|_{L^2(\Omega)}$ is a norm on $X^0_{\mathscr{D}_\mathbf{u}}$.

The gradient discretization (GD) of the Darcy continuous pressure model is introduced in [4] and defined by the vector space of d.o.f. $X^0_{\mathscr{D}_p}$ and

- two discrete gradient operators on the matrix and fracture domains

$$\nabla^m_{\mathscr{D}_p} : X^0_{\mathscr{D}_p} \to L^\infty(\Omega)^d, \qquad \nabla^f_{\mathscr{D}_p} : X^0_{\mathscr{D}_p} \to L^\infty(\Gamma)^{d-1};$$

- two function reconstruction operators on the matrix and fracture domains

$$\Pi^m_{\mathscr{D}_p} : X^0_{\mathscr{D}_p} \to L^\infty(\Omega), \qquad \Pi^f_{\mathscr{D}_p} : X^0_{\mathscr{D}_p} \to L^\infty(\Gamma),$$

which are piecewise constant [1, Definition 2.12].

A consequence of the piecewise-constant property is that, for any $g : \mathbb{R} \to \mathbb{R}$ and $v \in X^0_{\mathscr{D}_p}$, we can define $g(v) \in X^0_{\mathscr{D}_p}$ component-wise and we have $\Pi^\rho_{\mathscr{D}_p} g(v) = g(\Pi^\rho_{\mathscr{D}_p} v)$ for $\rho \in \{m, f\}$. Fixing a continuous function $d_0 : \Gamma \to (0, +\infty)$ with zero limits at the tips of Γ, the vector space $X^0_{\mathscr{D}_p}$ is endowed with $\|v\|_{\mathscr{D}_p} = \|\nabla^m_{\mathscr{D}_p} v\|_{L^2(\Omega)^d} + \|d_0^{\frac{3}{2}} \nabla^f_{\mathscr{D}_p} v\|_{L^2(\Gamma)^{d-1}}$, assumed to define a norm on $X^0_{\mathscr{D}_p}$.

This spatial GD is extended into a space-time GD by complementing it with

- a discretisation $0 = t_0 < t_1 < \cdots < t_N = T$ of the time interval $[0, T]$;
- interpolators $P_{\mathscr{D}_p} : V_0 \to X^0_{\mathscr{D}_p}$ and $P^m_{\mathscr{D}_p} : L^2(\Omega) \to X^0_{\mathscr{D}_p}$ of initial conditions.

The spatial operators are extended into space-time operators as follows. Let χ represent either p or $\mathbf{u}$. If $w = (w_n)_{n=0}^N \in (X^0_{\mathscr{D}_\chi})^{N+1}$, and $\Psi_{\mathscr{D}_\chi}$ is a spatial GDM operator, its space-time extension is defined by

$$\Psi_{\mathscr{D}_\chi} w(0, \cdot) = \Psi_{\mathscr{D}_\chi} w_0 \text{ and}, \forall n \in \{0, \ldots, N-1\}, \ \forall t \in (t_n, t_{n+1}], \ \Psi_{\mathscr{D}_\chi} w(t, \cdot) = \Psi_{\mathscr{D}_\chi} w_{n+1}.$$

where, for convenience, the same notation is kept for the spatial and space-time operators. We also define the discrete time derivative as follows: for $f : [0, T] \to L^1(\Omega)$ piecewise constant on the time discretisation, with $f_n = f_{|(t_{n-1}, t_n]}$, and using the same n and t as above, $\delta_t f(t) = \frac{f_{n+1} - f_n}{t_{n+1} - t_n}$.

The gradient scheme for (1) consists in writing the weak formulation (2)–(3) with continuous spaces and operators substituted by their discrete counterparts, after a formal integration by part: find $p^\alpha \in (X^0_{\mathscr{D}_p})^{N+1}$, $\alpha \in \{\mathrm{nw}, \mathrm{w}\}$, and $\mathbf{u} \in (X^0_{\mathscr{D}_\mathbf{u}})^{N+1}$, such that for all $\varphi^\alpha \in (X^0_{\mathscr{D}_p})^{N+1}$, $\mathbf{v} \in (X^0_{\mathscr{D}_\mathbf{u}})^{N+1}$ and $\alpha \in \{\mathrm{nw}, \mathrm{w}\}$,

$$\begin{aligned}
&\int_0^T\int_\Omega \Big(\delta_t\Big(\phi_{\mathscr{D}} \Pi^m_{\mathscr{D}_p} s^\alpha_m\Big)\Pi^m_{\mathscr{D}_p}\varphi^\alpha + \eta^\alpha_m(\Pi^m_{\mathscr{D}_p} s^\alpha_m)\mathbb{K}_m \nabla^m_{\mathscr{D}_p} p^\alpha \cdot \nabla^m_{\mathscr{D}_p}\varphi^\alpha\Big) d\mathbf{x}dt \\
&+\int_0^T\int_\Gamma \delta_t\Big(d_{f,\mathscr{D}_\mathbf{u}}\Pi^f_{\mathscr{D}_p} s^\alpha_f\Big)\Pi^f_{\mathscr{D}_p}\varphi^\alpha d\sigma(\mathbf{x}) \\
&+\int_0^T\int_\Gamma \eta^\alpha_f(\Pi^f_{\mathscr{D}_p} s^\alpha_f)\frac{d^3_{f,\mathscr{D}_\mathbf{u}}}{12}\nabla^f_{\mathscr{D}_p} p^\alpha \cdot \nabla^f_{\mathscr{D}_p}\varphi^\alpha d\mathbf{x}dt \\
&=\int_0^T\int_\Omega h^\alpha_m \Pi^m_{\mathscr{D}_p}\varphi^\alpha d\mathbf{x}dt + \int_0^T\int_\Gamma h^\alpha_f \Pi^f_{\mathscr{D}_p}\varphi^\alpha d\sigma(\mathbf{x})dt,
\end{aligned} \tag{4a}$$

$$\begin{aligned}
&\int_0^T\int_\Omega \Big(\sigma_{\mathscr{D}_u}(\mathbf{u}) : \varepsilon_{\mathscr{D}_\mathbf{u}}(\mathbf{v}) - b(\Pi^m_{\mathscr{D}_p} p^E_m)\mathrm{div}_{\mathscr{D}_\mathbf{u}}(\mathbf{v})\Big) d\mathbf{x}dt \\
&\qquad + \int_0^T\int_\Gamma (\Pi^f_{\mathscr{D}_p} p^E_f)[\![\mathbf{v}]\!]_{\mathscr{D}_\mathbf{u}} d\sigma(\mathbf{x})dt = \int_0^T\int_\Omega \mathbf{f}\cdot \Pi_{\mathscr{D}_\mathbf{u}}\mathbf{v} d\mathbf{x}dt,
\end{aligned} \tag{4b}$$

with the closure equations

$$\left\{\begin{array}{l}
p_c = p^{\mathrm{nw}} - p^{\mathrm{w}}, \quad s^\alpha_m = S^\alpha_m(p_c), \quad s^\alpha_f = S^\alpha_f(p_c), \\
p^E_m = \displaystyle\sum_{\alpha\in\{\mathrm{nw},\mathrm{w}\}} p^\alpha s^\alpha_m - U_m(p_c), \quad p^E_f = \displaystyle\sum_{\alpha\in\{\mathrm{nw},\mathrm{w}\}} p^\alpha s^\alpha_f - U_f(p_c), \\
\phi_{\mathscr{D}} - \Pi^m_{\mathscr{D}_p}\bar{\phi}^0_m = b\,\mathrm{div}_{\mathscr{D}_\mathbf{u}}(\mathbf{u}-\mathbf{u}^0) + \frac{1}{M}\Pi^m_{\mathscr{D}_p}(p^E_m - p^{E,0}_m).
\end{array}\right. \tag{4c}$$

The initial conditions are given by $p^\alpha_0 = P_{\mathscr{D}_p}\bar{p}^\alpha_0$ $(\alpha \in \{\mathrm{nw}, \mathrm{w}\})$, $\phi^0_m = P^m_{\mathscr{D}_p}\bar{\phi}^0$, and the initial displacement $\mathbf{u}^0$ is the solution of (4b) with the equivalent pressures obtained from the initial pressures $(p^\alpha_0)_{\alpha\in\{\mathrm{nw},\mathrm{w}\}}$.

3 Convergence Result

Let $(\mathscr{D}^l_p)_{l\in\mathbb{N}}$ and $(\mathscr{D}^l_\mathbf{u})_{l\in\mathbb{N}}$ be sequences of GDs. We state here the assumptions on these sequences which ensure that the solutions to the corresponding schemes converge. Most of these assumptions are adaptation of classical GDM assumptions [1], except for the chain-rule and cut-off properties, whose role is briefly discussed at the end of the paper; we note that all these assumptions hold for standard discretisations used in porous media flows.

Coercivity, consistency and limit-conformity of $(\mathscr{D}_p^l)_{l\in\mathbb{N}}$: these properties are omitted since they are similar to those in [4], the only change being the use in the definition of consistency of the L^r-norm with $r > 8$, instead of the L^2-norm, for the gradient in the fractures, and the use of fracture fluxes $\mathbf{q}_f$ vanishing at the fracture tips in the definition of the limit-conformity.

Chain rule estimate on $(\mathscr{D}_p^l)_{l\in\mathbb{N}}$: for any Lipschitz-continuous function $F : \mathbb{R} \to \mathbb{R}$, there is $C_F \geq 0$ such that, for all $l \in \mathbb{N}$ and $v \in X^0_{\mathscr{D}_p^l}$, $\|\nabla^m_{\mathscr{D}_p^l} F(v)\|_{L^2(\Omega)^d} \leq C_F \|\nabla^m_{\mathscr{D}_p^l} v\|_{L^2(\Omega)^d}$.

Cut-off property of $(\mathscr{D}_p^l)_{l\in\mathbb{N}}$: for any compact set $K \subset \Omega\backslash\Gamma$ and $l \in \mathbb{N}$, there exists $\psi^l \in X_{\mathscr{D}_p^l}$ such that, for l large enough and $C \geq 0$ not depending on l: $\Pi^m_{\mathscr{D}_p^l}\psi^l \geq 0$ on Ω; $\Pi^m_{\mathscr{D}_p^l}\psi^l \geq 1$ on K; $\|\nabla^m_{\mathscr{D}_p^l}\psi^l\|_{L^2(\Omega)^d} \leq C$; $\Pi^f_{\mathscr{D}_p^l}\psi^l = 0$; and $\nabla^f_{\mathscr{D}_p^l}\psi^l = 0$.

Coercivity of $(\mathscr{D}_\mathbf{u}^l)_{l\in\mathbb{N}}$. It holds

$$\sup_{l\in\mathbb{N}} \max_{\mathbf{v}\in X^0_{\mathscr{D}_\mathbf{u}^l}\backslash\{0\}} \frac{\|\Pi_{\mathscr{D}_\mathbf{u}^l}\mathbf{v}\|_{L^2(\Omega)^d} + \|[\![\mathbf{v}]\!]_{\mathscr{D}_\mathbf{u}^l}\|_{L^4(\Gamma)}}{\|\mathbf{v}\|_{\mathscr{D}_\mathbf{u}^l}} < +\infty. \tag{5}$$

Consistency of $(\mathscr{D}_\mathbf{u}^l)_{l\in\mathbb{N}}$. For all $\bar{\mathbf{u}} \in \mathbf{U}_0$, it holds $\lim_{l\to+\infty}\mathscr{S}_{\mathscr{D}_\mathbf{u}^l}(\bar{\mathbf{u}}) = 0$ where

$$\mathscr{S}_{\mathscr{D}_\mathbf{u}^l}(\bar{\mathbf{u}}) = \min_{\mathbf{v}\in X^0_{\mathscr{D}_\mathbf{u}^l}} \Big[\|\mathbb{\epsilon}_{\mathscr{D}_\mathbf{u}^l}(\mathbf{v}) - \mathbb{\epsilon}(\bar{\mathbf{u}})\|_{L^2(\Omega,\mathscr{S}_d(\mathbb{R}))} + \|\Pi_{\mathscr{D}_\mathbf{u}^l}\mathbf{v} - \bar{\mathbf{u}}\|_{L^2(\Omega)^d} + \|[\![\mathbf{v}]\!]_{\mathscr{D}_\mathbf{u}^l} - [\![\bar{\mathbf{u}}]\!]\|_{L^4(\Gamma)}\Big].$$

Limit Conformity of $(\mathscr{D}_\mathbf{u}^l)_{l\in\mathbb{N}}$. Let $C^\infty_\Gamma(\Omega\setminus\overline{\Gamma}, \mathscr{S}_d(\mathbb{R}))$ denote the vector space of smooth functions $\sigma(\mathbf{x})$ from $\Omega\setminus\overline{\Gamma}$ to $\mathscr{S}_d(\mathbb{R})$ defined as above, and such that $\sigma^+(\mathbf{x})\mathbf{n}^+ + \sigma^-(\mathbf{x})\mathbf{n}^- = \mathbf{0}$ and $(\sigma^+(\mathbf{x})\mathbf{n}^+)\times\mathbf{n}^+ = \mathbf{0}$ for a.e. $\mathbf{x}\in\Gamma$. For all $\sigma \in C^\infty_\Gamma(\Omega\setminus\overline{\Gamma}, \mathscr{S}_d(\mathbb{R}))$, it holds $\lim_{l\to+\infty}\mathscr{W}_{\mathscr{D}_\mathbf{u}^l}(\sigma) = 0$ where

$$\mathscr{W}_{\mathscr{D}_\mathbf{u}^l}(\sigma) = \max_{\mathbf{v}\in X^0_{\mathscr{D}_\mathbf{u}^l}\backslash\{0\}} \frac{1}{\|\mathbf{v}\|_{\mathscr{D}_\mathbf{u}^l}}\Big[\int_\Omega \Big(\sigma : \mathbb{\epsilon}_{\mathscr{D}_\mathbf{u}^l}(\mathbf{v}) + \Pi_{\mathscr{D}_\mathbf{u}^l}\mathbf{v}\,\mathrm{div}(\sigma)\Big)d\mathbf{x} - \int_\Gamma \Big((\sigma\mathbf{n}^+)\cdot\mathbf{n}^+[\![\mathbf{v}]\!]_{\mathscr{D}_\mathbf{u}^l}d\sigma(\mathbf{x})\Big].$$

Compactness of $(\mathscr{D}_\mathbf{u}^l)_{l\in\mathbb{N}}$. For any sequence $(\mathbf{v}^l)_{l\in\mathbb{N}}$ with $\mathbf{v}^l \in X^0_{\mathscr{D}_\mathbf{u}^l}$ for all $l\in\mathbb{N}$ such that $\sup_{l\in\mathbb{N}}\|\mathbf{v}^l\|_{\mathscr{D}_\mathbf{u}^l} < +\infty$, the sequences $(\Pi_{\mathscr{D}_\mathbf{u}^l}\mathbf{v}^l)_{l\in\mathbb{N}}$ and $([\![\mathbf{v}^l]\!]_{\mathscr{D}_\mathbf{u}^l})_{l\in\mathbb{N}}$ are relatively compact in $L^2(\Omega)^d$ and in $L^s(\Gamma)$ for all $s < 4$, respectively.

We can now state the convergence result.

Theorem 1 *Let t^l_n, $n = 0, \cdots, N^l$ and $l \in \mathbb{N}$, be a sequence of time discretizations such that* $\lim_{l \to +\infty} \max_{n=0,\cdots,N^l-1}(t^l_{n+1} - t^l_n) = 0$. *Let* $0 < \phi_{m,min} \le \phi_{m,max} < +\infty$ *and assume that, for each $l \in \mathbb{N}$, the gradient scheme* (4a)–(4b) *has a solution $p^\alpha_l \in X^0_{\mathscr{D}^l_p}$, $\alpha \in \{\mathrm{nw}, \mathrm{w}\}$, $\mathbf{u}^l \in X^0_{\mathscr{D}^l_\mathbf{u}}$ such that*

(i) $d_{f,\mathscr{D}^l_\mathbf{u}}(t, \mathbf{x}) \ge d_0(\mathbf{x})$ *for a.e.* $(t, \mathbf{x}) \in (0, T) \times \Gamma$,
(ii) $\phi_{m,min} \le \phi_{\mathscr{D}^l}(t, \mathbf{x}) \le \phi_{m,max}$ *for a.e.* $(t, \mathbf{x}) \in (0, T) \times \Omega$.

Then, there exist $\bar{p}^\alpha \in L^2(0, T; V_0)$, $\alpha \in \{\mathrm{nw}, \mathrm{w}\}$, and $\bar{\mathbf{u}} \in L^\infty(0, T; \mathbf{U}_0)$ solutions of the weak formulation (2)–(3) *such that for $\alpha \in \{\mathrm{nw}, \mathrm{w}\}$ and up to a subsequence*

$$\left\{\begin{array}{l}
\Pi^m_{\mathscr{D}^l_p} p^\alpha_l \rightharpoonup \bar{p}^\alpha \textit{ in } L^2(0, T; L^2(\Omega)),\\
\Pi^f_{\mathscr{D}^l_p} p^\alpha_l \rightharpoonup \gamma \bar{p}^\alpha \textit{ in } L^2(0, T; L^2(\Gamma)),\\
\Pi_{\mathscr{D}^l_\mathbf{u}} \mathbf{u}^l \rightharpoonup \bar{\mathbf{u}} \textit{ in } L^\infty(0, T; L^2(\Omega)^d) \textit{ weak } \star,\\
d_{f,\mathscr{D}^l_\mathbf{u}} \to \bar{d}_f \textit{ in } L^\infty(0, T; L^p(\Gamma)) \textit{ for } 2 \le p < 4,\\
\phi_{\mathscr{D}^l} \rightharpoonup \bar{\phi}_m \textit{ in } L^\infty(0, T; L^2(\Omega)) \textit{ weak } \star,\\
\Pi^m_{D^l_p} S^\alpha_m(p^l_c) \to S^\alpha_m(\bar{p}_c) \textit{ in } L^2(0, T; L^2(\Omega)),\\
\Pi^f_{D^l_p} S^\alpha_f(p^l_c) \to S^\alpha_f(\gamma \bar{p}_c) \textit{ in } L^2(0, T; L^2(\Gamma)).
\end{array}\right.$$

The proof of Theorem 1 hinges on the following steps:

- Inferring energy estimates by using suitable test functions;
- Obtaining weak estimates on time derivatives;
- Using the discontinuous Ascoli–Arzelà compactness theorem [1, Theorem C.11] to prove convergences;
- Identifying the limit fields.

We report here the energy estimate satisfied by the discrete unknowns. For a piecewise constant function v on $[0, T]$ with $v(t) = v_{n+1}$ for all $t \in (t_n, t_{n+1}]$, $n = 0, \cdots, N - 1$ and the initial value $v(0) = v_0$, we define the piecewise constant function $\hat{v}$ such that $\hat{v}(t) = v_n$ for all $t \in (t_n, t_{n+1}]$. Upon choosing $\varphi^\alpha = p^\alpha$ in (4a) and $\mathbf{v} = \delta_t \mathbf{u}$ in (4b), using the fact that $\delta_t(uv)(t) = \hat{u}(t)\delta_t v(t) + v(t)\delta_t u(t)$, summing the corresponding equations, and using the closure equations (4c) along with the assumptions we made on the data, we obtain the following estimate for the solutions of (4): there is a real number $C > 0$ depending on the data such that

$$\int_0^T \int_\Omega \delta_t(\phi_{\mathscr{D}} U_m(\Pi^m_{\mathscr{D}_p} p_c))\, d\mathbf{x}dt + \int_0^T \int_\Gamma \delta_t(d_{f,\mathscr{D}_\mathbf{u}} U_f(\Pi^f_{\mathscr{D}_p} p_c))\, d\sigma(\mathbf{x})dt$$
$$+ \int_0^T \int_\Omega \delta_t \left(\frac{1}{2} \left(\mathbb{\sigma}_{\mathscr{D}_\mathbf{u}}(\mathbf{u}) : \mathbb{\epsilon}_{\mathscr{D}_\mathbf{u}}(\mathbf{u}) \right) + \frac{1}{2M} (\Pi^m_{\mathscr{D}_p} p^E_m)^2 \right) d\mathbf{x}dt$$
$$+ \sum_{\alpha \in \{\mathrm{w},\mathrm{nw}\}} \int_0^T \int_\Omega |\nabla^m_{\mathscr{D}_p} p^\alpha|^2\, d\mathbf{x}dt + \sum_{\alpha \in \{\mathrm{w},\mathrm{nw}\}} \int_0^T \int_\Gamma d^3_{f,\mathscr{D}_\mathbf{u}} |\nabla^f_{\mathscr{D}_p} p^\alpha|^2\, d\sigma(\mathbf{x})dt \qquad (6)$$

$$\leq C\left(\int_0^T\int_\Omega \mathbf{f}\cdot\delta_t \Pi_{\mathscr{D}_{\mathbf{u}}}\mathbf{u}\,d\mathbf{x}dt + \sum_{\alpha\in\{\mathrm{w},\mathrm{nw}\}}\int_0^T\int_\Omega h_m^\alpha \Pi_{\mathscr{D}_p}^m p^\alpha\, d\mathbf{x}dt \right.$$
$$\left. + \sum_{\alpha\in\{\mathrm{w},\mathrm{nw}\}}\int_0^T\int_\Gamma h_f^\alpha \Pi_{\mathscr{D}_p}^f p^\alpha\, d\sigma(\mathbf{x})dt\right).$$

The left-hand side of this inequality is made of positive terms (up to initial conditions, that appear in the telescopic sums corresponding to the first three terms), with enough quadratic growth in the unknowns to compensate the linear dependency of the right-hand side on these unknowns.

The chain-rule estimates and cut-off properties of $(\mathscr{D}_p^l)_{l\in\mathbb{N}}$ are used to prove estimates on the time-translates of $\Pi_{D_p^l}^m S_m^\alpha(p_c^l)$ (which are crucial in establishing the strong convergence of this quantity). These estimates require to separate the matrix and fracture components (hence the need for using cut-off test functions in the scheme), and is based on a dual estimate that requires to use $S_m^\alpha(p_c^l)$ as a test function and estimate its gradient (which follows from gradient estimates on p_c^l and the chain-rule estimates).

Acknowledgements We are grateful to Andra and to the Australian Research Council's Discovery Projects (project DP170100605) funding scheme for partially supporting this work.

References

1. Droniou, J., Eymard, R., Gallouët, T., Guichard, C., Herbin, R.: The gradient discretisation method. Mathematics & Applications. Springer, vol. 82 (2018). https://doi.org/10.1007/978-3-319-79042-8
2. Girault, V., Wheeler, M., Ganis, B., Mear, M.: A lubrication fracture model in a poro-elastic medium. Math. Models Methods Appl. Sci. **25**, 4 (2015)
3. Coussy, O.: Poromechanics. Wiley (2004)
4. Brenner, K., Groza, M., Guichard, C., Lebeau, G., Masson, R.: Gradient discretization of hybrid-dimensional darcy flows in fractured porous media. Numerische Mathematik **134**(3), 569–609 (2016)

A New Optimal $L^\infty(H^1)$–Error Estimate of a SUSHI Scheme for the Time Fractional Diffusion Equation

Abdallah Bradji

Abstract We consider a finite volume scheme, using the general mesh of [8], for the **TFDE** (time fractional diffusion equation) in any space dimension. The time discretization is performed using a uniform mesh. We prove a new discrete $L^\infty(H^1)$–*a priori estimate*. Such *a priori estimate* is proved thanks to the use of the new tool of the discrete Laplace operator developed recently in [7]. Thanks to this *a priori estimate*, we prove a new optimal convergence order in the discrete $L^\infty(H^1)$–norm. These results improve the ones of [1, 4] which dealt respectively with finite volume and GDM (Gradient Discretization Method) for the **TFDE**. In [4], we only proved *a priori estimate* and error estimate in the discrete $L^\infty(L^2)$–norm whereas in [1] we proved *a priori estimate* and error estimate in the discrete $L^2(H^1)$–norm. The a priori estimate as well as the error estimate presented here were stated without proof for the first time in [3, Remark 1, p. 443] in the context of the general framework of GDM and [2, Remark 1, p. 205] in the context of finite volume methods. They also were mentioned, as future works, in [1, Remark 4.1].

Keywords Time fractional diffusion equation · SUSHI · *A priori estimate* · $L^\infty(H^1)$–error estimate

MSC 2010 65M08 · 65M12 · 65M15

1 Problem to Be Solved and Motivation

We consider the following time fractional diffusion equation:

$$\partial_t^\alpha u(\boldsymbol{x}, t) - \Delta u(\boldsymbol{x}, t) = f(\boldsymbol{x}, t), \qquad (\boldsymbol{x}, t) \in \Omega \times (0, T), \tag{1}$$

A. Bradji (✉)
LMA Laboratory, University of Annaba, Annaba, Algeria
e-mail: abdallah.bradji@gmail.com; abdallah.bradji@etu.univ-amu.fr
URL: https://www.i2m.univ-amu.fr/perso/abdallah.bradji/

R. Klöfkorn et al. (eds.), *Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples*, Springer Proceedings in Mathematics & Statistics 323,
https://doi.org/10.1007/978-3-030-43651-3_27

where Ω is an open polygonal bounded subset in $\mathbb{R}^d$, $T > 0$, $0 < \alpha < 1$, and f is a given function. Here the operator ∂_t^α is the Caputo derivative defined by:

$$\partial_t^\alpha u(t) = \frac{1}{\Gamma(1-\alpha)} \int_0^t (t-s)^{-\alpha} u_t(s) ds. \tag{2}$$

Initial condition is given by, for a given function u^0 defined on Ω

$$u(\boldsymbol{x}, 0) = u^0(\boldsymbol{x}), \qquad \boldsymbol{x} \in \Omega. \tag{3}$$

Homogeneous Dirichlet boundary conditions are given by

$$u(\boldsymbol{x}, t) = 0, \qquad (\boldsymbol{x}, t) \in \partial\Omega \times (0, T). \tag{4}$$

Fractional differential equations have been successfully used in theory and they appear in many areas of application, see [1, 9].

We consider in this note a cell-centered finite volume scheme, using the general class of meshes introduced in [8], for **TFDE**. The first aim of this contribution is to prove a $L^\infty(H^1)$–*a priori estimate* and the second one is to use this *a priori estimate* to derive an optimal convergence order in the discrete norm of $L^\infty(H^1)$. This improves the results of our previous works [1, 3, 4] which dealt with *a priori estimate* and error estimate in the discrete norms of $L^\infty(L^2)$ or $L^2(H^1)$. As mentioned above in the Abstract, the a priori estimate as well as the error estimate presented here were stated without proof for the first time in [3, Remark 1, p. 443] in the context of the general framework of GDM and in [2, Remark1, p. 205] in the context of finite volume methods. The proof of *a priori estimate* and error estimate in a discrete norm of $L^\infty(H^1)$ is not straightforward in the sense that the usual techniques do not lead to such results, we refer to Remark 1 below.

2 Space, Time Discretizations, and the Definition of a Discrete Gradient

Definition 1 (*Space discretization, cf.* [8]) Let Ω be a polyhedral open bounded subset of $\mathbb{R}^d$, where $d \in \mathbb{N} \setminus \{0\}$, and $\partial\Omega = \overline{\Omega} \setminus \Omega$ its boundary. A discretization of Ω, denoted by $\mathscr{D}$, is defined as the triplet $\mathscr{D} = (\mathscr{M}, \mathscr{E}, \mathscr{P})$, where:

1. $\mathscr{M}$ is a finite family of non empty connected open disjoint subsets of Ω (the "control volumes") such that $\overline{\Omega} = \cup_{K\in\mathscr{M}} \overline{K}$. For any $K \in \mathscr{M}$, let $\partial K = \overline{K} \setminus K$ be the boundary of K; let $\mathrm{m}(K) > 0$ denote the measure of K and h_K denote the diameter of K.

2. $\mathscr{E}$ is a finite family of disjoint subsets of $\overline{\Omega}$ (the "edges" of the mesh), such that, for all $\sigma \in \mathscr{E}$, σ is a non empty open subset of a hyperplane of $\mathbb{R}^d$, whose $(d-1)$–dimensional measure is strictly positive. We also assume that, for all $K \in \mathscr{M}$, there exists a subset $\mathscr{E}_K$ of $\mathscr{E}$ such that $\partial K = \cup_{\sigma \in \mathscr{E}_K} \overline{\sigma}$. For any $\sigma \in \mathscr{E}$, we denote by $\mathscr{M}_\sigma = \{K, \sigma \in \mathscr{E}_K\}$. We then assume that, for any $\sigma \in \mathscr{E}$, either $\mathscr{M}_\sigma$ has exactly one element and then $\sigma \subset \partial\,\Omega$ (the set of these interfaces, called boundary interfaces, denoted by $\mathscr{E}_{\rm ext}$) or $\mathscr{M}_\sigma$ has exactly two elements (the set of these interfaces, called interior interfaces, denoted by $\mathscr{E}_{\rm int}$). For all $\sigma \in \mathscr{E}$, we denote by $\boldsymbol{x}_\sigma$ the barycentre of σ. For all $K \in \mathscr{M}$ and $\sigma \in \mathscr{E}$, we denote by $\mathbf{n}_{K,\sigma}$ the unit vector normal to σ outward to K.
3. $\mathscr{P}$ is a family of points of Ω indexed by $\mathscr{M}$, denoted by $\mathscr{P} = (\boldsymbol{x}_K)_{K\in\mathscr{M}}$, such that for all $K \in \mathscr{M}$, $\boldsymbol{x}_K \in K$ and K is assumed to be $\boldsymbol{x}_K$–star-shaped, which means that for all $\boldsymbol{x} \in K$, the property $[\boldsymbol{x}_K, \boldsymbol{x}] \subset K$ holds. Denoting by $d_{K,\sigma}$ the Euclidean distance between $\boldsymbol{x}_K$ and the hyperplane including σ, one assumes that $d_{K,\sigma} > 0$. We then denote by $\mathscr{D}_{K,\sigma}$ the cone with vertex $\boldsymbol{x}_K$ and basis σ.

The time discretization is performed with a constant time step $k = \dfrac{T}{N+1}$, where $N \in \mathbb{N}^\star$, and we shall denote $t_n = nk$, for $n \in [\![0, N+1]\!]$. We denote by ∂^1 the discrete first time derivative given by $\partial^1 v^{j+1} = \dfrac{v^{j+1} - v^j}{k}$.

Throughout this paper, the letter C stands for a positive constant independent of the parameters of the space and time discretizations.

We use the finite volume space considered in [7, Definition 5.1, p. 2037], that is the space $\mathscr{H}_\mathscr{D} \subset L^2(\Omega)$ of functions which are constant on each control volume K of $\mathscr{M}$. We associate any $\sigma \in \mathscr{E}_{\rm int}$ with a family of real numbers $\left(\beta_\sigma^K\right)_{K\in\mathscr{M}}$ (this family contains in general at most $d+1$ nonzero elements) such that

$$1 = \sum_{K\in\mathscr{M}} \beta_\sigma^K \qquad \text{and} \qquad \boldsymbol{x}_\sigma = \sum_{K\in\mathscr{M}} \beta_\sigma^K \boldsymbol{x}_K. \tag{5}$$

Then, for any $u \in \mathscr{H}_\mathscr{D}$, we set $u_\sigma = \displaystyle\sum_{K\in\mathscr{M}} \beta_\sigma^K\, u_K$, for all $\sigma \in \mathscr{E}_{\rm int}$ and $u_\sigma = 0$, for all $\sigma \in \mathscr{E}_{\rm ext}$.

In order to analyze the convergence, we need to consider the size of the discretization $\mathscr{D}$ defined by $h_\mathscr{D} = \sup\{\text{diam}(K),\ K \in \mathscr{M}\}$ and the regularity of the mesh given by (see [8, (4.1)–(4.2), p. 1025])

$$\theta_\mathscr{D} = \max\left(\max_{\sigma\in\mathscr{E}_{\rm int}, K, L\in\mathscr{M}} \frac{d_{K,\sigma}}{d_{L,\sigma}},\ \max_{K\in\mathscr{M}, \sigma\in\mathscr{E}_K} \frac{h_K}{d_{K,\sigma}},\ \max_{K\in\mathscr{M}, \sigma\in\mathscr{E}_K\cap\mathscr{E}_{\rm int}} \frac{\sum_{L\in\mathscr{M}} |\beta_\sigma^L|\, |\boldsymbol{x}_\sigma - \boldsymbol{x}_L|^2}{h_K^2} \right). \tag{6}$$

The scheme we present uses the discrete gradient of [8] which is given by: For $u \in \mathscr{H}_\mathscr{D}$ and for $K \in \mathscr{M}$

$$\nabla_{\mathscr{D}} u(\boldsymbol{x}) = \nabla_K u + \left(\frac{\sqrt{d}}{d_{K,\sigma}} \left(u_\sigma - u_K - \nabla_K u \cdot (\boldsymbol{x}_\sigma - \boldsymbol{x}_K)\right) \right) \mathbf{n}_{K,\sigma}, \quad \text{a.e. } \boldsymbol{x} \in \mathscr{D}_{K,\sigma}, \tag{7}$$

where $\nabla_K u = \dfrac{1}{\mathrm{m}(K)} \displaystyle\sum_{\sigma \in \mathscr{E}_K} \mathrm{m}(\sigma)\, (u_\sigma - u_K)\, \mathbf{n}_{K,\sigma}$.

3 Formulation of a Finite Volume Scheme and Statement of Its Known Convergence Results

Taking $t = t_{n+1}$ in (1) yields $\partial_t^\alpha u(t_{n+1}) - \Delta u(t_{n+1}) = f(t_{n+1})$. A consistent approximation for $\partial_t^\alpha u(t_{n+1})$ can be defined as a linear combination of the discrete time derivatives $\{\partial^1 u(t_{j+1}),\ j \in [\![0, n]\!]\}$ and it is given by (see [9, (4.1)–(4.2), p. 836], [3], and references therein):

$$\partial_t^\alpha u(t_{n+1}) = \sum_{j=0}^{n} k \lambda_j^{n+1} \partial^1 u(t_{j+1}) + \mathbb{T}_1^{n+1}(u), \tag{8}$$

where

$$\lambda_j^{n+1} = \frac{(n-j+1)^{1-\alpha} - (n-j)^{1-\alpha}}{k^\alpha \Gamma(2-\alpha)} \quad \text{and} \quad |\mathbb{T}_1^{n+1}(u)| \le C k^{2-\alpha} \|u\|_{\mathscr{C}^2([0,T])}. \tag{9}$$

The following lemma summarizes some properties of the coefficients λ_j^{n+1} given by (9).

Lemma 1 (Properties of the coefficients λ_j^{n+1}, cf. [9]) *For any $n \in [\![0, N]\!]$ and for any $j \in [\![0, n]\!]$, let λ_j^{n+1} be defined by (9). The following properties hold:*

$$\frac{k^{-\alpha}}{\Gamma(2-\alpha)} = \lambda_n^{n+1} > \cdots > \lambda_0^{n+1} \ge \lambda_0 = \frac{T^{-\alpha}}{\Gamma(1-\alpha)} \quad \text{and} \quad \sum_{j=0}^{n} k \lambda_j^{n+1} \le \frac{T^{1-\alpha}}{\Gamma(2-\alpha)}. \tag{10}$$

We set now a finite volume scheme which is a slight modification for the one given in [4]: the finite volume space considered in [4] is larger than the one used here, that is $\mathscr{H}_{\mathscr{D}}$. Indeed, the finite volume space of [4] includes in addition to the unknowns on the centers it contains also unknowns on edges of the control volumes.

Definition 2 (*Formulation of a finite volume scheme for* (1)–(4)) Let Ω be a polyhedral open bounded subset of $\mathbb{R}^d$, where $d \in \mathbb{N} \setminus \{0\}$, and $\partial\Omega = \overline{\Omega} \setminus \Omega$ its boundary. Let $\mathscr{D} = (\mathscr{M}, \mathscr{E}, \mathscr{P})$ be a discretization in the sense of Definition 1. For any $n \in [\![0, N]\!]$ and for any $j \in [\![0, n]\!]$, let λ_j^{n+1} be defined by (9). We define the following finite volume scheme as approximation for (1)–(4):

1. **Approximation of initial conditions** (3). The discretization of initial conditions (3) can be performed as: Find $u^0_{\mathscr{D}} \in \mathscr{H}_{\mathscr{D}}$ such that for all $v \in \mathscr{H}_{\mathscr{D}}$

$$\left(\nabla_{\mathscr{D}} u^0_{\mathscr{D}}, \nabla_{\mathscr{D}} v\right)_{L^2(\Omega)} = -\left(\Delta u^0, v\right)_{L^2(\Omega)}. \tag{11}$$

2. **Approximation of (1) and (4).** For any $n \in [\![0, N]\!]$, find $u^{n+1}_{\mathscr{D}} \in \mathscr{H}_{\mathscr{D}}$ such that, for all $v \in \mathscr{H}_{\mathscr{D}}$

$$\sum_{j=0}^{n} k\lambda_j^{n+1}\left(\partial^1 u^{j+1}_{\mathscr{D}}, v\right)_{L^2(\Omega)} + \left(\nabla_{\mathscr{D}} u^{n+1}_{\mathscr{D}}, \nabla_{\mathscr{D}} v\right)_{L^2(\Omega)} = (f(t_{n+1}), v)_{L^2(\Omega)}. \tag{12}$$

The proof of the $L^\infty(L^2)$–error estimate of [4] can be modified slightly to get the following $L^\infty(L^2)$–error estimate:

$$\max_{n=0}^{n=N+1} \|u(t_n) - u^n_{\mathscr{D}}\|_{L^2(\Omega)} \le C(k^{2-\alpha} + h_{\mathscr{D}})\|u\|_{\mathscr{C}^2([0,T];\,\mathscr{C}^2(\overline{\Omega}))}. \tag{13}$$

It is also proved (recall that the scheme (11)–(12) can be viewed as a particular of the one considered in [1] for **TFDE**) in [1] that the following $L^2(H^1)$–error estimate holds:

$$\left(\sum_{n=0}^{N+1} k\|\nabla u(t_n) - \nabla_{\mathscr{D}} u^n_{\mathscr{D}}\|^2_{L^2(\Omega)}\right)^{\frac{1}{2}} \le C(k^{2-\alpha} + h_{\mathscr{D}})\|u\|_{\mathscr{C}^2([0,T];\,\mathscr{C}^2(\overline{\Omega}))}. \tag{14}$$

4 The Main Results: $L^\infty(H^1)$–a Priori Estimate and $L^\infty(H^1)$–error Estimate

Our aim in this contribution is to prove that the error estimate (14) is uniform in time which is confirmed numerically in [1]. We first set this new error estimate in the following theorem:

Theorem 1 (New $L^\infty(H^1)$–error estimate for the scheme (11)–(12)) *Let $\alpha \in (0, 1)$ be given. Let Ω be a polyhedral open bounded subset of $\mathbb{R}^d$, where $d \in \mathbb{N} \setminus \{0\}$, and $\partial\Omega = \overline{\Omega} \setminus \Omega$ its boundary. Assume that the solution of (1)–(4) satisfies $u \in \mathscr{C}^2([0, T]; \mathscr{C}^2(\overline{\Omega}))$. Let $k = \dfrac{T}{N+1}$, where $N \in \mathbb{N} \setminus \{0\}$. We shall denote $t_n = nk$, for $n \in [\![0, N+1]\!]$. For any $n \in [\![0, N]\!]$ and for any $j \in [\![0, n]\!]$, let λ_j^{n+1} be defined by (9). Let $\mathscr{D} = (\mathscr{M}, \mathscr{E}, \mathscr{P})$ be a discretization in the sense of Definition 1. Assume that $\theta_{\mathscr{D}}$ satisfies $\theta \ge \theta_{\mathscr{D}}$.*

Then, there exists a unique solution $\left(u_{\mathscr{D}}^n\right)_{n=0}^{N+1} \in \mathscr{H}_{\mathscr{D}}^{N+2}$ *for scheme (11)–(12) and the following* $L^\infty(H^1)$*–error estimate holds:*

$$\max_{n=0}^{n=N+1} \|\nabla u(t_n) - \nabla_{\mathscr{D}} u_{\mathscr{D}}^n\|_{L^2(\Omega)} \le C(k^{2-\alpha} + h_{\mathscr{D}})\|u\|_{\mathscr{C}^2([0,T];\,\mathscr{C}^2(\overline{\Omega}))}. \tag{15}$$

To prove Theorem 1, we use the following definition of the discrete Laplace operator:

Definition 3 (*Discrete Laplace operator, cf.* [7, Definition 5.3, p. 2038]) Let Ω be a polyhedral open bounded subset of $\mathbb{R}^d$, where $d \in \mathbb{N} \setminus \{0\}$, and $\partial\Omega = \overline{\Omega} \setminus \Omega$ its boundary. Let $\mathscr{D} = (\mathscr{M}, \mathscr{E}, \mathscr{P})$ be a discretization in the sense of Definition 1. Let $\nabla_{\mathscr{D}}$ be the discrete gradient given by (7). Let $u \in \mathscr{H}_{\mathscr{D}}$, the discrete Laplace operator of u denoted by $\Delta_{\mathscr{D}} u$ is the element of $\mathscr{H}_{\mathscr{D}}$ given by

$$-\left(\Delta_{\mathscr{D}} u, v\right)_{L^2(\Omega)} = \left(\nabla_{\mathscr{D}} u, \nabla_{\mathscr{D}} v\right)_{L^2(\Omega)}, \qquad \forall v \in \mathscr{H}_{\mathscr{D}}. \tag{16}$$

To prove Theorem 1, we will also need to use the following new discrete $L^\infty(H^1)$–*a priori estimate*:

Lemma 2 (New $L^\infty(H^1)$–a priori estimate for the discrete problem) *Under the same hypotheses of Theorem 1, assume that there exists* $\left(\eta_{\mathscr{D}}^n\right)_{n=0}^{N+1} \in \mathscr{H}_{\mathscr{D}}^{N+2}$ *such that for all* $n \in [\![0, N]\!]$ *and for all* $v \in \mathscr{H}_{\mathscr{D}}$

$$\sum_{j=0}^{n} k\lambda_j^{n+1} \left(\partial^1 \eta_{\mathscr{D}}^{j+1}, v\right)_{L^2(\Omega)} + \left(\nabla_{\mathscr{D}} \eta_{\mathscr{D}}^{n+1}, \nabla_{\mathscr{D}} v\right)_{L^2(\Omega)} = \left(\mathscr{S}^{n+1}, v\right)_{L^2(\Omega)}, \tag{17}$$

where $\mathscr{S}^{n+1} \in L^2(\Omega)$*, for all* $n \in [\![0, N]\!]$*, and* $\eta_{\mathscr{D}}^0 = 0$.

Then, the following $L^\infty(H^1)$*–a priori estimate holds:*

$$\max_{n=0}^{N+1} \|\nabla_{\mathscr{D}} \eta_{\mathscr{D}}^n\|_{L^2(\Omega)} \le \frac{\mathscr{S}}{\sqrt{\lambda_0}}, \tag{18}$$

where $\mathscr{S} = \max_{n=0}^{N} \|\mathscr{S}^{n+1}\|_{L^2(\Omega)}$ *and* λ_0 *is the lower bound which appears in (10).*

Proof Using the definition (16) of the discrete Laplace operator, the hypothesis (17) can be written as

$$\sum_{j=0}^{n} \lambda_j^{n+1} \left(\eta_{\mathscr{D}}^{j+1} - \eta_{\mathscr{D}}^{j}, v\right)_{L^2(\Omega)} - \left(\Delta_{\mathscr{D}} \eta_{\mathscr{D}}^{n+1}, v\right)_{L^2(\Omega)} = \left(\mathscr{S}^{n+1}, v\right)_{L^2(\Omega)}. \tag{19}$$

Taking $v = -\Delta_{\mathscr{D}} \eta_{\mathscr{D}}^{n+1}$ in (19) and using again the definition (16) imply that

$$\sum_{j=0}^{n} \lambda_j^{n+1} \left(\nabla_{\mathscr{D}} (\eta_{\mathscr{D}}^{j+1} - \eta_{\mathscr{D}}^{j}), \nabla_{\mathscr{D}} \eta_{\mathscr{D}}^{n+1}\right)_{L^2(\Omega)} + \left\| \Delta_{\mathscr{D}} \eta_{\mathscr{D}}^{n+1} \right\|_{L^2(\Omega)}^2 = -\left(\mathscr{S}^{n+1}, \Delta_{\mathscr{D}} \eta_{\mathscr{D}}^{n+1}\right)_{L^2(\Omega)}. \tag{20}$$

Using the Cauchy Schwarz inequality together with inequality $xy \le x^2/2 + y^2/2$, (20) implies that

$$\sum_{j=0}^{n} \lambda_j^{n+1} \left(\nabla_{\mathscr{D}} (\eta_{\mathscr{D}}^{j+1} - \eta_{\mathscr{D}}^{j}), \nabla_{\mathscr{D}} \eta_{\mathscr{D}}^{n+1}\right)_{L^2(\Omega)} + \frac{1}{2} \left\| \Delta_{\mathscr{D}} \eta_{\mathscr{D}}^{n+1} \right\|_{L^2(\Omega)}^2 \le \frac{(\mathscr{S})^2}{2}. \tag{21}$$

Re-ordering the sum, inequality (21) gives

$$\lambda_n^{n+1} \|\nabla_{\mathscr{D}} \eta_{\mathscr{D}}^{n+1}\|_{L^2(\Omega)}^2 + \frac{1}{2} \left\| \Delta_{\mathscr{D}} \eta_{\mathscr{D}}^{n+1} \right\|_{L^2(\Omega)}^2 \le \sum_{j=1}^{n} (\lambda_j^{n+1} - \lambda_{j-1}^{n+1}) \left(\nabla_{\mathscr{D}} \eta_{\mathscr{D}}^{j}, \nabla_{\mathscr{D}} \eta_{\mathscr{D}}^{n+1}\right)_{L^2(\Omega)} + \frac{(\mathscr{S})^2}{2}. \tag{22}$$

Using again the Cauchy Schwarz inequality together with inequality $xy \le x^2/2 + y^2/2$ and the property (10) (which implies that $\lambda_j^{n+1} - \lambda_{j-1}^{n+1} > 0$), inequality (22) leads to

$$\|\nabla_{\mathscr{D}} \eta_{\mathscr{D}}^{n+1}\|_{L^2(\Omega)}^2 \le \frac{1}{\lambda_n^{n+1}} \left(\sum_{j=1}^{n} (\lambda_j^{n+1} - \lambda_{j-1}^{n+1}) \|\nabla_{\mathscr{D}} \eta_{\mathscr{D}}^{j}\|_{L^2(\Omega)}^2 + (\mathscr{S})^2 \right). \tag{23}$$

We prove by mathematical induction on n that, for all $n \in [\![1, N+1]\!]$

$$\|\nabla_{\mathscr{D}} \eta_{\mathscr{D}}^{n}\|_{L^2(\Omega)}^2 \le \frac{(\mathscr{S})^2}{\lambda_0}. \tag{24}$$

Taking $n = 0$ in (23) and using (10) yield (24) for $n = 1$. Assume now that (24) holds for $n \le m$ and prove it for $n = m + 1$. Taking $n = m$ in (23) and using the fact that $\lambda_0^{m+1} > \lambda_0$ (see (10)) implies that

$$\|\nabla_{\mathscr{D}} \eta_{\mathscr{D}}^{m+1}\|_{L^2(\Omega)}^2 \le \frac{(\mathscr{S})^2}{\lambda_m^{m+1}} \left(\frac{\lambda_m^{m+1} - \lambda_0^{m+1}}{\lambda_0} + 1 \right) \le \frac{(\mathscr{S})^2}{\lambda_m^{m+1}} \left(\frac{\lambda_m^{m+1} - \lambda_0}{\lambda_0} + 1 \right) = \frac{(\mathscr{S})^2}{\lambda_0}.$$

This completes the proof of Lemma 2. □

***Proof Sketch for Theorem* 1**

1. **Existence and uniqueness for scheme** (11)–(12). Are given for instance in [1].
2. **Proof of estimate** (15). To prove (15), we compare (11)–(12) with the following auxiliary scheme: For any $n \in [\![0, N+1]\!]$, find $\bar{u}_{\mathscr{D}}^n \in \mathscr{H}_{\mathscr{D}}$ such that

$$\left(\nabla_{\mathscr{D}} \bar{u}_{\mathscr{D}}^n, \nabla_{\mathscr{D}} v\right)_{L^2(\Omega)} = (-\Delta u(t_n), v)_{L^2(\Omega)}, \qquad \forall v \in \mathscr{H}_{\mathscr{D}}. \tag{25}$$

- Comparison between the solution of (25) and the solution of problem (1)–(4). The following convergence results hold, see [5, 8]:

$$\max_{n=0}^{n=N+1} \|\nabla u(t_n) - \nabla_{\mathcal{D}} \bar{u}_{\mathcal{D}}^n\|_{L^2(\Omega)} + \max_{n=1}^{n=N+1} \|\partial^1 (u(t_n) - \bar{u}_{\mathcal{D}}^n)\|_{L^2(\Omega)} \le C h_{\mathcal{D}} \|u\|_{\mathcal{C}^1([0,T];\, \mathcal{C}^2(\overline{\Omega}))}. \tag{26}$$

- Comparison between the solution of (11)–(12) and the auxiliary scheme (25). Let us define the *auxiliary* error $\eta_{\mathcal{D}}^n = u_{\mathcal{D}}^n - \bar{u}_{\mathcal{D}}^n$. Comparing (25) with scheme (11) and using the fact that $u(0) = u^0$ (subject of (3)) imply that $\eta_{\mathcal{D}}^0 = 0$. Writing now scheme (25) in the level $n+1$ and subtracting the result from (12) to get

$$\sum_{j=0}^{n} k\lambda_j^{n+1} \left(\partial^1 u_{\mathcal{D}}^{j+1}, v\right)_{L^2(\Omega)} + \left(\nabla_{\mathcal{D}} \eta_{\mathcal{D}}^{n+1}, \nabla_{\mathcal{D}} v\right)_{L^2(\Omega)} = (f(t_{n+1}) + \Delta u(t_{n+1}), v)_{L^2(\Omega)}. \tag{27}$$

Subtracting $\sum_{j=0}^{n} k\lambda_j^{n+1} \left(\partial^1 \bar{u}_{\mathcal{D}}^{j+1}, v\right)_{L^2(\Omega)}$ from both sides of (27), replacing $f(t_{n+1}) + \Delta u(t_{n+1})$ by $\partial_t^\alpha u(t_{n+1})$ (which stems from (1)), and using (8) we find that $\left(\eta_{\mathcal{D}}^n\right)_{n=0}^{N+1} \in (\mathcal{H}_{\mathcal{D}})^{N+2}$ is satisfying the hypothesis (17) of Lemma 2 with $\mathcal{S}^{n+1} = \sum_{j=0}^{n} k\lambda_j^{n+1} \partial^1 \left(u(t_{j+1}) - \bar{u}_{\mathcal{D}}^{j+1}\right) + \mathbb{T}_1^{n+1}(u)$. We are able then to apply the discrete *a priori estimate* (18) of Lemma 2 (recall that $\eta_{\mathcal{D}}^0 = 0$) to get $\max_{n=0}^{N+1} \|\nabla_{\mathcal{D}} \eta_{\mathcal{D}}^n\|_{L^2(\Omega)} \le \dfrac{\mathcal{S}}{\sqrt{\lambda_0}}$. Gathering now this with (10), error estimate (26), and (9) yields

$$\max_{n=0}^{N+1} \|\nabla_{\mathcal{D}} \eta_{\mathcal{D}}^n\|_{L^2(\Omega)} \le C(h_{\mathcal{D}} + k^{2-\alpha}) \|u\|_{\mathcal{C}^2([0,T];\, \mathcal{C}^2(\overline{\Omega}))}. \tag{28}$$

Gathering (28) with the fact that $\nabla u(t_n) - \nabla_{\mathcal{D}} u_{\mathcal{D}}^n = \nabla u(t_n) - \nabla_{\mathcal{D}} \bar{u}_{\mathcal{D}}^n - \nabla_{\mathcal{D}} \eta_{\mathcal{D}}^n$ and the error estimate (26) imply the desired estimate (15). This completes the proof of Theorem 1. □

Remark 1 (Both Lemma 2 and Theorem 1 are not straightforward) The *a priori estimate* of Lemma 2 is not straightforward in the sense that the usual techniques do not lead to the optimal estimate (18) (which is an unconditional estimate) in the $L^\infty(H^1)$–norm. The estimate (18) led to prove the optimal convergence order (15). Indeed, if we take for instance the obvious choice $v = \eta_{\mathcal{D}}^{n+1}$ in (17) instead of the choice $v = -\Delta_{\mathcal{D}} \eta_{\mathcal{D}}^{n+1}$ made in the proof of Lemma 2 and use [3, (27)–(28), p. 443], we get

$$\|\nabla_{\mathcal{D}} \eta_{\mathcal{D}}^{n+1}\|_{L^2(\Omega)^d}^2 \le \sum_{j=1}^{n} (\lambda_j^{n+1} - \lambda_{j-1}^{n+1}) \|\Pi_{\mathcal{M}} \eta^j\|_{L^2(\Omega)}^2 + C\,(\mathcal{S})^2 \le \frac{C\,(\mathcal{S})^2}{\lambda_0} \lambda_n^{n+1}. \tag{29}$$

This gives, using the fact that λ_n^{n+1} is of order $k^{-\alpha}$, $\|\nabla_{\mathscr{D}}\eta_{\mathscr{D}}^{n+1}\|_{L^2(\Omega)^d} \le Ck^{-\frac{\alpha}{2}}\mathscr{S}$ which is less accurate then that of (18). In fact estimate $\|\nabla_{\mathscr{D}}\eta_{\mathscr{D}}^{n+1}\|_{L^2(\Omega)} \le Ck^{-\frac{\alpha}{2}}\mathscr{S}$ is a conditional convergence. Such conditional convergence can be found for instance in the stability result [10, Theorem 3.1, p. 1540] where the energy $\|\cdot\|_1$ is given by [10, (3.14), p. 1539], that is $\|\cdot\|_1^2 = \|\cdot\|_{L^2}^2 + k^\alpha\Gamma(2-\alpha)\|\nabla\cdot\|_{L^2}^2$ and in the error estimate [10, (3.17), Theorem 3.2, p. 1540]. In the context of discontinuous Galerkin methods, we find similar conditional convergence results in [12]. The subject of error estimate in energy norm has not attracted the attention it merits yet, see [1, 9, 11].

5 Conclusion and Perspectives

We considered a cell-centered finite volume scheme (see [8, Sect. 3.2, p. 1022]) on the general mesh introduced in [8] to approximate **TFDE** in any space dimension. The formulation of this scheme involves the discrete gradient developed in [8]. One of the main features of this discrete gradient is that it is stable and consistent. The discretization in time is uniform. The scheme is a slight modification of the one of [4] which can be viewed as a pure hybrid scheme. We proved a new $L^\infty(H^1)$– a priori estimate. The proof of this result is not straightforward and it uses the discrete Laplace operator as a technical tool. The $L^\infty(H^1)$– *a priori estimate* allowed to prove an optimal convergence rate in the discrete norm of $L^\infty(H^1)$. Both the $L^\infty(H^1)$– a priori estimate and $L^\infty(H^1)$– error estimate improve our previous results of [4] in which we only proved $L^\infty(L^2)$ error estimate and also improve the results of [1] which dealt with $L^2(H^1)$– *a priori estimate* and $L^2(H^1)$– error estimate. We plan to extend the results of Lemma 2 to the general case $\eta_{\mathscr{D}}^0 \neq 0$. We plan also to extend this contribution to schemes with parameters (in particular schemes which use the Crank-Nicolson method) which use the general framework GDM [6] as discretization in space and we allow a large class of discretizations in time (not only the uniform case).

References

1. Bradji, A.: A new analysis for the convergence of the gradient discretization method for multidimensional time fractional diffusion and diffusion-wave equations. Comput. Math. Appl. **79**(2), 500–520 (2020)
2. Bradji, A.: A second order time accurate SUSHI method for the time-fractional diffusion equation. In: Nikolov, G. et al. (ed.) Numerical Methods and Applications. 9th International Conference, NMA 2018, Borovets, Bulgaria, August 20–24, 2018. Revised Selected Papers. Lecture Notes in Computer Science, vol. 11189, pp. 197-206. Cham: Springer (2019)
3. Bradji, A.: Notes on the convergence order of gradient schemes for time fractional differential equations. C. R. Math. Acad. Sci. Paris **356**(4), 439–448 (2018)
4. Bradji, A., Fuhrmann, J.: Convergence order of a finite volume scheme for the time-fractional diffusion equation. In: Numerical Analysis and Its Applications. Lecture Notes in Computer Science, vol. 10187, pp. 33–45. Springer, Cham (2017)

5. Bradji, A., Fuhrmann, J.: Some abstract error estimates of a finite volume scheme for a nonstationary heat equation on general nonconforming multidimensional spatial meshes. Appl. Math. **58**(1), 1–38 (2013)
6. Droniou, J., Eymard, R., Gallouët, T., Guichard, C., Herbin, R.: The Gradient Discretisation Method. Mathématiques et Applications, vol. 82. Springer Nature Switzerland AG, Basel, Switzerland (2018)
7. Eymard, R., Gallouët, T., Herbin, R., Linke, A.: Finite volume schemes for the biharmonic problem on general meshes. Math. Comput. **81**(280), 2019–2048 (2012)
8. Eymard, R., Gallouët, T., Herbin, R.: Discretization of heterogeneous and anisotropic diffusion problems on general nonconforming meshes. IMA J. Numer. Anal. **30**(4), 1009–1043 (2010)
9. Jin, B., Lazarov, R., Liu, Y., Zhou, Z.: The Galerkin finite element method for a multi-term time-fractional diffusion equation. J. Comput. Phys. **281**, 825–843 (2015)
10. Lin, Y., Xu, C.: Finite difference/spectral approximations for the time-fractional diffusion equation. J. Comput. Phys. **225**, 1533–1552 (2007)
11. Sidi Ammi, M.R., Jamiai, I., Torres, D.F.M.: A finite element approximation for a class of Caputo time-fractional diffusion equations. Comput. Math. Appl. **78**(5), 1334–1344 (2019)
12. Xu, Q., Zheng, Z.: Discontinuous Galerkin method for time fractional diffusion equation. J. Inf. Comput. Sci. **10**(11), 3253–3264 (2013)

Note on the Convergence of a Finite Volume Scheme for a Second Order Hyperbolic Equation with a Time Delay in Any Space Dimension

Fayssal Benkhaldoun and Abdallah Bradji

Abstract In this note, we establish a finite volume scheme for a model of a second order hyperbolic equation with a time delay in any space dimension. This model is considered in [10, 11] where some exponential stability estimates and oscillatory behaviour are proved. The scheme we shall present uses, as space discretization, the general class of nonconforming finite volume meshes of [5]. In addition to the proof of the existence and uniqueness of the discrete solution, we develop a new discrete a priori *estimate*. Thanks to this a priori *estimate*, we prove error estimates in discrete seminorms of $L^\infty(H_0^1)$, $L^\infty(L^2)$, and $W^{1,\infty}(L^2)$. This work can be viewed as extension to the previous ones [2, 4] which dealt with the analysis of finite volume methods for respectively semilinear parabolic equations with a time delay and the wave equation.

Keywords Hyperbolic delay equation · SUSHI · Discrete a priori *estimate* · Convergence

MSC 2010 65M08 · 65M12 · 65M15 · 35L20

1 Problem to Be Solved and Motivation

We consider the following second order hyperbolic equation with a time delay (see [10, p. 1563] and [11, 12]):

F. Benkhaldoun
LAGA, University of Paris 13, Paris, France
e-mail: fayssal@math.univ-paris13.fr
URL: https://www.math.univ-paris13.fr/~fayssal/

A. Bradji (✉)
LMA Laboratory, University of Annaba, Annaba, Algeria
e-mail: abdallah.bradji@gmail.com; abdallah.bradji@etu.univ-amu.fr
URL: https://www.i2m.univ-amu.fr/perso/abdallah.bradji/

R. Klöfkorn et al. (eds.), *Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples*, Springer Proceedings in Mathematics & Statistics 323,
https://doi.org/10.1007/978-3-030-43651-3_28

$$u_{tt}(\boldsymbol{x},t) - \Delta u(\boldsymbol{x},t) + \alpha u_t(\boldsymbol{x},t) + \beta u_t(\boldsymbol{x},t-\tau) = f(\boldsymbol{x},t), \qquad (\boldsymbol{x},t) \in \Omega \times (0,T), \quad (1)$$

where Ω is an open polygonal bounded subset in $\mathbb{R}^d$, f is a given function defined on $\Omega \times (0,T)$, and $T > 0$, $\alpha \geq 0$, $\beta \geq 0$, and $\tau > 0$ (the time delay) are given. Initial conditions are given by, for given functions u^0 and u^1 defined respectively on Ω and $\Omega \times (-\tau, 0)$

$$u(\boldsymbol{x},0) = u^0(\boldsymbol{x}), \quad u_t(\boldsymbol{x},t) = u^1(\boldsymbol{x},t), \qquad \boldsymbol{x} \in \Omega, \quad -\tau \leq t \leq 0. \qquad (2)$$

Homogeneous Dirichlet boundary conditions are given by

$$u(\boldsymbol{x},t) = 0, \qquad (\boldsymbol{x},t) \in \partial\Omega \times (0,T). \qquad (3)$$

Delay differential equations occur in several applications such as ecology, biology, medicine, see [1, 8] and references therein. We also refer to [12] where we find an explanation for the delay equations. However, the numerical methods which are carried out with Partial (or Ordinary) Differential Equations are not enough to deal with Delay Partial Differential Equations, cf. [1, pp. 9–19]. In addition, numerical methods for the delay equations are well developed for the case of Ordinary Differential Equations but the subject of numerical analysis for Delay Partial Differential Equations has not attracted the attention it merits yet, see [1, 13]. In this note, we consider a finite volume scheme, based on the uses of SUSHI [5] (Scheme Using Stabilization and Hybrid Interfaces), for the model (1)–(3) in any space dimension. Such model is considered for instance in [10, (1.12)–(1.16), p. 1563] where some exponential stability estimates are proved. Equation (1) generalizes then the wave equation treated in our previous work [4] and differs in the two terms: $u_t(t)$ and the delay term $u_t(t-\tau)$. The time stepping of the scheme we shall present is Euler implicit. We prove the existence and unique along with a full convergence analysis for the scheme. The convergence analysis is performed thanks to a well developed discrete a priori *estimate*. One of the main features of SUSHI is that the control volumes can only be assumed to be polyhedral. In addition to this, the formulation of SUSHI involves a consistent and stable discrete gradient.

2 Space and Time Discretizations and Some Preliminaries

Definition 1 (*Space discretization, cf.* [5]) Let Ω be a polyhedral open bounded subset of $\mathbb{R}^d$, where $d \in \mathbb{N} \setminus \{0\}$, and $\partial\Omega = \overline{\Omega} \setminus \Omega$ its boundary. A discretization of Ω, denoted by $\mathscr{D}$, is defined as the triplet $\mathscr{D} = (\mathscr{M}, \mathscr{E}, \mathscr{P})$, where:

1. $\mathscr{M}$ is a finite family of non empty connected open disjoint subsets of Ω (the "control volumes") such that $\overline{\Omega} = \cup_{K \in \mathscr{M}} \overline{K}$. For any $K \in \mathscr{M}$, let $\partial K = \overline{K} \setminus K$ be the boundary of K; let $\mathrm{m}(K) > 0$ denote the measure of K and h_K denote the diameter of K.

2. $\mathscr{E}$ is a finite family of disjoint subsets of $\overline{\Omega}$ (the "edges" of the mesh), such that, for all $\sigma \in \mathscr{E}$, σ is a non empty open subset of a hyperplane of $\mathbb{R}^d$, whose $(d-1)$–dimensional measure is strictly positive. We also assume that, for all $K \in \mathscr{M}$, there exists a subset $\mathscr{E}_K$ of $\mathscr{E}$ such that $\partial K = \cup_{\sigma \in \mathscr{E}_K} \overline{\sigma}$. For any $\sigma \in \mathscr{E}$, we denote by $\mathscr{M}_\sigma = \{K, \sigma \in \mathscr{E}_K\}$. We then assume that, for any $\sigma \in \mathscr{E}$, either $\mathscr{M}_\sigma$ has exactly one element and then $\sigma \subset \partial \Omega$ (the set of these interfaces, called boundary interfaces, denoted by $\mathscr{E}_{\text{ext}}$) or $\mathscr{M}_\sigma$ has exactly two elements (the set of these interfaces, called interior interfaces, denoted by $\mathscr{E}_{\text{int}}$). For all $\sigma \in \mathscr{E}$, we denote by $\boldsymbol{x}_\sigma$ the barycentre of σ. For all $K \in \mathscr{M}$ and $\sigma \in \mathscr{E}$, we denote by $\mathbf{n}_{K,\sigma}$ the unit vector normal to σ outward to K.
3. $\mathscr{P}$ is a family of points of Ω indexed by $\mathscr{M}$, denoted by $\mathscr{P} = (\boldsymbol{x}_K)_{K \in \mathscr{M}}$, such that for all $K \in \mathscr{M}$, $\boldsymbol{x}_K \in K$ and K is assumed to be $\boldsymbol{x}_K$–star-shaped, which means that for all $\boldsymbol{x} \in K$, the property $[\boldsymbol{x}_K, \boldsymbol{x}] \subset K$ holds. Denoting by $d_{K,\sigma}$ the Euclidean distance between $\boldsymbol{x}_K$ and the hyperplane including σ, one assumes that $d_{K,\sigma} > 0$. We then denote by $\mathscr{D}_{K,\sigma}$ the cone with vertex $\boldsymbol{x}_K$ and basis σ.

The time discretization is performed with a constrained time step-size k such that $\frac{\tau}{k} \in \mathbb{N}$. We set then $k = \frac{\tau}{M}$, where $M \in \mathbb{N} \setminus \{0\}$. Denote by N the integer part of $\frac{T}{k}$, i.e. $N = \left[\frac{T}{k}\right]$. We shall denote $t_n = nk$, for $n \in [\![-M, N]\!]$. As particular cases $t_{-M} = -\tau$, $t_0 = 0$, and $t_N \leq T$. One of the advantages of this time discretization is that the point $t = 0$ is a mesh point which is suitable since we have Eq. (1) defined for $t \in (0, T)$ and the second initial condition in (2) is defined for $t \in (-\tau, 0)$. We denote by ∂^1 and ∂^2 the discrete first and second time derivatives given by $\partial^1 v^{j+1} = \frac{v^{j+1} - v^j}{k}$ and $\partial^2 v^{j+1} = \partial^1(\partial^1 v^{j+1})$.

Throughout this paper, the letter C stands for a positive constant independent of the parameters of discretizations.

We define the discrete space $\mathscr{X}_{\mathscr{D},0}$ as the set of all $v = \big((v_K)_{K \in \mathscr{M}}\,,\ (v_\sigma)_{\sigma \in \mathscr{E}}\big)$, where $v_K, v_\sigma \in \mathbb{R}$ and $v_\sigma = 0$ for all $\sigma \in \mathscr{E}_{\text{ext}}$. Let $H_{\mathscr{M}}(\Omega) \subset L^2(\Omega)$ be the space of functions which are constant on each control volume K of the mesh $\mathscr{M}$. For all $v \in \mathscr{X}_{\mathscr{D}}$, we denote by $\Pi_{\mathscr{M}} v \in H_{\mathscr{M}}(\Omega)$ the function defined by $\Pi_{\mathscr{M}} v(\boldsymbol{x}) = v_K$, for a.e. $\boldsymbol{x} \in K$, for all $K \in \mathscr{M}$. In order to analyze the convergence, we consider the size of the discretization $\mathscr{D}$ defined by $h_{\mathscr{D}} = \sup\{\text{diam}(K),\ K \in \mathscr{M}\}$ and the regularity of the mesh given by

$$\theta_{\mathscr{D}} = \max\left(\max_{\sigma \in \mathscr{E}_{\text{int}}, K, L \in \mathscr{M}} \frac{d_{K,\sigma}}{d_{L,\sigma}}, \max_{K \in \mathscr{M}, \sigma \in \mathscr{E}_K} \frac{h_K}{d_{K,\sigma}}\right). \tag{4}$$

The scheme we consider is based on the discrete gradient of [5]. For $u \in \mathscr{X}_{\mathscr{D}}$, we define, for all $K \in \mathscr{M}$

$$\nabla_{\mathscr{D}} u(\boldsymbol{x}) = \nabla_K u + \left(\frac{\sqrt{d}}{d_{K,\sigma}}\left(u_\sigma - u_K - \nabla_K u \cdot (\boldsymbol{x}_\sigma - \boldsymbol{x}_K)\right)\right)\mathbf{n}_{K,\sigma}, \quad \text{a.e. } \boldsymbol{x} \in \mathscr{D}_{K,\sigma}, \tag{5}$$

where $\nabla_K u = \dfrac{1}{\mathrm{m}(K)} \displaystyle\sum_{\sigma \in \mathscr{E}_K} \mathrm{m}(\sigma)\,(u_\sigma - u_K)\,\mathbf{n}_{K,\sigma}$. We define the bilinear form defined on $\mathscr{X}_{\mathscr{D}} \times \mathscr{X}_{\mathscr{D}}$ by $\langle u, v\rangle_F = \displaystyle\int_\Omega \nabla_{\mathscr{D}} u(\boldsymbol{x}) \cdot \nabla_{\mathscr{D}} v(\boldsymbol{x}) d\boldsymbol{x}$

3 Formulation of a New Finite Volume Scheme for the Delay Problem (1)–(3)

We now set a formulation of an implicit finite volume scheme for problem (1)–(3). The unknowns of this scheme are the set $\left\{u_{\mathscr{D}}^n;\, n \in [\![-M, N]\!]\right\}$ which are expected to approximate the set of the unknowns $\left\{u(t_n);\, n \in [\![-M, N]\!]\right\}$.

1. **Approximation of initial conditions** (2). The discretization of initial conditions (2) can be performed as: Find $u_{\mathscr{D}}^n$ for $n \in [\![-M, 0]\!]$ such that for all $v \in \mathscr{X}_{\mathscr{D},0}$

$$\langle u_{\mathscr{D}}^0, v\rangle_F = -\left(\Delta u^0, \Pi_{\mathscr{M}} v\right)_{L^2(\Omega)} \text{ and } \langle \partial^1 u_{\mathscr{D}}^n, v\rangle_F = -\left(\Delta u^1(t_n), \Pi_{\mathscr{M}} v\right)_{L^2(\Omega)},\ \forall n \in [\![-M+1, 0]\!]. \tag{6}$$

2. **Approximation of** (1) **and** (3). For any $n \in [\![0, N-1]\!]$, find $u_{\mathscr{D}}^{n+1} \in \mathscr{X}_{\mathscr{D},0}$ such that, for all $v \in \mathscr{X}_{\mathscr{D},0}$

$$\begin{aligned}&\left(\partial^2 \Pi_{\mathscr{M}} u_{\mathscr{D}}^{n+1}, \Pi_{\mathscr{M}} v\right)_{L^2(\Omega)} + \langle u_{\mathscr{D}}^{n+1}, v\rangle_F + \alpha\left(\partial^1 \Pi_{\mathscr{M}} u_{\mathscr{D}}^{n+1}, \Pi_{\mathscr{M}} v\right)_{L^2(\Omega)}\\ &+\beta\left(\partial^1 \Pi_{\mathscr{M}} u_{\mathscr{D}}^{n+1-M}, \Pi_{\mathscr{M}} v\right)_{L^2(\Omega)} = \left(f(t_{n+1}), \Pi_{\mathscr{M}} v\right)_{L^2(\Omega)}.\end{aligned} \tag{7}$$

4 Convergence Order of Scheme (6)–(7)

In addition to the new scheme (6)–(7), we present also its existence, uniqueness, and convergence order.

Theorem 1 (New error estimates for scheme *(6)–(7)*) *Let Ω be a polyhedral open bounded subset of $\mathbb{R}^d$, where $d \in \mathbb{N} \setminus \{0\}$, and $\partial\Omega = \overline{\Omega} \setminus \Omega$ its boundary. Assume that the solution of (1)–(3) satisfies $u \in \mathscr{C}^3([-\tau, T]; \mathscr{C}^2(\overline{\Omega}))$. Let $k = \dfrac{\tau}{M}$, where $M \in \mathbb{N} \setminus \{0\}$. Denote by N the integer part of $\dfrac{T}{k}$. We shall denote $t_n = nk$, for $n \in [\![-M, N]\!]$. As particular cases $t_{-M} = -\tau$ and $t_0 = 0$. Let $\mathscr{D} = (\mathscr{M}, \mathscr{E}, \mathscr{P})$ be a discretization in the sense of Definition 1. Assume that $\theta_{\mathscr{D}}$ (given by (4)) satisfies $\theta \geq \theta_{\mathscr{D}}$. Let $\nabla_{\mathscr{D}}$ be the discrete gradient given by (5). Then, there exists a unique solution $\left(u_{\mathscr{D}}^n\right)_{n=-M}^{N} \in \mathscr{X}_{\mathscr{D},0}^{M+N+1}$ for scheme (6)–(7) and the following error estimates hold:*

- *$L^\infty(L^2)$ and $L^\infty(H_0^1)$ error estimates.*

$$\max_{n=0}^{n=N}\|u(t_n)-\Pi_{\mathscr{M}}\,u_{\mathscr{D}}^n\|_{L^2(\Omega)}+\max_{n=0}^{n=N}\|\nabla u(t_n)-\nabla_{\mathscr{D}}\,u_{\mathscr{D}}^n\|_{L^2(\Omega)}\le C(k+h_{\mathscr{D}})\|u\|_{\mathscr{C}^3([0,T];\,\mathscr{C}^2(\overline{\Omega}))}. \tag{8}$$

- *$W^{1,\infty}(L^2)$–estimate.*

$$\max_{n=-M+1}^{n=N}\left\|u_t(t_n)-\Pi_{\mathscr{M}}\partial^1 u_{\mathscr{D}}^n\right\|_{L^2(\Omega)}\le C(k+h_{\mathscr{D}})\|u\|_{\mathscr{C}^3([0,T];\,\mathscr{C}^2(\overline{\Omega}))}. \tag{9}$$

To prove Theorem 1, we need to use the following new discrete a priori *estimate*:

Lemma 1 (New a priori estimate for the discrete problem) *Under the same hypotheses of Theorem 1, assume that there exists $\left(\eta_{\mathscr{D}}^n\right)_{n=-M}^N\in\left(\mathscr{X}_{\mathscr{D},0}\right)^{M+N+1}$ such that for all $n\in[\![0,N-1]\!]$ and for all $v\in\mathscr{X}_{\mathscr{D},0}$*

$$\begin{aligned}&\left(\partial^2\Pi_{\mathscr{M}}\eta_{\mathscr{D}}^{n+1},\Pi_{\mathscr{M}}v\right)_{L^2(\Omega)}+\langle\eta_{\mathscr{D}}^{n+1},v\rangle_F+\alpha\left(\partial^1\Pi_{\mathscr{M}}\eta_{\mathscr{D}}^{n+1},\Pi_{\mathscr{M}}v\right)_{L^2(\Omega)}\\&+\beta\left(\partial^1\Pi_{\mathscr{M}}\eta_{\mathscr{D}}^{n+1-M},\Pi_{\mathscr{M}}v\right)_{L^2(\Omega)}=\left(\mathscr{S}^{n+1},\Pi_{\mathscr{M}}v\right)_{L^2(\Omega)},\end{aligned} \tag{10}$$

where $\mathscr{S}^{n+1}\in L^2(\Omega)$, for all $n\in[\![0,N-1]\!]$. Then, the following estimate holds, for all $J\in[\![1,N]\!]$

$$\mathbb{E}_{\mathscr{D}}^J\le C\left((\mathscr{S})^2+\max_{n=1-M}^{0}\|\partial^1\Pi_{\mathscr{M}}\eta_{\mathscr{D}}^n\|_{L^2(\Omega)}^2+\|\nabla_{\mathscr{D}}\eta_{\mathscr{D}}^0\|_{L^2(\Omega)}^2\right), \tag{11}$$

where $\mathbb{E}_{\mathscr{D}}^J=\|\partial^1\Pi_{\mathscr{M}}\eta_{\mathscr{D}}^J\|_{L^2(\Omega)}^2+\|\nabla_{\mathscr{D}}\eta_{\mathscr{D}}^J\|_{L^2(\Omega)}^2$ and $\mathscr{S}=\max_{n=0}^{N-1}\|\mathscr{S}^{n+1}\|_{L^2(\Omega)}$.

Proof The following rules will be useful, for $\alpha_{\mathscr{D}}^n=\Pi_{\mathscr{M}}\,\partial^1\eta_{\mathscr{D}}^n$

$$\begin{aligned}\left(\Pi_{\mathscr{M}}\,\partial^2\,\eta_{\mathscr{D}}^{n+1},\Pi_{\mathscr{M}}\,\partial^1\,\eta_{\mathscr{D}}^{n+1}\right)_{L^2(\Omega)}=&\frac{1}{2k}\left(\alpha_{\mathscr{D}}^{n+1}-\alpha_{\mathscr{D}}^n,\alpha_{\mathscr{D}}^{n+1}-\alpha_{\mathscr{D}}^n\right)_{L^2(\Omega)}\\&+\frac{1}{2k}\left\{\left(\alpha_{\mathscr{D}}^{n+1},\alpha_{\mathscr{D}}^{n+1}\right)_{L^2(\Omega)}-\left(\alpha_{\mathscr{D}}^n,\alpha_{\mathscr{D}}^n\right)_{L^2(\Omega)}\right\}\end{aligned}$$

and $\langle\,\eta_{\mathscr{D}}^{n+1},\partial^1\,\eta_{\mathscr{D}}^{n+1}\rangle_F=\frac{1}{2k}\left\{\langle\,\eta_{\mathscr{D}}^{n+1}-\eta_{\mathscr{D}}^n,\eta_{\mathscr{D}}^{n+1}-\eta_{\mathscr{D}}^n\rangle_F+\langle\,\eta_{\mathscr{D}}^{n+1},\eta_{\mathscr{D}}^{n+1}\rangle_F-\langle\eta_{\mathscr{D}}^n,\eta_{\mathscr{D}}^n\rangle_F\right\}$.

Taking $v=\partial^1\eta_{\mathscr{D}}^{n+1}$ in (10) and using the previous two rules to get, for all $n\in[\![0,N-1]\!]$

$$\begin{aligned}&\|\partial^1\Pi_{\mathscr{M}}\,\eta_{\mathscr{D}}^{n+1}\|_{L^2(\Omega)}^2-\|\partial^1\Pi_{\mathscr{M}}\,\eta_{\mathscr{D}}^n\|_{L^2(\Omega)}^2+\langle\eta_{\mathscr{D}}^{n+1},\eta_{\mathscr{D}}^{n+1}\rangle_F-\langle\,\eta_{\mathscr{D}}^n,\eta_{\mathscr{D}}^n\rangle_F\\&\le 2k\left(\mathscr{S}^{n+1},\partial^1\Pi_{\mathscr{M}}\,\eta_{\mathscr{D}}^{n+1}\right)_{L^2(\Omega)}-2k\beta\left(\partial^1\Pi_{\mathscr{M}}\eta_{\mathscr{D}}^{n+1-M},\Pi_{\mathscr{M}}\partial^1\eta_{\mathscr{D}}^{n+1}\right)_{L^2(\Omega)}.\end{aligned}$$

Summing the previous inequality over $n\in[\![0,J-1]\!]$, where $J\in[\![1,N]\!]$, and using the Cauchy Schwarz inequality yield

$$\begin{aligned}\mathbb{E}_{\mathcal{D}}^J &\le 2\mathcal{S}\sum_{n=0}^{J-1} k\|\partial^1 \mathit{\Pi}_{\mathcal{M}}\, \eta_{\mathcal{D}}^{n+1}\|_{L^2(\Omega)} + 2\beta\sum_{n=0}^{J-1} k\|\partial^1\mathit{\Pi}_{\mathcal{M}}\,\eta_{\mathcal{D}}^{n+1-M}\|_{L^2(\Omega)}\|\mathit{\Pi}_{\mathcal{M}}\,\partial^1\eta_{\mathcal{D}}^{n+1}\|_{L^2(\Omega)}\\ &\quad + \|\partial^1\mathit{\Pi}_{\mathcal{M}}\,\eta_{\mathcal{D}}^0\|^2_{L^2(\Omega)} + \|\nabla_{\mathcal{D}}\,\eta_{\mathcal{D}}^0\|^2_{L^2(\Omega)}.\end{aligned} \tag{12}$$

Using twice inequality $ab \le \epsilon a^2 + b^2/\epsilon$, with $\epsilon = k/(4\tau)$, and the fact that $k/(2\tau) = 1/(2M) \le 1/2$, (12) implies that

$$\mathbb{E}_{\mathcal{D}}^J \le C\left(\sum_{n=1}^{J-1} k\|\partial^1\mathit{\Pi}_{\mathcal{M}}\,\eta_{\mathcal{D}}^n\|^2_{L^2(\Omega)} + (\mathcal{S})^2 + \max_{n=1-M}^{0}\|\partial^1\mathit{\Pi}_{\mathcal{M}}\,\eta_{\mathcal{D}}^n\|^2_{L^2(\Omega)} + \|\nabla_{\mathcal{D}}\,\eta_{\mathcal{D}}^0\|^2_{L^2(\Omega)}\right). \tag{13}$$

This implies in particular that

$$\|\partial^1\mathit{\Pi}_{\mathcal{M}}\,\eta_{\mathcal{D}}^J\|^2_{L^2(\Omega)} \le C\left(\sum_{n=1}^{J-1} k\|\partial^1\mathit{\Pi}_{\mathcal{M}}\,\eta_{\mathcal{D}}^n\|^2_{L^2(\Omega)} + (\mathcal{S})^2 + \max_{n=1-M}^{0}\|\partial^1\mathit{\Pi}_{\mathcal{M}}\,\eta_{\mathcal{D}}^n\|^2_{L^2(\Omega)} + \|\nabla_{\mathcal{D}}\,\eta_{\mathcal{D}}^0\|^2_{L^2(\Omega)}\right).$$

Applying a discrete version for the Gronwall lemma (see for instance [3, Lemma 4.7, p. 1303] and references therein) to this inequality yields, for all $J \in [\![1, N]\!]$

$$\|\partial^1\mathit{\Pi}_{\mathcal{M}}\eta_{\mathcal{D}}^J\|^2_{L^2(\Omega)} \le C\left((\mathcal{S})^2 + \max_{n=1-M}^{0}\|\partial^1\mathit{\Pi}_{\mathcal{M}}\eta_{\mathcal{D}}^n\|^2_{L^2(\Omega)} + \|\nabla_{\mathcal{D}}\eta_{\mathcal{D}}^0\|^2_{L^2(\Omega)}\right).$$

This with (13) lead to the the desired estimate (11). □

Proof Sketch for Theorem 1

1. **Existence and uniqueness for scheme** (6)–(7). The existence and uniqueness can be justified using the fact that schemes (6) and (7) are finite dimensional linear systems.
2. **Proof of estimates** (8)–(9). To prove (8)–(9), we compare (6)–(7) with the following auxiliary scheme: For any $n \in [\![-M, N]\!]$, find $\bar{u}_{\mathcal{D}}^n \in \mathcal{X}_{\mathcal{D},0}$ such that

$$\langle \bar{u}_{\mathcal{D}}^n, v\rangle_F = (-\Delta u(t_n), \mathit{\Pi}_{\mathcal{M}}\, v)_{L^2(\Omega)}, \quad \forall v \in \mathcal{X}_{\mathcal{D},0}. \tag{14}$$

– Comparison between the solutions of (14) and (1)–(3). The following convergence results hold, see [3–5]:

- Discrete $L^\infty(L^2)$ and $L^\infty(H^1)$–error estimates. For all $n \in [\![-M, N]\!]$

$$\|u(t_n) - \mathit{\Pi}_{\mathcal{M}}\bar{u}_{\mathcal{D}}^n\|_{L^2(\Omega)} + \|\nabla u(t_n) - \nabla_{\mathcal{D}}\bar{u}_{\mathcal{D}}^n\|_{\left(L^2(\Omega)\right)^d} \le Ch_{\mathcal{D}}\|u\|_{\mathcal{C}([0,T];\ \mathcal{C}^2(\overline{\Omega}))}. \tag{15}$$

- $\mathcal{W}^{j,\infty}(L^2)$–error estimate, for $j \in \{1, 2\}$.

$$\begin{aligned}\max_{n=-M+2}^{N}\|u_{tt}(t_n) - \partial^2\mathit{\Pi}_{\mathcal{M}}\bar{u}_{\mathcal{D}}^n\|_{L^2(\Omega)} + \max_{n=-M+1}^{N}\|u_t(t_n) - \partial^1\mathit{\Pi}_{\mathcal{M}}\bar{u}_{\mathcal{D}}^n\|_{L^2(\Omega)}\\ \le C(h_{\mathcal{D}} + k)\|u\|_{\mathcal{C}^3([0,T];\ \mathcal{C}^2(\overline{\Omega}))}.\end{aligned} \tag{16}$$

- Comparison between the schemes (6)–(7) and (14). Let us define the error $\eta^n_{\mathscr{D}} = u^n_{\mathscr{D}} - \bar{u}^n_{\mathscr{D}}$. Comparing (14) with the first scheme in (6) and using the fact that $u(0) = u^0$ (see (2)) imply that $\eta^0_{\mathscr{D}} = 0$. Writing scheme (14) in the level $n+1$ and subtracting the result from (7) to get, for all $n \in [\![0, N-1]\!]$ and for all $v \in \mathscr{X}_{\mathscr{D},0}$

$$\left(\partial^2 \Pi_{\mathscr{M}} u^{n+1}_{\mathscr{D}}, \Pi_{\mathscr{M}} v\right)_{L^2(\Omega)} + \langle \eta^{n+1}_{\mathscr{D}}, v\rangle_F + \alpha \left(\partial^1 \Pi_{\mathscr{M}} u^{n+1}_{\mathscr{D}}, \Pi_{\mathscr{M}} v\right)_{L^2(\Omega)}$$
$$+\beta \left(\partial^1 \Pi_{\mathscr{M}} u^{n+1-M}_{\mathscr{D}}, \Pi_{\mathscr{M}} v\right)_{L^2(\Omega)} = \left(f(t_{n+1}) + \Delta u(t_{n+1}), \Pi_{\mathscr{M}} v\right)_{L^2(\Omega)}.$$

Subtracting $\left(\partial^2 \Pi_{\mathscr{M}} \bar{u}^{n+1}_{\mathscr{D}} + \alpha \partial^1 \Pi_{\mathscr{M}} \bar{u}^{n+1}_{\mathscr{D}} + \beta \partial^1 \Pi_{\mathscr{M}} \bar{u}^{n+1-M}_{\mathscr{D}}, \Pi_{\mathscr{M}} v\right)_{L^2(\Omega)}$ from the both sides of the previous equation and replacing $f(t_{n+1}) + \Delta u(t_{n+1})$ by $u_{tt}(t_{n+1}) + \alpha u_t(t_{n+1}) + \beta u_t(t_{n+1-M})$ (which stems from (1)), we get, for all $v \in \mathscr{X}_{\mathscr{D},0}$

$$\left(\partial^2 \Pi_{\mathscr{M}} \eta^{n+1}_{\mathscr{D}}, \Pi_{\mathscr{M}} v\right)_{L^2(\Omega)} + \langle \eta^{n+1}_{\mathscr{D}}, v\rangle_F + \alpha \left(\partial^1 \Pi_{\mathscr{M}} \eta^{n+1}_{\mathscr{D}}, \Pi_{\mathscr{M}} v\right)_{L^2(\Omega)}$$
$$+\beta \left(\partial^1 \Pi_{\mathscr{M}} \eta^{n+1-M}_{\mathscr{D}}, \Pi_{\mathscr{M}} v\right)_{L^2(\Omega)} = \left(\mathscr{S}^{n+1}, \Pi_{\mathscr{M}} v\right)_{L^2(\Omega)}, \tag{17}$$

with $\mathscr{S}^n = u_{tt}(t_n) - \partial^2 \Pi_{\mathscr{M}} \bar{u}^n_{\mathscr{D}} + \alpha(u_t(t_n) - \partial^1 \Pi_{\mathscr{M}} \bar{u}^n_{\mathscr{D}}) + \beta(u_t(t_{n-M}) - \partial^1 \Pi_{\mathscr{M}} \bar{u}^{n-M}_{\mathscr{D}})$.

Since $\left(\eta^n_{\mathscr{D}}\right)_{n=-M}^N \in \left(\mathscr{X}_{\mathscr{D},0}\right)^{N+M+1}$ is satisfying (17), hence it satisfies the hypothesis (10) of Lemma 1. Applying now the discrete a priori *estimate* (11) of Lemma 1 and using the property $\eta^0_{\mathscr{D}} = 0$ and (16) together with the triangle inequality to get, for all $J \in [\![1, N]\!]$

$$\mathbb{E}^J_{\mathscr{D}} \le C(h_{\mathscr{D}} + k)^2 \|u\|^2_{\mathscr{C}^3([0,T];\, \mathscr{C}^2(\overline{\Omega}))} + \sum_{n=1-M}^{0} k \|\partial^1 \Pi_{\mathscr{M}} \eta^n_{\mathscr{D}}\|^2_{L^2(\Omega)} + \|\partial^1 \Pi_{\mathscr{M}} \eta^0_{\mathscr{D}}\|^2_{L^2(\Omega)}. \tag{18}$$

Let us estimate the terms $\|\partial^1 \Pi_{\mathscr{M}} \eta^n_{\mathscr{D}}\|_{L^2(\Omega)}$ involved in rhs (the right hand side) of (18). We have, for all $n \in [\![1-M, 0]\!]$, $\partial^1 \Pi_{\mathscr{M}} \eta^n_{\mathscr{D}} = \partial^1 \Pi_{\mathscr{M}} u^n_{\mathscr{D}} - u_t(t_n) + u_t(t_n) - \partial^1 \Pi_{\mathscr{M}} \bar{u}^n_{\mathscr{D}}$. This with the triangle inequality and estimate (16)

$$\|\partial^1 \Pi_{\mathscr{M}} \eta^n_{\mathscr{D}}\|_{L^2(\Omega)} \le \|\partial^1 \Pi_{\mathscr{M}} u^n_{\mathscr{D}} - u_t(t_n)\|_{L^2(\Omega)} + C(h_{\mathscr{D}} + k)\|u\|_{\mathscr{C}^2([0,T];\, \mathscr{C}^2(\overline{\Omega}))}. \tag{19}$$

Since, for all $n \in [\![1-M, 0]\!]$, $\partial^1 u^n_{\mathscr{D}}$ satisfies (see the second scheme in (6)) the same scheme (14) but with $u^1(t_n)$ in the rhs instead of $u(t_n)$, we are able then to apply estimates (15) on the second scheme in (6) to get, for all $n \in [\![1-M, 0]\!]$,

$$\|u^1(t_n) - \Pi_{\mathscr{M}}\, \partial^1 u^n_{\mathscr{D}}\|_{L^2(\Omega)} + \|\nabla u^1(t_n) - \nabla_{\mathscr{D}}\, \partial^1 u^n_{\mathscr{D}}\|_{\left(L^2(\Omega)\right)^d} \le C h_{\mathscr{D}} \|u\|_{\mathscr{C}^1([0,T];\, \mathscr{C}^2(\overline{\Omega}))}.$$

This with the fact that $u^1(t_n) = u_t(t_n)$, for all $n \in [\![1-M, 0]\!]$ (see (2)), imply that

$$\|u_t(t_n) - \mathit{\Pi}_{\mathscr{M}}\, \partial^1 u^n_{\mathscr{D}}\, \|_{L^2(\Omega)} + \|\nabla u_t(t_n) - \nabla_{\mathscr{D}}\, \partial^1 u^n_{\mathscr{D}}\, \|_{\left(L^2(\Omega)\right)^d} \le C h_{\mathscr{D}}\, \|\, u\|_{\mathscr{C}^1([0,T];\ \mathscr{C}^2(\overline{\Omega}))}. \tag{20}$$

Gathering this and (19) implies that $\max\limits_{n=1-M}^{n=0} \|\partial^1 \mathit{\Pi}_{\mathscr{M}}\, \eta^n_{\mathscr{D}}\|_{L^2(\Omega)} \le C(h_{\mathscr{D}} + k)$ $\|u\|_{\mathscr{C}^2([0,T];\ \mathscr{C}^2(\overline{\Omega}))}$. This with (18) yield, for all $J \in [\![1, N]\!]$

$$\|\partial^1 \mathit{\Pi}_{\mathscr{M}}\, \eta^J_{\mathscr{D}}\|^2_{L^2(\Omega)} + \|\nabla_{\mathscr{D}}\eta^J_{\mathscr{D}}\|^2_{L^2(\Omega)} \le C(h_{\mathscr{D}} + k)^2 \|u\|^2_{\mathscr{C}^3([0,T];\ \mathscr{C}^2(\overline{\Omega}))}. \tag{21}$$

Using now (15), (16), the discrete Poincaré inequality [5, Lemma 5.4], and (21) yield the desired estimates (8)–(9) when $n \in [\![1, N]\!]$. The case of (9) when $n \in [\![-M + 1, 0]\!]$ is a result of (20). The case when $n = 0$ in (8) can be deduced from the property $\eta^0_{\mathscr{D}} = 0$ and estimate (15). □

5 Some Numerical Tests

We consider $\Omega = (0, 1)^2$ meshed with the rectangular meshes described as in [7, p. 756–758], with uniform meshes with mesh size h, that is a mesh $\mathscr{D} = (\mathscr{M}, \mathscr{E}, \mathscr{P})$ given by

- $\mathscr{M}$ is a set of rectangles $\mathscr{M} = \{K_{ij} =](i-1)h, ih[\times](j-1)h, jh[, (i, j) \in [\![1, N]\!] \times [\![1, N]\!]\}$, where $N \in \mathbb{N} \setminus \{0\}$ is given and $h = 1/N$.
- $\mathscr{E}$ is the set of the edges of the elements K_{ij} of $\mathscr{M}$.
- The family $\mathscr{P}$ is the set of points $((i - \frac{1}{2})h, (j - \frac{1}{2})h)$, where $(i, j) \in [\![1, N]\!] \times [\![1, N]\!]$.

For the sake of simplicity, we will consider the discrete gradient described in [6, (211)–(212), p. 333]. The exact solution is given by $u(x, y, t) = \dfrac{1}{2\pi^2} \sin(\pi x)$ $\sin(\pi y) \cos(\pi\sqrt{2}t)$ for $(x, y, t) \in \Omega \times (-\dfrac{1}{\sqrt{2}}, 1)$. By this way u is satisfying (1) with $\alpha = \beta = 1$ and $\tau = \dfrac{1}{\sqrt{2}}$. We report the following results obtained using Scilab https://www.scilab.org/:

k		$\lvert \text{Error}\rvert_{L^\infty(H^1)}$ when $h = 1/40$	
		Error	Order
$\tau/100$		0.006393027957	–
$\tau/200$		0.003232991714	–
$\tau/400$		0.001627837907	0.977229560961
$\tau/800$		0.000820899933	0.992181855205

h		$\lvert \text{Error}\rvert_{L^\infty(H^1)}$ when $k = \tau/2000$	
		Error	Order
1/3		0.007766996341	–
1/6		0.002579934118	–
1/12		0.000848217611	1.582714905051
1/24		0.000405860137	1.968918167065

The results of the table in left show that the convergence order in time is one in $|\cdot|_{L^\infty(H^1)}$ which supports (8). The table in right shows that the numerical order in space is not only one as stated in (8) but it approaches two. This confirms the observation stated in [5, Lines 31–33, p. 1022] where the numerical tests show a better rate of convergence of the gradient in the case of uniform squares.

6 Conclusion and Perspectives

We established a new finite volume scheme for a model of a second hyperbolic equation with a time delay involved in the time derivative of the exact solution in any dimension. Such model is already considered in [10, p. 1563] where some exponential stability results are proved. A convergence analysis for the numerical scheme is carried out. One of the interesting paths to be followed in the future is to use the a priori estimate of Lemma 1 to prove a well posedness for the discrete problem and to prove the convergence of the family of the approximate solutions towards the unique solution of a weak formulation. This yields then the existence of the exact solution of a weak formulation for problem (1)–(3). We also plan to extend the results to the model of hyperbolic equation in which the delay is involved in the boundary, see [10, (1.1)–(1.5), p. 1561]. Another task is to consider the case of time dependent delay.

References

1. Bellen, A., Zennaro, M.: Numerical methods for delay differential equations. Numerical Mathematics and Scientific Computation. Oxford University Press, Oxford (2003)
2. Bradji, A., Ghoudi, T.: Some convergence results of a multidimensional finite volume scheme for a semilinear parabolic equation with a time delay. Numerical Methods and Applications. Lecture Notes in Computer Science, vol. 11189, pp. 351–359. Springer, Cham (2019)
3. Bradji, A.: Convergence analysis of some high-order time accurate schemes for a finite volume method for second order hyperbolic equations on general nonconforming multidimensional spatial meshes. Numer. Methods Partial. Differ. Equ. **29**(4), 1278–1321 (2013)
4. Bradji, A.: A theoretical analysis of a new finite volume scheme for second order hyperbolic equations on general nonconforming multidimensional spatial meshes. Numer. Methods Partial. Differ. Equ. **29**(1), 1–39 (2013)
5. Eymard, R., Gallouët, T., Herbin, R.: Discretization of heterogeneous and anisotropic diffusion problems on general nonconforming meshes. IMA J. Numer. Anal. **30**(4), 1009–1043 (2010)
6. Eymard, R., Gallouët, T., Herbin, R.: A cell-centred finite-volume approximation for anisotropic diffusion operators on unstructured meshes in any space dimension. IMA J. Numer. Anal. **26**, 326–353 (2006)
7. Eymard, R., Gallouët T., Herbin, R.: Finite volume methods. Ciarlet, P.G., Lions, J.L. (eds.) Handbook of Numerical Analysis, vol. VII , 723–1020 (2000)
8. Kuang, Y.: Delay differential equations: with applications in population dynamics. Mathematics in Science and Engineering, vol. 191. Academic Press, Boston, MA (1993)
9. Li, D., Zhang, C.: L^∞-error estimates of discontinuous Galerkin methods for delay differential equations. Appl. Numer. Math. **82**, 1–10 (2014)
10. Nicaise, S., Pignotti, C.: Stability and instability results of the wave equation with a delay term in the boundary or internal feedbacks. SIAM J. Control Optim. **45**(5), 1561–1585 (2006)

11. Parhi, N., Kirane, M.: Oscillatory behaviour of solutions of coupled hyperbolic differential equations. Analysis **14**(1), 43–56 (1994)
12. Raposo, C., Nguyen, H., Ribeiro, J.-O., Barros, V.: Well-posedness and exponential stability for a wave equation with nonlocal time-delay condition. Electron. J. Differ. Equ. Paper No. 279, 11 pp (2017)
13. Zhang, Q., Zhang, C.: A new linearized compact multisplitting scheme for the nonlinear convection-reaction-diffusion equations with delay

A Cell-Centered Finite Volume Method for the Navier–Stokes/Biot Model

Sergio Caucao, Tongtong Li, and Ivan Yotov

Abstract We develop a cell-centered finite volume method for the Navier–Stokes/Biot model, based on a fully mixed formulation with weakly symmetric stresses. The multipoint stress mixed finite element method is employed for the Navier–Stokes and elasticity equations, while the multipoint flux mixed finite element method is used for Darcy's flow. These methods allow for local elimination of the fluid and poroelastic stresses, vorticity, and rotation, resulting in a positive definite finite volume scheme for the fluid and structure velocities and the Darcy pressure, coupled via Lagrange multipliers on the interface to impose the transmission conditions.

Keywords Navier–Stokes/Biot · Mixed finite element · Multipoint flux · Multipoint stress · Finite volume method

MSC (2010) 65M08 · 65M60 · 74S05 · 76S05 · 76D05

1 Introduction

Modeling of the interaction between a free fluid and adjacent poroelastic media has been a subject of increased studies in recent years, due to its many applications, including flows in naturally fractured aquifers or reservoirs, hydraulic fracturing, and

S. Caucao is supported in part by BECAS CHILE para postdoctorado en el extranjero (convocatoria 2018). T. Li and I. Yotov are supported in part by NSF grant DMS 1818775.

S. Caucao · T. Li · I. Yotov (✉)
Department of Mathematics, University of Pittsburgh,
Pittsburgh, PA 15260, USA
e-mail: yotov@math.pitt.edu

S. Caucao
e-mail: sac304@pitt.edu

T. Li
e-mail: tol24@pitt.edu

R. Klöfkorn et al. (eds.), *Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples*, Springer Proceedings in Mathematics & Statistics 323,
https://doi.org/10.1007/978-3-030-43651-3_29

arterial flows. The free fluid flow is usually modeled by the Stokes or the Navier–Stokes equations, while the fluid flow within the poroelastic media is modeled by the Biot system of poroelasticity. The latter couples the Darcy model for fluid flow with elasticity. The two regions are coupled across the interface with kinematic and equilibrium transmission conditions. The coupled model is referred to as fluid-poroelastic structure interaction. Some recent works on the mathematical and numerical modeling of this problem can be found in [1, 3]. Due to the large size of the resulting algebraic system, the efficiency of its solution is of critical importance. The methods in [1, 3] are based on combinations of standard and mixed finite element (MFE) methods and require solving a large saddle point problem at each time step. Recently in [6], a fully mixed finite element method for the Stokes-Biot model is studied, that can be reduced to a cell-centered scheme within each region. Here we develop an extension of this approach to the Navier–Stokes/Biot model. The efficiency of the solution of this problem is even more important, since the algebraic system is nonlinear. The approach is based on coupling the multipoint stress MFE method [2] for a stress-velocity-vorticity formulation of Navier–Stokes and a stress-displacement-rotation formulation of elasticity with the the multipoint flux MFE method for Darcy flow [8]. These methods utilize the first order Brezzi-Douglas-Marini space [4] for the stresses and the Darcy velocity, along with a vertex quadrature rule for some of the bilinear forms. This allows for efficient local elimination of the stresses, Darcy velocity, vorticity, and rotation, resulting in a positive definite finite volume scheme for the fluid and structure velocities and the Darcy pressure, coupled via Lagrange multipliers on the interface to impose the transmission conditions.

We end this section by introducing some definitions and notation. For a domain $\mathscr{O} \subseteq \mathrm{R}^n$, $n \in \{2, 3\}$, and $p \in [1, +\infty]$, we denote by $\mathrm{L}^p(\mathscr{O})$ and $\mathrm{W}^{s,p}(\mathscr{O})$ the usual Lebesgue and Sobolev spaces. If $p = 2$ we write $\mathrm{H}^s(\mathscr{O})$ in place of $\mathrm{W}^{s,2}(\mathscr{O})$. Let $(\cdot, \cdot)_{\mathscr{O}}$ be the $L^2(\mathscr{O})$-inner product. For $\Gamma \subset \partial\mathscr{O}$, let $\langle \cdot, \cdot \rangle_\Gamma$ be the $L^2(\Gamma)$ inner product or duality pairing. By $\mathbf{M}$ and $\mathbb{M}$ we will denote the vectorial and tensorial counterparts of the generic scalar functional space M. For any vector field $\mathbf{v} = (v_i)_{i=1,n}$, we set the gradient, divergence, and tensor-product operators, as

$$\nabla\mathbf{v} := \left(\frac{\partial v_i}{\partial x_j}\right)_{i,j=1,n}, \quad \mathrm{div}(\mathbf{v}) := \sum_{j=1}^{n} \frac{\partial v_j}{\partial x_j}, \quad \text{and} \quad \mathbf{v} \otimes \mathbf{w} := (v_i w_j)_{i,j=1,n}.$$

Furthermore, for any tensor field $\boldsymbol{\tau} := (\tau_{ij})_{i,j=1,n}$ and $\boldsymbol{\zeta} := (\zeta_{ij})_{i,j=1,n}$, we let $\mathbf{div}(\boldsymbol{\tau})$ be the divergence operator div acting along the rows of $\boldsymbol{\tau}$, and define the transpose, the trace, the tensor inner product, and the deviatoric tensor, respectively, as

$$\boldsymbol{\tau}^{\mathrm{t}} := (\tau_{ji})_{i,j=1,n}, \quad \mathrm{tr}(\boldsymbol{\tau}) := \sum_{i=1}^{n} \tau_{ii}, \quad \boldsymbol{\tau} : \boldsymbol{\zeta} := \sum_{i,j=1}^{n} \tau_{ij}\zeta_{ij}, \quad \text{and} \quad \boldsymbol{\tau}^{\mathrm{d}} := \boldsymbol{\tau} - \frac{1}{n}\,\mathrm{tr}(\boldsymbol{\tau})\,\mathbf{I},$$

where $\mathbf{I}$ is the identity matrix in $\mathrm{R}^{n\times n}$. In addition, we recall the space

$$\mathbf{H}(\mathrm{div};\mathscr{O}) := \Big\{\mathbf{w}\in\mathbf{L}^2(\mathscr{O}) : \quad \mathrm{div}(\mathbf{w})\in\mathrm{L}^2(\mathscr{O})\Big\}.$$

The space of matrix valued functions with rows in $\mathbf{H}(\mathrm{div};\mathscr{O})$ is denoted by $\mathbb{H}(\mathbf{div};\mathscr{O})$.

2 Model Problem

Let $\Omega\subset\mathrm{R}^n$, $n\in\{2,3\}$ be a Lipschitz domain, which is subdivided into two non-overlapping regions: fluid region Ω_f and poroelastic region Ω_p. Let $\Gamma_{fp}=\partial\Omega_f\cap\partial\Omega_p$ denote the (nonempty) interface between these regions and let $\Gamma_f=\partial\Omega_f\setminus\Gamma_{fp}$ and $\Gamma_p=\partial\Omega_p\setminus\Gamma_{fp}$ denote the external parts of the boundary $\partial\Omega$. We denote by $\mathbf{n}_f$ and $\mathbf{n}_p$ the unit normal vectors which point outward from $\partial\Omega_f$ and $\partial\Omega_p$, respectively, noting that $\mathbf{n}_f=-\mathbf{n}_p$ on Γ_{fp}. Let $(\mathbf{u}_\star, p_\star)$ be the velocity-pressure pair in $\Omega_\star$ with $\star\in\{f,p\}$, and let $\boldsymbol{\eta}_p$ be the displacement in Ω_p. In addition, μ stands for the fluid viscosity, ρ_f is the density, $\mathbf{f}_\star$ is the body force term, and $q_\star$ is external source or sink term.

We assume that the flow in Ω_f is governed by the Navier–Stokes equations with constant density and viscosity, which are written in the following nonstandard pseudostress-velocity-pressure formulation:

$$\begin{aligned}\boldsymbol{\sigma}_f &= -p_f\,\mathbf{I}+2\,\mu\,\mathbf{e}(\mathbf{u}_f)-\rho_f\,(\mathbf{u}_f\otimes\mathbf{u}_f),\quad \mathrm{div}(\mathbf{u}_f)=q_f \quad\text{in}\quad \Omega_f\times(0,T],\\ \rho_f\left(\frac{\partial\,\mathbf{u}_f}{\partial\,t}+(\nabla\mathbf{u}_f)\,\mathbf{u}_f\right)&-\mathbf{div}\big(-p_f\,\mathbf{I}+2\,\mu\,\mathbf{e}(\mathbf{u}_f)\big)=\mathbf{f}_f\quad\text{in}\quad\Omega_f\times(0,T],\end{aligned}\tag{1}$$

with boundary conditions $\boldsymbol{\sigma}_f\mathbf{n}_f=\mathbf{0}$ on $\Gamma_f^{\mathrm{N}}\times(0,T]$, $\mathbf{u}_f=\mathbf{0}$ on $\Gamma_f^{\mathrm{D}}\times(0,T]$, where $\boldsymbol{\sigma}_f$ is the nonlinear pseudostress tensor, $\mathbf{e}(\mathbf{u}_f):=\big(\nabla\mathbf{u}_f+(\nabla\mathbf{u}_f)^{\mathrm{t}}\big)/2$ stands for the deformation rate tensor, $\Gamma_f=\Gamma_f^{\mathrm{D}}\cup\Gamma_f^{\mathrm{N}}$, and $T>0$ is the final time.

As in [5], we first observe that, due to tr $\mathbf{e}(\mathbf{u}_f)=\mathrm{div}(\mathbf{u}_f)=q_f$, there hold

$$\mathbf{div}(\mathbf{u}_f\otimes\mathbf{u}_f)=(\nabla\mathbf{u}_f)\,\mathbf{u}_f+q_f\,\mathbf{u}_f,\quad \mathrm{tr}(\boldsymbol{\sigma}_f)=-n\,p_f+2\,\mu\,q_f-\rho_f\,\mathrm{tr}(\mathbf{u}_f\otimes\mathbf{u}_f).\tag{2}$$

In particular, the pressure p_f can be written in terms of $\mathbf{u}_f$, $\boldsymbol{\sigma}_f$ and q_f as

$$p_f=-\frac{1}{n}\,\big(\mathrm{tr}(\boldsymbol{\sigma}_f)+\rho_f\,\mathrm{tr}(\mathbf{u}_f\otimes\mathbf{u}_f)-2\,\mu\,q_f\big)\,,\tag{3}$$

and hence, eliminating the pressure p_f, which can be recovered by (3), and employing the identities (2), problem (1) can be rewritten as

$$\begin{aligned} \boldsymbol{\sigma}_f^{\mathrm{d}} &= 2\,\mu\,\mathbf{e}(\mathbf{u}_f) - \rho_f\,(\mathbf{u}_f \otimes \mathbf{u}_f)^{\mathrm{d}} - \frac{2\,\mu}{n}\,q_f\,\mathbf{I} \quad \text{in} \quad \Omega_f \times (0,T], \\ \rho_f\,\frac{\partial\,\mathbf{u}_f}{\partial\,t} - \rho_f\,q_f\,\mathbf{u}_f - \mathbf{div}(\boldsymbol{\sigma}_f) &= \mathbf{f}_f \quad \text{in} \quad \Omega_f \times (0,T]. \end{aligned} \tag{4}$$

Next, in order to impose weakly the symmetry of $\boldsymbol{\sigma}_f$, we introduce

$$\boldsymbol{\gamma}_f := \frac{1}{2}\left(\nabla\mathbf{u}_f - (\nabla\mathbf{u}_f)^{\mathrm{t}}\right),$$

which represents the vorticity (or skew-symmetric part of the velocity gradient). Instead of (4), in the sequel we consider the problem with unknowns $\boldsymbol{\sigma}_f$, $\boldsymbol{\gamma}_f$ and $\mathbf{u}_f$,

$$\begin{aligned} \frac{1}{2\,\mu}\,\boldsymbol{\sigma}_f^{\mathrm{d}} &= \nabla\mathbf{u}_f - \boldsymbol{\gamma}_f - \frac{\rho_f}{2\,\mu}\,(\mathbf{u}_f \otimes \mathbf{u}_f)^{\mathrm{d}} - \frac{1}{n}\,q_f\,\mathbf{I} \quad \text{in} \quad \Omega_f \times (0,T], \\ \boldsymbol{\sigma}_f &= \boldsymbol{\sigma}_f^{\mathrm{t}}, \quad \rho_f\,\frac{\partial\,\mathbf{u}_f}{\partial\,t} - \rho_f\,q_f\,\mathbf{u}_f - \mathbf{div}(\boldsymbol{\sigma}_f) = \mathbf{f}_f \quad \text{in} \quad \Omega_f \times (0,T]. \end{aligned} \tag{5}$$

Next, let $\boldsymbol{\sigma}_e$ and $\boldsymbol{\sigma}_p$ be the elastic and poroelastic stress tensors, respectively,

$$\boldsymbol{\sigma}_e := \lambda_p\,\mathrm{div}(\boldsymbol{\eta}_p)\,\mathbf{I} + 2\,\mu_p\,\mathbf{e}(\boldsymbol{\eta}_p), \quad \boldsymbol{\sigma}_p := \boldsymbol{\sigma}_e - \alpha_p\,p_p\,\mathbf{I} \quad \text{in} \quad \Omega_p \times (0,T], \tag{6}$$

where $0 < \lambda_{\min} \le \lambda_p(\mathbf{x}) \le \lambda_{\max}$ and $0 < \mu_{\min} \le \mu_p(\mathbf{x}) \le \mu_{\max}$ are the Lamé parameters and $0 \le \alpha_p \le 1$ is the Biot–Willis constant. The poroelasticity region Ω_p is governed by the quasi-static Biot system:

$$\begin{aligned} &-\mathbf{div}(\boldsymbol{\sigma}_p) = \mathbf{f}_p, \quad \mu\,\mathbf{K}^{-1}\mathbf{u}_p + \nabla\,p_p = \mathbf{0} \quad \text{in} \quad \Omega_p \times (0,T], \\ &\frac{\partial}{\partial t}\left(s_0\,p_p + \alpha_p\,\mathrm{div}(\boldsymbol{\eta}_p)\right) + \mathrm{div}(\mathbf{u}_p) = q_p \quad \text{in} \quad \Omega_p \times (0,T], \end{aligned} \tag{7}$$

with boundary conditions $\mathbf{u}_p \cdot \mathbf{n}_p = 0$ on $\Gamma_p^{\mathrm{N}} \times (0,T]$, $p_p = 0$ on $\Gamma_p^{\mathrm{D}} \times (0,T]$, $\boldsymbol{\eta}_p = \mathbf{0}$ on $\Gamma_p \times (0,T]$, where $\Gamma_p = \Gamma_p^{\mathrm{N}} \cup \Gamma_p^{\mathrm{D}}$, $s_0 \ge 0$ is a storage coefficient and $\mathbf{K}$ is the symmetric and uniformly positive definite rock permeability tensor.

Next, we introduce the transmission conditions on the interface $\Gamma_{fp} \times (0,T]$ [1, 3]:

$$\begin{aligned} &\mathbf{u}_f \cdot \mathbf{n}_f + \left(\frac{\partial\,\boldsymbol{\eta}_p}{\partial t} + \mathbf{u}_p\right) \cdot \mathbf{n}_p = 0, \quad \boldsymbol{\sigma}_f\mathbf{n}_f + \boldsymbol{\sigma}_p\mathbf{n}_p = \mathbf{0}, \\ &(\boldsymbol{\sigma}_f\mathbf{n}_f) \cdot \mathbf{n}_f = -\,p_p, \quad (\boldsymbol{\sigma}_f\mathbf{n}_f) \cdot \mathbf{t}_{f,j} = -\,\mu\,\alpha_{\mathrm{BJS}}\sqrt{\mathbf{K}_j^{-1}}\left(\mathbf{u}_f - \frac{\partial\,\boldsymbol{\eta}_p}{\partial t}\right) \cdot \mathbf{t}_{f,j}, \end{aligned} \tag{8}$$

where $\mathbf{t}_{f,j}$, $1 \le j \le n-1$, is an orthogonal system of unit tangent vectors on Γ_{fp}, $\mathbf{K}_j = (\mathbf{K}\,\mathbf{t}_{f,j}) \cdot \mathbf{t}_{f,j}$, and $\alpha_{\mathrm{BJS}} \ge 0$ is an experimentally determined friction coefficient. The first and second equations in (8) correspond to mass conservation and conservation of momentum on Γ_{fp}, respectively, whereas the third and fourth are balance of normal fluid stress and the Beaver–Joseph–Saffman (BJS) slip with friction condition, respectively. Note that the third condition, which also arises in Stokes-Darcy couplings, implies a pressure jump on the interface. Finally, the above system of equations is complemented by the initial conditions $\mathbf{u}_f(\mathbf{x}, 0) = \mathbf{u}_{f,0}(\mathbf{x})$ in Ω_f and $p_p(\mathbf{x}, 0) = p_{p,0}(\mathbf{x})$ in Ω_p.

3 Weak Formulation

In this section we proceed analogously to [1, Sect. 3] (see also [7]) and derive a weak formulation of the coupled problem given by (5)–(8). Similarly to [5], we employ suitable Banach spaces to deal with the nonlinear stress tensor and velocity of the Navier–Stokes equation, together with the subspace of skew-symmetric tensors of $\mathbb{L}^2(\Omega_f)$ for the vorticity:

$$\begin{aligned}
\mathbb{X}_f &:= \Big\{ \boldsymbol{\tau}_f \in \mathbb{L}^2(\Omega_f) : \quad \mathbf{div}(\boldsymbol{\tau}_f) \in \mathbf{L}^{4/3}(\Omega_f) \quad \text{and} \quad \boldsymbol{\tau}_f \mathbf{n}_f = \mathbf{0} \quad \text{on} \quad \Gamma_f^{\mathrm{N}} \Big\}, \\
\mathbf{V}_f &:= \mathbf{L}^4(\Omega_f), \quad \mathbb{Q}_f := \Big\{ \boldsymbol{\chi}_f \in \mathbb{L}^2(\Omega_f) : \quad \boldsymbol{\chi}_f^{\mathrm{t}} = -\boldsymbol{\chi}_f \Big\}.
\end{aligned}$$

In turn, in order to deal with the unknowns in the Biot region we introduce the Hilbert spaces:

$$\begin{gathered}
\mathbb{X}_p := \mathbb{H}(\mathbf{div}; \Omega_p), \quad \mathbf{V}_s := \mathbf{L}^2(\Omega_p), \quad \mathbb{Q}_p := \Big\{ \boldsymbol{\chi}_p \in \mathbb{L}^2(\Omega_p) : \quad \boldsymbol{\chi}_p^{\mathrm{t}} = -\boldsymbol{\chi}_p \Big\}, \\
\mathbf{V}_p := \Big\{ \mathbf{v}_p \in \mathbf{H}(\mathrm{div}; \Omega_p) : \quad \mathbf{v}_p \cdot \mathbf{n} = 0 \quad \text{on} \quad \Gamma_p^{\mathrm{N}} \Big\}, \quad \mathrm{W}_p := \mathrm{L}^2(\Omega_p).
\end{gathered}$$

Finally, as in [1, 3, 7], we introduce the spaces of traces $\Lambda_p := \mathrm{H}^{1/2}(\Gamma_{fp})$, $\boldsymbol{\Lambda}_f := [\mathrm{H}^{1/2}(\Gamma_{fp})]^n$, and $\boldsymbol{\Lambda}_s := \Big[\mathrm{H}_{00}^{1/2}(\Gamma_{fp})\Big]^n := \Big\{ v|_{\Gamma_{fp}} : \quad v \in (\mathrm{H}^1(\Omega_p))^n, v = 0 \text{ on } \Gamma_p \Big\}$. Next, inspired by [1], we introduce the structure velocity $\mathbf{u}_s := \partial_t \boldsymbol{\eta}_p \in \mathbf{V}_s$ and the Lagrange multipliers

$$\boldsymbol{\varphi} := \mathbf{u}_f|_{\Gamma_{fp}} \in \boldsymbol{\Lambda}_f, \quad \boldsymbol{\theta} := \mathbf{u}_s|_{\Gamma_{fp}} \in \boldsymbol{\Lambda}_s, \quad \text{and} \quad \lambda := p_p|_{\Gamma_{fp}} \in \Lambda_p.$$

We employ a mixed elasticity formulation with weak stress symmetry, introducing as a new unknown the structure rotation operator

$$\boldsymbol{\gamma}_p := \frac{1}{2}\left(\nabla \mathbf{u}_s - (\nabla \mathbf{u}_s)^{\mathrm{t}}\right) \in \mathbb{Q}_p,$$

and the symmetric and positive definite compliance tensor A,

$$A(\boldsymbol{\tau}) := \frac{1}{2\,\mu_p}\left(\boldsymbol{\tau} - \frac{\lambda_p}{2\,\mu_p + n\,\lambda_p}\,\mathrm{tr}(\boldsymbol{\tau})\,\mathbf{I}\right), \quad A^{-1}(\boldsymbol{\tau}) = 2\,\mu_p\,\boldsymbol{\tau} + \lambda_p\,\mathrm{tr}(\boldsymbol{\tau})\,\mathbf{I}. \tag{9}$$

From the definition of the elastic and poroelastic stress tensors $\boldsymbol{\sigma}_e, \boldsymbol{\sigma}_p$, see (6) and the relation $A(\boldsymbol{\sigma}_e) = \mathbf{e}(\boldsymbol{\eta}_p)$, we deduce the identities

$$\mathrm{div}(\boldsymbol{\eta}_p) \;=\; A(\alpha_p\,p_p\,\mathbf{I}) : \mathbf{I} + A(\boldsymbol{\sigma}_p) : \mathbf{I} \tag{10}$$

and

$$\partial_t\,A(\boldsymbol{\sigma}_p) \;=\; \nabla\,\mathbf{u}_s - \boldsymbol{\gamma}_p - \partial_t\,A\,(\alpha_p\,p_p\,\mathbf{I}). \tag{11}$$

Then, similarly to [1, 3, 7], we test the first equation of (5), the second equation of (7), and (11) with arbitrary $\boldsymbol{\tau}_f \in \mathbb{X}_f$, $\mathbf{v}_p \in \mathbf{V}_p$, and $\boldsymbol{\tau}_p \in \mathbb{X}_p$, integrate by parts, utilize the fact that $\boldsymbol{\sigma}_f^{\mathrm{d}} : \boldsymbol{\tau}_f = \boldsymbol{\sigma}_f^{\mathrm{d}} : \boldsymbol{\tau}_f^{\mathrm{d}}$, test the third equation of (7) with $w_p \in \mathrm{W}_p$ employing (10), and impose the remaining equations weakly, as well as the symmetry of $\boldsymbol{\sigma}_f, \boldsymbol{\sigma}_p$, and the transmission conditions in (8) to obtain the following variational problem. Find $(\boldsymbol{\sigma}_f, \mathbf{u}_f, \boldsymbol{\gamma}_f, \boldsymbol{\varphi}, \boldsymbol{\sigma}_p, \mathbf{u}_s, \boldsymbol{\gamma}_p, \boldsymbol{\theta}, \mathbf{u}_p, p_p, \lambda) : [0, T] \mapsto \mathbb{X}_f \times \mathbf{V}_f \times \mathbb{Q}_f \times \boldsymbol{\Lambda}_f \times \mathbb{X}_p \times \mathbf{V}_s \times \mathbb{Q}_p \times \boldsymbol{\Lambda}_s \times \mathbf{V}_p \times W_p \times \Lambda_p$ such that for all $(\boldsymbol{\tau}_f, \mathbf{v}_f, \boldsymbol{\chi}_f, \boldsymbol{\psi}, \boldsymbol{\tau}_p, \mathbf{v}_s, \boldsymbol{\chi}_p, \boldsymbol{\phi}, \mathbf{v}_p, w_p, \xi)$,

$$\begin{aligned}
&\frac{1}{2\,\mu}\,(\boldsymbol{\sigma}_f^{\mathrm{d}}, \boldsymbol{\tau}_f^{\mathrm{d}})_{\Omega_f} - \langle \boldsymbol{\varphi}, \boldsymbol{\tau}_f\mathbf{n}_f\rangle_{\Gamma_{fp}} + (\mathbf{u}_f, \mathbf{div}\,\boldsymbol{\tau}_f)_{\Omega_f} \\
&\qquad + \frac{\rho_f}{2\,\mu}\,((\mathbf{u}_f \otimes \mathbf{u}_f)^{\mathrm{d}}, \boldsymbol{\tau}_f)_{\Omega_f} + (\boldsymbol{\gamma}_f, \boldsymbol{\tau}_f)_{\Omega_f} = -\frac{1}{n}\,(q_f, \mathrm{tr}(\boldsymbol{\tau}_f))_{\Omega_f}, \\
&\rho_f\,(\partial_t\,\mathbf{u}_f, \mathbf{v}_f)_{\Omega_f} - \rho_f\,(g_f\,\mathbf{u}_f, \mathbf{v}_f)_{\Omega_f} - (\mathbf{div}\,\boldsymbol{\sigma}_f, \mathbf{v}_f)_{\Omega_f} = (\mathbf{f}_f, \mathbf{v}_f)_{\Omega_f}, \\
&(\boldsymbol{\sigma}_f, \boldsymbol{\chi}_f)_{\Omega_f} = 0, \\
&(\partial_t\,A(\boldsymbol{\sigma}_p + \alpha_p\,p_p\,\mathbf{I}), \boldsymbol{\tau}_p)_{\Omega_p} - \langle \boldsymbol{\theta}, \boldsymbol{\tau}_p\mathbf{n}_p\rangle_{\Gamma_{fp}} + (\mathbf{u}_s, \mathbf{div}\,\boldsymbol{\tau}_p)_{\Omega_p} + (\boldsymbol{\gamma}_p, \boldsymbol{\tau}_p)_{\Omega_p} = 0, \\
&(\mathbf{div}\,\boldsymbol{\sigma}_p, \mathbf{v}_s)_{\Omega_p} = (\mathbf{f}_p, \mathbf{v}_s)_{\Omega_p}, \\
&(\boldsymbol{\sigma}_p, \boldsymbol{\chi}_p)_{\Omega_p} = 0, \\
&\mu\,(\mathbf{K}^{-1}\mathbf{u}_p, \mathbf{v}_p)_{\Omega_p} - (p_p, \mathrm{div}\,\mathbf{v}_p)_{\Omega_p} + \langle \lambda, \mathbf{v}_p \cdot \mathbf{n}_p\rangle_{\Gamma_{fp}} = 0, \\
&(s_0\,\partial_t\,p_p, w_p)_{\Omega_p} + \alpha_p\,(\partial_t\,A(\boldsymbol{\sigma}_p + \alpha_p\,p_p\,\mathbf{I}), w_p\,\mathbf{I})_{\Omega_p} + (w_p, \mathrm{div}\,\mathbf{u}_p)_{\Omega_p} = (q_p, w_p)_{\Omega_p}, \\
&\langle \boldsymbol{\varphi} \cdot \mathbf{n}_f + \big(\boldsymbol{\theta} + \mathbf{u}_p\big) \cdot \mathbf{n}_p, \xi\rangle_{\Gamma_{fp}} = 0, \\
&\langle \boldsymbol{\sigma}_p\mathbf{n}_p, \boldsymbol{\phi}\rangle_{\Gamma_{fp}} - \mu\,\alpha_{\mathtt{BJS}}\,\sum_{j=1}^{n-1}\langle\sqrt{\mathbf{K}_j^{-1}}\,(\boldsymbol{\varphi} - \boldsymbol{\theta}) \cdot \mathbf{t}_{f,j}, \boldsymbol{\phi} \cdot \mathbf{t}_{f,j}\rangle_{\Gamma_{fp}} + \langle \lambda, \boldsymbol{\phi} \cdot \mathbf{n}_p, \rangle_{\Gamma_{fp}} = 0, \\
&\langle \boldsymbol{\sigma}_f\mathbf{n}_f, \boldsymbol{\psi}\rangle_{\Gamma_{fp}} + \mu\,\alpha_{\mathtt{BJS}}\,\sum_{j=1}^{n-1}\langle\sqrt{\mathbf{K}_j^{-1}}\,(\boldsymbol{\varphi} - \boldsymbol{\theta}) \cdot \mathbf{t}_{f,j}, \boldsymbol{\psi} \cdot \mathbf{t}_{f,j}\rangle_{\Gamma_{fp}} + \langle \lambda . \boldsymbol{\psi} \cdot \mathbf{n}_f\rangle_{\Gamma_{fp}} = 0.
\end{aligned} \tag{12}$$

For the well posedness of the problem, compatible initial data is needed for all variables. It can be obtained from $\mathbf{u}_{f,0}$ and $p_{p,0}$ using that the equations without time derivatives hold at $t = 0$, see [1, 6].

4 Numerical Method

We employ a mixed finite element approximation of the weak formulation (12). Let $\mathscr{T}_h^f$ and $\mathscr{T}_h^p$ be affine finite element partitions of Ω_f and Ω_p, respectively, which may be non-matching along the interface Γ_{fp}. For the spatial discretization, we consider the conforming finite element spaces $\mathbb{X}_{fh} \times \mathbf{V}_{fh} \times \mathbb{Q}_{fh} = \mathbb{BDM}_1 - \mathbf{P}_0 - \mathbb{P}_1$, $\mathbb{X}_{ph} \times \mathbf{V}_{sh} \times \mathbb{Q}_{ph} = \mathbb{BDM}_1 - \mathbf{P}_0 - \mathbb{P}_1$, and $\mathbf{V}_{ph} \times \mathrm{W}_{ph} = \mathbf{BDM}_1 - \mathrm{P}_0$, where $\mathbf{BDM}_1$ denotes the first order Brezzi-Douglas-Marini space [4]. For the Lagrange multiplier spaces on Γ_{fp} we take $\boldsymbol{\Lambda}_{fh} = \mathbb{X}_{fh}\,\mathbf{n}_f$, $\boldsymbol{\Lambda}_{sh} = \mathbb{X}_{ph}\,\mathbf{n}_p$, and $\Lambda_{ph} = \mathbf{V}_{ph} \cdot \mathbf{n}_p$, resulting in $\boldsymbol{\Lambda}_{fh} \times \boldsymbol{\Lambda}_{sh} \times \Lambda_{ph} = \mathbf{P}_1^{\mathrm{dc}} - \mathbf{P}_1^{\mathrm{dc}} - \mathrm{P}_1^{\mathrm{dc}}$. For the time discretization we employ the backward Euler method. The straightforward application of the MFE method results, on each time step, in a large 11-field saddle point problem. In order to reduce the computational cost, we employ the vertex quadrature rule for some of the terms in (12), which allows for local elimination of certain variables. For a pair of tensor or vector valued functions (φ, ψ) and a linear operator L, define the quadrature rule

$$(L(\varphi), \psi)_{Q,\Omega_\star} := \sum_{E\in\mathscr{T}_h^\star} (L(\varphi), \psi)_{Q,E} = \sum_{E\in\mathscr{T}_h^\star} \frac{|E|}{s} \sum_{i=1}^{s} L(\varphi(\mathbf{r}_i)) : \psi(\mathbf{r}_i),$$

where $\star \in \{f, p\}$, $s = 3$ on triangles, $s = 4$ on tetrahedra or rectangles, and $\mathbf{r}_i$ are the vertices of E. The quadrature rule is applied to the terms

$$(\boldsymbol{\sigma}_f^{\mathrm{d}}, \boldsymbol{\tau}_f^{\mathrm{d}})_{\Omega_f},\ \ (\boldsymbol{\gamma}_f, \boldsymbol{\tau}_f)_{\Omega_f},\ \ (\boldsymbol{\sigma}_f, \boldsymbol{\chi}_f)_{\Omega_f},\ \ (\partial_t\, A(\boldsymbol{\sigma}_p + \alpha_p\, p_p\, \mathbf{I}), \boldsymbol{\tau}_p + \alpha_p w_p\, \mathbf{I})_{\Omega_p},$$
$$(\boldsymbol{\gamma}_p, \boldsymbol{\tau}_p)_{\Omega_p},\ \ (\boldsymbol{\sigma}_p, \boldsymbol{\chi}_p)_{\Omega_p},\ \ (\mathbf{K}^{-1}\mathbf{u}_p, \mathbf{v}_p)_{\Omega_p}.$$

Since the $\mathbf{BDM}_1$ degrees of freedom on each edge of face can be associated with the vertices, the quadrature rule results in block-diagonal stress and Darcy velocity matrices with one block per vertex. Therefore $\boldsymbol{\sigma}_f$, $\boldsymbol{\sigma}_p$, and $\mathbf{u}_p$ can be easily eliminated. The resulting matrices for the vorticity γ_f and the rotation γ_p are also block-diagonal, due the quadrature rule and the vertex degrees of freedom of these variables. They can also be eliminated, resulting in a cell-centered positive definite system for $\mathbf{u}_f$, $\mathbf{u}_s$, and p_p, coupled through the Lagrange multipliers $\boldsymbol{\varphi}$, $\boldsymbol{\theta}$, and λ. After solving this system, the rest of the variables are recovered from their elimination expressions. We refer to [6] for further details. The numerical method for the Stokes-Biot model is analyzed in [6], where first order convergence for all variables in their natural norms is shown. The analysis of the method presented in this paper for the nonlinear Navier–Stokes/Biot model will be developed in future work.

Table 1 EXAMPLE 1, Mesh sizes, errors, rates of convergences, and average number of Newton iterations

	$\Vert\mathbf{e}_{\boldsymbol{\sigma}_f}\Vert_{\ell^2(0,T;\mathbb{X}_f)}$		$\Vert\mathbf{e}_{\mathbf{u}_f}\Vert_{\ell^2(0,T;\mathbf{V}_f)}$		$\Vert\mathbf{e}_{\mathbf{u}_f}\Vert_{\ell^\infty(0,T;\mathbf{L}^2(\Omega_f))}$		$\Vert\mathbf{e}_{\boldsymbol{\gamma}_f}\Vert_{\ell^2(0,T;\mathbb{Q}_f)}$		$\Vert\mathbf{e}_{p_f}\Vert_{\ell^2(0,T;\mathrm{L}^2(\Omega_f))}$	
h_f	error	rate	error	rate	error	rate	error	rate	error	rate
0.1964	5.1E-01	–	3.4E-02	–	2.7E-01	–	3.2E-02	–	1.7E-01	–
0.0997	2.4E-01	1.1136	1.7E-02	0.9965	1.4E-01	1.0044	1.0E-02	1.6752	8.2E-02	1.0411
0.0487	1.2E-01	1.0327	8.5E-03	0.9978	6.8E-02	0.9943	4.2E-03	1.2504	3.9E-02	1.0249
0.0250	5.6E-02	1.0665	4.2E-03	1.0420	3.4E-02	1.0436	1.5E-03	1.4745	2.0E-02	1.0111
0.0136	2.8E-02	1.1521	2.1E-03	1.1458	1.7E-02	1.1449	6.5E-04	1.4287	1.0E-02	1.1489
0.0072	1.4E-02	1.0895	1.0E-03	1.1040	8.4E-03	1.0971	2.8E-04	1.3025	4.8E-03	1.1392
	$\Vert\mathbf{e}_{\boldsymbol{\sigma}_p}\Vert_{\ell^\infty(0,T;\mathbb{X}_p)}$		$\Vert\mathbf{e}_{\mathbf{u}_s}\Vert_{\ell^2(0,T;\mathbf{V}_s)}$		$\Vert\mathbf{e}_{\boldsymbol{\gamma}_p}\Vert_{\ell^2(0,T;\mathbb{Q}_p)}$		$\Vert\mathbf{e}_{\mathbf{u}_p}\Vert_{\ell^2(0,T;\mathbf{V}_p)}$		$\Vert\mathbf{e}_{p_p}\Vert_{\ell^\infty(0,T;\mathrm{W}_p)}$	
h_p	error	rate	error	rate	error	rate	error	rate	error	rate
0.2828	2.7E-01	–	4.3E-02	–	3.6E-02	–	1.0E-01	–	7.5E-02	–
0.1646	1.4E-01	1.2737	2.2E-02	1.2289	9.9E-03	2.3678	5.2E-02	1.2576	3.8E-02	1.2486
0.0779	6.7E-02	0.9651	1.1E-02	0.9623	2.3E-03	1.9774	2.5E-02	1.0003	1.9E-02	0.9335
0.0434	3.4E-02	1.1690	5.4E-03	1.1865	6.2E-04	2.1958	1.2E-02	1.2373	9.4E-03	1.2151
0.0227	1.7E-02	1.0635	2.7E-03	1.0668	2.0E-04	1.7255	5.9E-03	1.0816	4.7E-03	1.0659
0.0124	8.4E-03	1.1462	1.4E-03	1.1456	8.2E-05	1.5042	2.9E-03	1.1486	2.4E-03	1.1429
$\Vert\mathbf{e}_{\boldsymbol{\eta}_p}\Vert_{\ell^2(0,T;\mathbf{L}^2(\Omega_p))}$			$\Vert\mathbf{e}_{\boldsymbol{\varphi}}\Vert_{\ell^2(0,T;\mathbf{L}^2(\Gamma_{fp}))}$			$\Vert\mathbf{e}_{\boldsymbol{\theta}}\Vert_{\ell^2(0,T;\mathbf{L}^2(\Gamma_{fp}))}$		$\Vert\mathbf{e}_{\lambda}\Vert_{\ell^2(0,T;\mathrm{L}^2(\Gamma_{fp}))}$		
error	rate	h_{tf}	error	rate	h_{tp}	error	rate	error	rate	iter
2.7E-04	–	1/8	8.4E-03	–	1/5	1.0E-02	–	1.2E-03	–	4
1.4E-04	1.2275	1/16	2.1E-03	2.0195	1/10	3.3E-03	1.6431	3.2E-04	1.8656	4
6.7E-05	0.9623	1/32	4.7E-04	2.1340	1/20	6.1E-04	2.4481	7.7E-05	2.0334	4
3.4E-05	1.1865	1/64	1.2E-04	1.9659	1/40	1.7E-04	1.8741	1.9E-05	2.0006	4
1.7E-05	1.0668	1/128	2.8E-05	2.1140	1/80	3.9E-05	2.0897	4.9E-06	1.9817	4
8.4E-06	1.1456	1/256	7.7E-06	1.8636	1/160	9.0E-06	2.1194	1.2E-06	2.0796	4

5 Numerical Results

In this section we study numerically the convergence in space, using unstructured triangular grids. The total simulation time is $T = 0.01$ s and the time step is $\Delta t = 10^{-3}$ s, which is sufficiently small, so that the time discretization error does not affect the convergence rates. The domain is $\Omega = \Omega_f \cup \Gamma_{fp} \cup \Omega_p$, where $\Omega_f = (0, 1) \times (0, 1)$, $\Gamma_{fp} = (0, 1) \times \{0\}$, and $\Omega_p = (0, 1) \times (-1, 0)$. We take $\Gamma_f^D = (0, 1) \times \{1\}$ and $\Gamma_p^D = (0, 1) \times \{-1\}$. The solution in the Navier–Stokes region is

$$\mathbf{u}_f = \pi \cos(\pi t) \begin{pmatrix} -3x + \cos(y) \\ y + 1 \end{pmatrix}, \quad p_f = \exp(t) \sin(\pi x) \cos\left(\frac{\pi y}{2}\right) + 2\pi \cos(\pi t).$$

The Biot solution is chosen accordingly to satisfy the interface conditions (8):

$$p_p = \exp(t) \sin(\pi x) \cos\left(\frac{\pi y}{2}\right), \quad \mathbf{u}_p = -\frac{1}{\mu} \mathbf{K} \nabla p_p, \quad \boldsymbol{\eta}_p = \sin(\pi t) \begin{pmatrix} -3x + \cos(y) \\ y + 1 \end{pmatrix}.$$

We run a sequence of mesh refinements with non-matching grids along Γ_{fp}. The results are reported on Table 1. We note that the displacement at t_n is recovered by the formula $\boldsymbol{\eta}_p^n = \Delta t\, \mathbf{u}_s^n + \boldsymbol{\eta}_p^{n-1}$. As expected, we observe at least first order convergence for all subdomain variables in their natural norms. The Lagrange multiplier variables, which are approximated in $\mathbf{P}_1^{\mathrm{dc}} - \mathbf{P}_1^{\mathrm{dc}} - \mathrm{P}_1^{\mathrm{dc}}$, exhibit second order convergence in the L^2-norm on Γ_{fp}, which is consistent with the order of approximation.

References

1. Ambartsumyan, I., Ervin, V.J., Nguyen, T., Yotov, I.: A nonlinear Stokes–Biot model for the interaction of a non-Newtonian fluid with poroelastic media. ESAIM Math. Model. Numer. Anal. **53**, 1915–1955 (2019)
2. Ambartsumyan, I., Khattatov, E., Nordbotten, J., Yotov, I.: A multipoint stress mixed finite element method for elasticity on simplicial grids. SIAM J. Numer. Anal. (To appear)
3. Ambartsumyan, I., Khattatov, E., Yotov, I., Zunino, P.: A Lagrange multiplier method for a Stokes–Biot fluid-poroelastic structure interaction model. Numer. Math. **140**(2), 513–553 (2018)
4. Brezzi, F., Douglas Jr., J., Marini, L.D.: Two families of mixed finite elements for second order elliptic problems. Numer. Math. **47**(2), 217–235 (1985)
5. Camaño, J., García, C., Oyarzúa, R.: Analysis of a conservative mixed-FEM for the stationary Navier–Stokes problem. Preprint 2018-25, CI^2 MA, Universidad de Concepción, Chile (2018)
6. Caucao, S., Li, T., Yotov, I.: A multipoint stress-flux mixed finite element method for the Stokes–Biot model. In: preparation
7. Gatica, G., Márquez, A., Oyarzúa, R., Rebolledo, R.: Analysis of an augmented fully-mixed approach for the coupling of quasi-newtonian fluids and porous media. Comput. Methods Appl. Mech. Engrg. **270**, 76–112 (2014)
8. Wheeler, M.F., Yotov, I.: A multipoint flux mixed finite element method. SIAM J. Numer. Anal. **44**(5), 2082–2106 (2006)

Convergence Study of a DDFV Scheme for the Navier-Stokes Equations Arising in the Domain Decomposition Setting

Thierry Goudon, Stella Krell, and Giulia Lissoni

Abstract We consider DDFV discretization of the Navier-Stokes equations where the convection fluxes are computed by means of B-schemes, generalizing the classical centered and upwind discretizations. This study is motivated by the analysis of domain decomposition approaches. We investigate on numerical grounds the convergence of the method.

Keywords Navier-Stokes problem · DDFV schemes · B-schemes · Domain decomposition method

MSC2010 65M08 · 76D05 · 35Q35

1 Introduction

We consider the incompressible Navier-Stokes problem

$$\begin{cases} \partial_t \mathbf{u} + (\mathbf{u}\cdot\nabla)\mathbf{u} - \mathrm{div}(\sigma(\mathbf{u},\mathrm{p})) = \mathbf{f} & \text{in} \quad \Omega \times [0,T], \\ \mathrm{div}(\mathbf{u}) = 0 & \text{in} \quad \Omega \times [0,T], \\ \mathbf{u} = 0 & \text{on} \quad \partial\Omega \times [0,T], \\ \mathbf{u}(0) = \mathbf{u}_{init} & \text{in} \quad \Omega, \end{cases} \tag{1}$$

T. Goudon · S. Krell
Université Côte dAzur, CNRS, Inria, LJAD, Nice, France
e-mail: thierry.goudon@inria.fr

S. Krell
e-mail: stella.krell@univ-cotedazur.fr

G. Lissoni (✉)
Université de Nantes, LMJL, CNRS, Nantes, France
e-mail: giulia.lissoni@univ-nantes.fr

R. Klöfkorn et al. (eds.), *Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples*, Springer Proceedings in Mathematics & Statistics 323,
https://doi.org/10.1007/978-3-030-43651-3_30

where Ω is an open connected bounded polygonal domain of $\mathbb{R}^2$, $\mathbf{f} \in (L^2(\Omega))^2$ and $\mathbf{u}_{init} \in (L^\infty(\Omega))^2$ given. The unknowns $\mathbf{u} : \Omega \times [0, T] \to \mathbb{R}^2$ and $\mathrm{p} : \Omega \times [0, T] \to \mathbb{R}$ are respectively the velocity and the pressure; $\sigma(\mathbf{u}, \mathrm{p}) = \frac{2}{\mathrm{Re}}\mathrm{D}\mathbf{u} - \mathrm{pId}$ stands for the stress tensor, and $\mathrm{Re} > 0$ is the Reynolds number. Here and below, the strain rate tensor is defined by the symmetric part of the velocity gradient $\mathrm{D}\mathbf{u} = \frac{1}{2}(\nabla\mathbf{u} + {}^t\nabla\mathbf{u})$.

The Discrete Duality Finite Volume (DDFV) approach is quite appealing because it applies to very general meshes and it mimics at the discrete level the dual properties of the continuous differential operators. The introduction of the DDFV formalism dates back to [3, 5, 9], in order to approximate anisotropic diffusion problems on general meshes, including non-conformal and distorted meshes. DDFV schemes require unknowns on both the vertices and centers of primal control volumes; in particular, for the Stokes and Navier-Stokes problems it leads naturally to staggered discretizations of velocity and pressure; see [1, 4, 6, 10]. This work is motivated by the analysis of DDFV domain decomposition methods for (1). In contrast to direct methods, domain decomposition methods, in which the computational domain is decomposed into smaller subdomains, are naturally parallel; this makes those methods interesting for high performance computing perspectives. The classical Schwarz algorithm was proposed in 1870 by H. A. Schwarz for the Laplace problem and further studied in 1990 by P.-L. Lions, see [12, 13]. This approach has been adapted to many problems and motivates a huge literature.

In [7], we investigated non overlapping Schwarz algorithms in the DDFV framework for the Navier-Stokes system. The convergence analysis of the Schwarz iterations reveals a complex interplay between the design of the transmission conditions and the definition of the numerical fluxes. It turns out that the discrete limit problem does not coincide with the "standard" DDFV scheme on the entire domain; instead fluxes near the interface need to be modified. We are going to show, based on numerical experiments, that the modified scheme still provides a good approximation of the solution of (1) on Ω. Note that it is also possible to modify the fluxes of the domain decomposition method in order to restore a given DDFV scheme on Ω. These considerations rely on the formalism on B-schemes [2, 8] which allows us to consider general convection fluxes.

2 The DDFV Framework

We consider a domain Ω that can be seen as the union of two subdomains that share a common interface denoted by Γ.

Meshes: The complete description of the DDFV scheme for the 2D Navier-Stokes problem can be found in [6, 11]. A DDFV mesh is a pair $(\mathfrak{T}, \mathfrak{D})$; $\mathfrak{T}$ combines the primal mesh $\mathfrak{M} \cup \partial\mathfrak{M}$ (whose cells are denoted by κ), and the dual mesh $\mathfrak{M}^* \cup \partial\mathfrak{M}^*$, (whose cells κ^* are built around the vertices x_{κ^*} of the primal mesh).

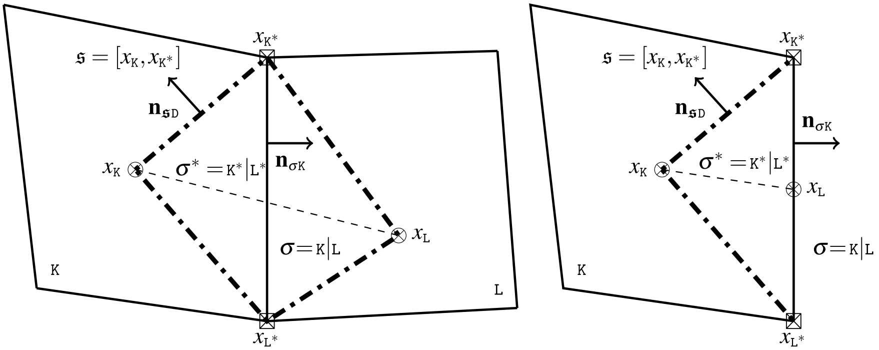


Fig. 1 *Left*: A diamond $\mathrm{D} = \mathrm{D}_{\sigma,\sigma^*}$ with $\sigma \notin \partial\Omega$. *Right*: A diamond $\mathrm{D} = \mathrm{D}_{\sigma,\sigma^*}$ with $\sigma \in \partial\Omega$

The primal mesh $\mathfrak{M}$ consists of disjoint polygons K called “primal cells”, whose union covers Ω. The symbol $\partial\mathfrak{M}$ denotes the set of edges of primal mesh included in $\partial\Omega$, that are considered as degenerated primal cells. We associate to each K a point x_{K}, called “center”. For the cells of the boundary, the point x_{K} is situated at the middle point of the edge. For all the neighbors volumes K and L, we suppose that $\partial\mathrm{K} \cap \partial\mathrm{L}$ is a segment that we call $\sigma = \mathrm{K}|\mathrm{L}$, edge of the primal mesh $\mathfrak{M}$.

From this primal mesh, we build the associated dual mesh. A dual cell K^* is associated to a vertex x_{K^*} of the primal mesh. The dual cells are obtained by joining the centers of the primal cells that have x_{K^*} as vertex. Then, the point x_{K^*} is called center of K^*. We will distinguish interior dual mesh, for which x_{K^*} does not belong to $\partial\Omega$, denoted by $\mathfrak{M}^*$ and the boundary dual mesh, for which x_{K^*} belongs to $\partial\Omega$, denoted by $\partial\mathfrak{M}^*$. We denote with $\sigma^* = \mathrm{K}^*|\mathrm{L}^*$ the edges of the dual mesh.

Next, $\mathfrak{D}$ stands for the diamond mesh, whose cells $\mathrm{D} = \mathrm{D}_{\sigma,\sigma^*}$ are built such that their principal diagonals are a primal edge σ and a dual edge σ^*. Thus a diamond is a quadrilateral with vertices $x_{\mathrm{K}}, x_{\mathrm{L}}, x_{\mathrm{K}^*}$ and x_{L^*}. Note that we have $\Omega = \bigcup_{D\in\mathfrak{D}} \mathfrak{D}$. We distinguish the diamonds that intersect the interface Γ as $\mathfrak{D}^\Gamma = \{D_{\sigma,\sigma^*} \in \mathfrak{D},\ \text{such that}\ \sigma \subset \Gamma\}$.

For a diamond cell D we note by m_{D} its measure, m_σ the length of the primal edge σ, m_{σ^*} the length of the dual edge σ^*, $\mathbf{n}_{\sigma\mathrm{K}}$ the unit vector normal to σ oriented from x_{K} to x_{L}, $\mathbf{n}_{\sigma^*\mathrm{K}^*}$ the unit vector normal to σ^* oriented from x_{K^*} to x_{L^*}. We denote also its sides by $\mathfrak{s}$ and their measure by $m_{\mathfrak{s}}$; see Fig. 1 for an illustration.

Finally, we denote by $\mathbf{f}_{\mathrm{K}}$ (resp. $\mathbf{f}_{\mathrm{K}^*}$) the mean-value of the source term $\mathbf{f}$ on $\mathrm{K} \in \mathfrak{M}$ (resp. on $\mathrm{K}^* \in \mathfrak{M}^* \cup \partial\mathfrak{M}^*$).

Unknowns: The DDFV method for Navier-Stokes problem uses staggered unknowns. We associate to every $\mathrm{K} \in \mathfrak{M} \cup \partial\mathfrak{M}$ an unknown $\mathbf{u}_{\mathrm{K}} \in \mathbb{R}^2$, to every $\mathrm{K}^* \in \mathfrak{M}^* \cup \partial\mathfrak{M}^*$ an unknown $\mathbf{u}_{\mathrm{K}^*} \in \mathbb{R}^2$ for the velocity and to every $\mathrm{D} \in \mathfrak{D}$ an unknown $\mathrm{p}^{\mathrm{D}} \in \mathbb{R}$ for the pressure. Those unknowns are collected in the families:

$$\mathbf{u}_{\mathfrak{T}} = \left((\mathbf{u}_{\mathrm{K}})_{\mathrm{K}\in(\mathfrak{M}\cup\partial\mathfrak{M})}, (\mathbf{u}_{\mathrm{K}^*})_{\mathrm{K}^*\in(\mathfrak{M}^*\cup\partial\mathfrak{M}^*)}\right) \in \left(\mathbb{R}^2\right)^{\mathfrak{T}} \quad \text{and} \quad \mathrm{p}_{\mathfrak{D}} = ((\mathrm{p}^{\mathrm{D}})_{\mathrm{D}\in\mathfrak{D}}) \in \mathbb{R}^{\mathfrak{D}}.$$

We define the subspace of $\left(\mathbb{R}^2\right)^{\mathfrak{T}}$ that takes into account Dirichlet boundary conditions:

$$\mathbb{E}_0 = \{\mathbf{u}_{\mathfrak{T}} \in \left(\mathbb{R}^2\right)^{\mathfrak{T}}, \text{ s. t. } \forall \mathrm{K} \in \partial\mathfrak{M},\ \mathbf{u}_{\mathrm{K}} = 0 \text{ and } \forall \mathrm{K}^* \in \partial\mathfrak{M}^*,\ \mathbf{u}_{\mathrm{K}^*} = 0\}.$$

For $\mathbf{v} \in (H^2(\Omega))^2$, we set $\mathbb{P}_c^{\mathfrak{T}}(\mathbf{v}) = \left((\mathbf{v}(x_{\mathrm{K}}))_{\mathrm{K}\in\mathfrak{M}\cup\partial\mathfrak{M}}, (\mathbf{v}(x_{\mathrm{K}^*}))_{\mathrm{K}^*\in\mathfrak{M}^*\cup\partial\mathfrak{M}^*}\right)$.

Discrete operators: We define a piecewise constant approximation of the gradient operator denoted by $\nabla^{\mathfrak{D}} : \left(\mathbb{R}^2\right)^{\mathfrak{T}} \to (\mathrm{M}_2(\mathbb{R}))^{\mathfrak{D}}$,

$$\nabla^{\mathrm{D}}\mathbf{u}_{\mathfrak{T}} := \frac{1}{2m_{\mathrm{D}}}\left[m_\sigma(\mathbf{u}_{\mathrm{L}} - \mathbf{u}_{\mathrm{K}}) \otimes \mathbf{n}_{\sigma\mathrm{K}} + m_{\sigma^*}(\mathbf{u}_{\mathrm{L}^*} - \mathbf{u}_{\mathrm{K}^*}) \otimes \mathbf{n}_{\sigma^*\mathrm{K}^*}\right], \quad \forall \mathrm{D} \in \mathfrak{D}.$$

To work with the Navier-Stokes problem, we also need to define the *discrete strain rate tensor* $\mathrm{D}^{\mathfrak{D}} : \mathbf{u}_{\mathfrak{T}} \in (\mathbb{R}^2)^{\mathfrak{T}} \mapsto (\mathrm{D}^{\mathrm{D}}\mathbf{u}_{\mathfrak{T}})_{\mathrm{D}\in\mathfrak{D}} \in (\mathrm{M}_2(\mathbb{R}))^{\mathfrak{D}}$, such that:

$$\mathrm{D}^{\mathrm{D}}\mathbf{u}_{\mathfrak{T}} = \frac{\nabla^{\mathrm{D}}\mathbf{u}_{\mathfrak{T}} + {}^t(\nabla^{\mathrm{D}}\mathbf{u}_{\mathfrak{T}})}{2}, \quad \text{for } \mathrm{D} \in \mathfrak{D},$$

the *discrete stress tensor* $\sigma^{\mathfrak{D}} : (\mathbf{u}_{\mathfrak{T}}, \mathrm{p}_{\mathfrak{D}}) \in \left(\mathbb{R}^2\right)^{\mathfrak{T}} \times \mathbb{R}^{\mathfrak{D}} \mapsto (\sigma^{\mathrm{D}}(\mathbf{u}_{\mathfrak{T}}, \mathrm{p}_{\mathfrak{D}}))_{\mathrm{D}\in\mathfrak{D}} \in (\mathrm{M}_2(\mathbb{R}))^{\mathfrak{D}}$

$$\sigma^{\mathrm{D}}(\mathbf{u}_{\mathfrak{T}}, \mathrm{p}_{\mathfrak{D}}) = -\left(\frac{2}{\mathrm{Re}}\mathrm{D}^{\mathrm{D}}\mathbf{u}_{\mathfrak{T}} - \mathrm{p}^{\mathrm{D}}\mathrm{Id}\right), \quad \text{for } \mathrm{D} \in \mathfrak{D},$$

and the *discrete divergence of a vector field of* $(\mathbb{R}^2)^{\mathfrak{T}}$ as $\mathrm{div}^{\mathfrak{D}} : \mathbf{u}_{\mathfrak{T}} \in (\mathbb{R}^2)^{\mathfrak{T}} \mapsto (\mathrm{div}^{\mathrm{D}}\mathbf{u}_{\mathfrak{T}})_{\mathrm{D}\in\mathfrak{D}} \in \mathbb{R}^{\mathfrak{D}}$ with $\mathrm{div}^{\mathrm{D}}\mathbf{u}_{\mathfrak{T}} = \mathrm{Tr}(\nabla^{\mathrm{D}}\mathbf{u}_{\mathfrak{T}})$ for any $\mathrm{D} \in \mathfrak{D}$.

To treat convection terms, it is convenient to define the scalar velocity fluxes $F_{\sigma\mathrm{K}}$ and $F_{\sigma^*\mathrm{K}^*}$; their definition comes from [11], up to the boundary terms. They are approximations of the fluxes: $\displaystyle\int_\sigma (\mathbf{u} \cdot \mathbf{n}_{\sigma\mathrm{K}}) \rightsquigarrow F_{\sigma\mathrm{K}}(\mathbf{u}_{\mathfrak{T}})$ and $\displaystyle\int_{\sigma^*} (\mathbf{u} \cdot \mathbf{n}_{\sigma^*\mathrm{K}^*}) \rightsquigarrow F_{\sigma^*\mathrm{K}^*}(\mathbf{u}_{\mathfrak{T}})$. By defining $m_{\mathfrak{s}}G_{\mathfrak{s},\mathrm{D}} = m_{\mathfrak{s}}\dfrac{\mathbf{u}_{\mathrm{K}} + \mathbf{u}_{\mathrm{K}^*}}{2} \cdot \mathbf{n}_{\mathfrak{s}\mathrm{D}}$, for the *primal edges*, we impose:

$$m_\sigma F_{\sigma\mathrm{K}} = -\sum_{\mathfrak{s}\in\partial\mathrm{D}\cap\mathrm{K}} m_{\mathfrak{s}}G_{\mathfrak{s},\mathrm{D}},$$

see Fig. 1. The velocity fluxes $F_{\sigma\mathrm{K}}$ and $F_{\sigma^*\mathrm{K}^*}$ are conservative, that is to say $F_{\sigma\mathrm{K}} = -F_{\sigma\mathrm{L}}$, $\forall\sigma = \mathrm{K}|\mathrm{L}$ and $F_{\sigma^*\mathrm{K}^*} = -F_{\sigma^*\mathrm{L}^*}$, $\forall\sigma^* = \mathrm{K}^*|\mathrm{L}^*$. Next, since $\int_{\mathrm{K}}(\mathbf{u} \cdot \nabla)\mathbf{v} = \sum_{\sigma\subset\partial\mathrm{K}} \int_\sigma (\mathbf{u} \cdot \mathbf{n}_{\sigma\mathrm{K}})\mathbf{v}$ holds for any $\mathrm{K} \in \mathfrak{M}$, we approximate the convection terms as follows

$$\int_{\mathrm{K}} (\mathbf{u} \cdot \nabla)\mathbf{v} \rightsquigarrow \sum_{\sigma\subset\partial\mathrm{K}} m_\sigma F_{\sigma\mathrm{K}}\left(\frac{\mathbf{v}_{\mathrm{K}} + \mathbf{v}_{\mathrm{L}}}{2}\right),$$

with a *centered* discretization for $\mathbf{v}$. For the *dual edges* the definition is similar.

3 DDFV Scheme for the Navier-Stokes Equations

The DDFV scheme under consideration is obtained by an implicit Euler time discretization, except for the nonlinear term, which is linearized by using a semi-implicit approximation. Let $N \in \mathbb{N}^*$. We note $\delta t = \frac{T}{N}$ and $t_n = n\delta t$ for $n \in \{0, \dots, N\}$. We look for $\mathbf{u}_{\mathfrak{T}}^{[0,T]} = (\mathbf{u}^n)_{n\in\{0,\dots N\}} \in \left(\mathbb{E}_0\right)^{N+1}$ and $\mathrm{p}_{\mathfrak{D}}^{[0,T]} = (\mathrm{p}^n)_{n\in\{1,\dots N\}} \in (\mathbb{R}^{\mathfrak{D}})^{N+1}$, and the scheme is initialized with $\mathbf{u}^0 = \mathbb{P}_c^{\mathfrak{T}}\mathbf{u}_0$ in $\mathbb{E}_0$.

To simplify the notations, we denote $(\mathbf{u}^{n+1}, \mathrm{p}^{n+1})$ with $(\mathbf{u}_{\mathfrak{T}}, \mathrm{p}_{\mathfrak{D}})$ and $(\mathbf{u}^n, \mathrm{p}^n)$ with $(\bar{\mathbf{u}}_{\mathfrak{T}}, \bar{\mathrm{p}}_{\mathfrak{D}})$ that at each time step are known. Given $(\bar{\mathbf{u}}_{\mathfrak{T}}, \bar{\mathrm{p}}_{\mathfrak{D}})$, we look for $(\mathbf{u}_{\mathfrak{T}}, \mathrm{p}_{\mathfrak{D}}) \in \mathbb{E}_0 \times \mathbb{R}^{\mathfrak{D}}$ such that:

$$\left\{\begin{aligned}
m_{\mathrm{K}}\frac{\mathbf{u}_{\mathrm{K}}}{\delta t} + \sum_{\sigma\in\partial\mathrm{K}} m_\sigma \mathscr{F}_{\sigma\mathrm{K}} &= m_{\mathrm{K}}\mathbf{f}_{\mathrm{K}} + m_{\mathrm{K}}\frac{\bar{\mathbf{u}}_{\mathrm{K}}}{\delta t} && \forall \mathrm{K} \in \mathfrak{M}\\
m_{\mathrm{K}^*}\frac{\mathbf{u}_{\mathrm{K}^*}}{\delta t} + \sum_{\sigma^*\in\partial\mathrm{K}^*} m_{\sigma^*} \mathscr{F}_{\sigma^*\mathrm{K}^*} &= m_{\mathrm{K}^*}\mathbf{f}_{\mathrm{K}^*} + m_{\mathrm{K}^*}\frac{\bar{\mathbf{u}}_{\mathrm{K}^*}}{\delta t} && \forall \mathrm{K}^* \in \mathfrak{M}^*\\
m_{\mathrm{D}}\mathrm{div}^{\mathrm{D}}(\mathbf{u}_{\mathfrak{T}}) &= 0 && \forall \mathrm{D} \in \mathfrak{D}\\
\sum_{\mathrm{D}\in\mathfrak{D}} m_{\mathrm{D}}\mathrm{p}^{\mathrm{D}} &= 0,
\end{aligned}\right. \qquad (\widetilde{\mathscr{P}})$$

The total fluxes $\mathscr{F}_{\sigma\mathrm{K}}$, $\mathscr{F}_{\sigma^*\mathrm{K}^*}$ read

$$m_\sigma \mathscr{F}_{\sigma\mathrm{K}} = -m_\sigma \sigma^{\mathrm{D}}(\mathbf{u}_{\mathfrak{T}}, \mathrm{p}_{\mathfrak{D}})\,\mathbf{n}_{\sigma\mathbf{K}} + \left[m_\sigma F_{\sigma\mathrm{K}}\left(\frac{\mathbf{u}_{\mathbf{K}} + \mathbf{u}_{\mathbf{L}}}{2}\right) + \frac{m_\sigma^2}{2\mathrm{Re}m_{\mathrm{D}}}B_{\sigma\mathrm{K}}(\mathbf{u}_{\mathbf{K}} - \mathbf{u}_{\mathbf{L}})\right],$$

$$\begin{aligned}
m_{\sigma^*}\mathscr{F}_{\sigma^*\mathrm{K}^*} = -m_{\sigma^*}\sigma^{\mathrm{D}}(\mathbf{u}_{\mathfrak{T}}, \mathrm{p}_{\mathfrak{D}})\,\mathbf{n}_{\sigma^*\mathbf{K}^*}&\\
+\left[m_{\sigma^*}F_{\sigma^*\mathrm{K}^*}\left(\frac{\mathbf{u}_{\mathbf{K}^*} + \mathbf{u}_{\mathbf{L}^*}}{2}\right) + \frac{m_{\sigma^*}^2}{2\mathrm{Re}m_{\mathrm{D}}}B_{\sigma^*\mathrm{K}^*}(\mathbf{u}_{\mathbf{K}^*} - \mathbf{u}_{\mathbf{L}^*})\right]&.
\end{aligned}$$

They are the sum of a “diffusion” term, discretized by means of the DDFV operators defined in Sect. 2, and a “convection” term, approximated through general B-schemes, as in [2, 8]. It means that the latter are written as a centered discretization plus a diffusive perturbation, which depends on a certain function B. The definition of the velocity fluxes $F_{\sigma\mathrm{K}}$, $F_{\sigma^*\mathrm{K}^*}$ comes from the literature and it can be found in Sect. 2; they are computed with the velocity of the previous time step. We now need to define the matrices $B_{\sigma\mathrm{K}}$, $B_{\sigma^*\mathrm{K}^*}$.

Definition of the diffusive perturbations to the convection fluxes

Our study is motivated by domain decomposition purposes: the domain Ω is seen as the union of two subdomains that share a common interface Γ. A specific definition of the total fluxes is required on the interface, as a trace of the iteration process [7]. The diamonds of Ω which cross the interface Γ are split into two boundary diamonds on the subdomains; they share the *primal edge* σ, which lies on the interface Γ, while the *dual edge* σ^* is divided into $\sigma^* \cap \mathrm{K}$ and $\sigma^* \cap \mathrm{L}$, see Fig. 1. The convergence of the Schwarz algorithm amounts to re-glue the two pieces of such diamonds. This entails the following properties on the *total fluxes*

$$\begin{aligned} m_\sigma \mathscr{F}_{\sigma\mathrm{K}} &= -m_\sigma \mathscr{F}_{\sigma\mathrm{L}} \\ m_{\sigma^*} \mathscr{F}_{\sigma^*\mathrm{K}^*} &= m_{\sigma^*\cap\mathrm{K}} \mathscr{F}_{\sigma^*\cap\mathrm{K},\mathrm{K}^*} + m_{\sigma^*\cap\mathrm{L}} \mathscr{F}_{\sigma^*\cap\mathrm{L},\mathrm{K}^*} \end{aligned} \tag{2}$$

which are not naturally satisfied. Relations (2) lead to algebraic constraints, which, in turn, modify the definition of the coefficients $B_{\sigma\mathrm{K}}$, $B_{\sigma^*\mathrm{K}^*}$ on the interface. In particular it leads to work with matrix-valued $B_{\sigma\mathrm{K}}$, $B_{\sigma^*\mathrm{K}^*}$. Therefore, for the *primal mesh*, we have

$$B_{\sigma\mathrm{K}} := \begin{cases} B\left(\dfrac{2\mathrm{Re} m_{\mathrm{D}}}{m_\sigma} F_{\sigma\mathrm{K}}\right) \mathrm{Id} & \forall \mathrm{D}_{\sigma,\sigma^*} \in \mathfrak{D} \setminus \mathfrak{D}^\Gamma \\ \dfrac{2\mathrm{Re} m_{\mathrm{D}}}{m_\sigma^2} \left(A_{\mathrm{K}} A_{\mathrm{L}} + \left(\dfrac{1}{2} m_\sigma F_{\sigma\mathrm{K}}\right)^2 \mathrm{Id} \right) A^{-1} - P & \forall \mathrm{D}_{\sigma,\sigma^*} \in \mathfrak{D}^\Gamma \end{cases} \tag{3}$$

with $P = \mathrm{Id} + \mathbf{n}_{\sigma\mathbf{K}} \otimes \mathbf{n}_{\sigma\mathbf{K}}$ for $\sigma = \mathrm{K}|\mathrm{L}$ and

$$A_{\mathrm{K}} := \frac{m_\sigma^2}{2\mathrm{Re} m_{\mathrm{D}\cap\mathrm{K}}} \left(P + B\left(\frac{2\mathrm{Re} m_{\mathrm{D}\cap\mathrm{K}}}{m_\sigma} F_{\sigma\mathrm{K}}\right) \mathrm{Id} \right), \quad A := A_{\mathrm{K}} + A_{\mathrm{L}}.$$

For the *dual mesh*, we have

$$B_{\sigma^*\mathrm{K}^*} = \begin{cases} B\left(\dfrac{2\mathrm{Re} m_{\mathrm{D}}}{m_{\sigma^*}} F_{\sigma^*\mathrm{K}^*}\right) \mathrm{Id}, & \forall \mathrm{D}_{\sigma,\sigma^*} \in \mathfrak{D} \setminus \mathfrak{D}^\Gamma \\ \dfrac{m_{\sigma^*\cap\mathrm{K}}}{m_{\sigma^*}} B\left(\dfrac{2\mathrm{Re} m_{\mathrm{D}\cap\mathrm{K}}}{m_{\sigma^*\cap\mathrm{K}}} F_{\sigma^*\cap\mathrm{K}}\right) \mathrm{Id} + \dfrac{m_{\sigma^*\cap\mathrm{L}}}{m_{\sigma^*}} B\left(\dfrac{2\mathrm{Re} m_{\mathrm{D}\cap\mathrm{L}}}{m_{\sigma^*\cap\mathrm{L}}} F_{\sigma^*\cap\mathrm{L}}\right) \mathrm{Id}, & \forall \mathrm{D}_{\sigma,\sigma^*} \in \mathfrak{D}^\Gamma \end{cases} \tag{4}$$

where, for $\mathrm{D}_{\sigma,\sigma^*} \in \mathfrak{D}^\Gamma$, we set $m_{\sigma^*\cap\mathrm{K}} F_{\sigma^*\cap\mathrm{K}} = -m_{\mathfrak{s}} G_{\mathfrak{s},\mathrm{D}} - \dfrac{1}{2} \displaystyle\sum_{\sigma \subset \mathrm{K}^* \cap \Gamma} m_{\sigma\cap\mathrm{K}^*} \bar{\mathbf{u}}_{\mathrm{K}^*} \cdot \mathbf{n}_{\sigma\mathbf{K}}$.

Therefore the details of the fluxes depend on the function B which appears in these definitions. On the interior diamonds $\mathfrak{D} \setminus \mathfrak{D}^\Gamma$, for both *primal* and *dual meshes*, standard choices are $B(s) = 0$ which leads to the centered scheme, or $B(s) = \frac{1}{2}|s|$, which corresponds to the upwind scheme. We refer the reader to [11] for the analysis of the DDFV scheme for (1) with the upwind scheme on the entire domain Ω. This result generalizes as follows.

Theorem 1 *Let $\mathfrak{T}$ be a mesh that satisfies inf-sup stability condition and let B be an even Lipschitz continuous function such that $B(s) \geq 0$, $\forall \mathfrak{s} \in \mathbb{R}$. Then, the problem $\widetilde{(\mathscr{P})}$ is well-posed.*

The hypothesis of *inf-sup stability* ([1]) on the mesh can be dropped by stabilizing the incompressibility constraint. For the proof, we refer the reader to [7].

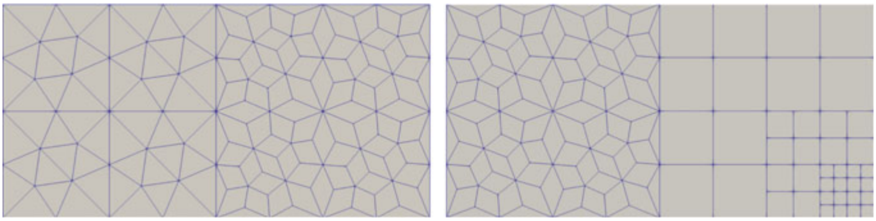
4 Numerical Results

In this Section, the scheme $(\widetilde{\mathscr{P}})$ is validated by some numerical experiments. The computational domain is $\Omega = [-1, 1] \times [0, 1]$ and the interface Γ is placed at $x = 0$. For the tests, we give the expression of the exact solution ($\mathbf{u}$, p), from which we deduce the source term $\mathbf{f}$. We compare the L^2-norm of the error (difference between a centered projection of the exact solution and the approximated solution obtained with DDFV scheme) for the velocity (denoted Ervel), the velocity gradient (Ergradvel) and the pressure (Erpre). The error estimates are discussed by working with a family of meshes (see Fig. 2), obtained by refining successively and uniformly the original mesh. The sub-index in the name of the mesh denotes the level of refinement, i.e. Mesh_1^k represents the coarse mesh of a family of refined meshes $(\text{Mesh}_m^k)_m$. More precisely, Mesh_m^k is obtained by dividing by two all the edges of Mesh_{m-1}^k. The meshes in those examples are non conformal.

We consider the following exact solutions to (1):

$$\begin{aligned} \mathbf{u}(t, x, y) &= \begin{pmatrix} -2\pi \cos(\pi x) \sin(2\pi y) \exp(-5\eta t \pi^2), \\ \pi \sin(\pi x) \cos(2\pi y) \exp(-5\eta t \pi^2) \end{pmatrix}, \\ \mathrm{p}(t, x, y) &= -\frac{\pi^2}{4}(4\cos(2\pi x) + \cos(4\pi y)) \exp(-10 t \eta \pi^2). \end{aligned} \tag{5}$$

The final time is $T = 0.3$ and we fix $\delta t = 1.5 \times 10^{-3}$, $\eta = \text{Re} = 1$, and $B(s) = \frac{1}{2}|s|$. In Tables 1 and 2, we observe convergence of order 1 for the L^2 norm of the velocity, the H^1 norm of the velocity and for the L^2 norm of the pressure. Those results are comparable to the ones presented in [11]. This underlines that the presence of the interface Γ and the modified fluxes that appear in (3), (4) do not influence the convergence results. The solution of $(\widetilde{\mathscr{P}})$ is a good approximation of the solution of (1).



(a) Mesh_1^1

(b) Mesh_1^2

Fig. 2 Coarse level of refinement of the meshes on Ω

Table 1 Test (5) on Mesh^1_m, $m = 1, \dots 5$

Mesh	NbCell	Ervel	Ratio	Ergradvel	Ratio	Erpre	Ratio
Mesh^1_1	896	2.414E−002	–	8.568E−002	–	1.178	–
Mesh^1_2	3300	6.921E−003	1.80	3.726E−002	1.20	0.507	1.21
Mesh^1_3	12644	2.938E−003	1.23	1.861E−002	1.00	0.186	1.44
Mesh^1_4	49476	1.493E−003	0.97	9.281E−003	1.00	6.850E−002	1.44
Mesh^1_5	195716	6.802E−004	**1.13**	4.594E−003	**1.01**	2.772E−002	**1.30**

Table 2 Test (5) on Mesh^2_m, $m = 1, \dots 5$

Mesh	NbCell	Ervel	Ratio	Ergradvel	Ratio	Erpre	Ratio
Mesh^2_1	924	8.288E−002	–	0.147	–	5.130	–
Mesh^2_2	3332	1.923E−002	2.10	5.596E−002	1.39	2.025	1.34
Mesh^2_3	12612	4.691E−003	2.03	2.425E−002	1.20	0.674	1.58
Mesh^2_4	49028	1.811E−003	1.37	1.135E−002	1.09	0.214	1.65
Mesh^2_5	193284	7.725E−004	**1.23**	5.460E−003	**1.05**	7.083E−002	**1.59**

References

1. Boyer, F., Krell, S., Nabet, F.: Inf-sup stability of the discrete duality finite volume method for the 2D Stokes problem. Math. Comput. **84**, 2705–2742 (2015)
2. Chainais-Hillairet, C., Droniou, J.: Finite volume schemes for non-coercive elliptic problems with Neumann boundary conditions. IMA J. Numer. Anal. **31**, 61–85 (2011)
3. Coudière, Y., Vila, J.-P., Villedieu, P.: Convergence rate of a finite volume scheme for a two dimensional convection-diffusion problem. ESAIM: Math. Mod. Numer. Anal. **33**(3), 493–516 (1999)
4. Delcourte, S., Omnes, P.: A discrete duality finite volume discretization of the vorticity-velocity-pressure formulation of the 2D Stokes problem on almost arbitrary two-dimensional grids. Numer. Methods PDEs 1–30 (2015)
5. Domelevo, K., Omnes, P.: A finite volume method for the Laplace equation on almost arbitrary two-dimensional grids. ESAIM: Math. Mod. Numer. Anal. **39**(6), 1203–1249 (2005)
6. Goudon, T., Krell, S., Lissoni, G.: DDFV method for Navier-Stokes problem with outflow boundary conditions. Numer. Math. **142**(1), 55–102 (2019)
7. Goudon, T., Krell, S., Lissoni, G.: Non-overlapping Schwarz algorithms for the incompressible Navier-Stokes equations with DDFV discretizations. Tech. report, Univ. Côte d'Azur, Inria, CNRS, LJAD, 2019. https://hal.archives-ouvertes.fr/hal-02448007
8. Halpern, L., Hubert, F.: A finite volume Ventcell-Schwarz algorithm for advection-diffusion equations. SIAM J. Numer. Anal. **52**(3), 1269–1291 (2014)
9. Hermeline, F.: A finite volume method for the approximation of diffusion operators on distorted meshes. J. Comput. Phys. **160**(2), 481–499 (2000)
10. Krell, S.: Stabilized DDFV schemes for Stokes problem with variable viscosity on general 2D meshes. Numer. Methods PDEs **27**(6), 1666–1706 (2011)
11. Krell, S.: Stabilized DDFV schemes for the incompressible Navier-Stokes equations. In: Finite Volumes for Complex Applications VI. Problems & Perspectives, pp. 605–612 (2011)

12. Lions, P.L.: On the Schwarz alternating method. III. A variant for nonoverlapping subdomains. In: Third International Symposium on Domain Decomposition Methods for Partial Differential Equations, SIAM, Philadelphia, PA, 1990, pp. 202–223
13. Schwarz, H.A.: *Über einen grenzübergang durch alternierendes verfahren.* Vierteljahrsschrift der Naturforschenden Gesellschaft in Zurich **15**, 272–286 (1870)

Interface Conditions for Arbitrary Flows in Coupled Porous-Medium and Free-Flow Systems

Elissa Eggenweiler and Iryna Rybak

Abstract Physically consistent interface conditions are important for accurate mathematical modelling and numerical simulation of flow and transport processes in coupled free-flow and porous-medium systems. Traditional coupling concepts are valid for simplified cases only, such as flows parallel to the fluid-porous interface or very specific boundary value problems. This severely limits the range of applications that can be accurately modelled. Evidently, there is a need for more general interface conditions to couple free flow to porous-medium flow. In this paper, we propose new coupling conditions for arbitrary flow directions and periodic porous media. These conditions are derived by the theory of homogenisation and boundary layers and are applicable to general filtration problems. The derived set of coupling conditions are validated by comparison of pore-scale to macroscale numerical simulations.

Keywords Porous media · Free flow · Interface conditions · Finite volumes

MSC (2010) 65N08 · 65N30 · 76D07 · 76S05 · 76M50

1 Introduction

Fluid flows in coupled free-flow and porous-medium domains appear routinely in environmental settings, technical applications and biological systems. The correct specification of interface conditions at the fluid-porous interface is crucial for physically consistent model formulation and accurate numerical simulation of applications. Flows parallel to the porous layer are well studied in the last decades, however the correct choice of coupling conditions for arbitrary flows is still an open question.

E. Eggenweiler (✉) · I. Rybak
Institute of Applied Analysis and Numerical Simulation, University of Stuttgart,
Pfaffenwaldring 57, 70569 Stuttgart, Germany
e-mail: elissa.eggenweiler@mathematik.uni-stuttgart.de

I. Rybak
e-mail: rybak@mathematik.uni-stuttgart.de

R. Klöfkorn et al. (eds.), *Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples*, Springer Proceedings in Mathematics & Statistics 323,
https://doi.org/10.1007/978-3-030-43651-3_31

The Stokes–Darcy problem containing the Stokes equations in the free-flow domain, Darcy's law in the porous medium and an appropriate set of coupling conditions at the fluid-porous interface is the most studied one in the literature, both from modelling and numerical sides [1, 3, 5, 9]. Usually, the conservation of mass, the balance of normal forces and a variant of the Beavers–Joseph interface condition [2, 10, 13] on the tangential velocity are considered. The last condition was postulated for flows parallel to the fluid-porous interface, but it is often applied for other flow regimes, e.g. for industrial filtration [4]. We have shown in [6, 12] that the Beavers–Joseph condition is not valid for such flow situations. Recently, an alternative set of interface conditions has been proposed for the forced infiltration [3], however these interface conditions are restricted to a very specific boundary value problem. The coupling conditions presented in [1] contain several unknown parameters which still need to be fitted. Coupling conditions developed in [11] are not validated for arbitrary flows to the interface. Therefore, a need exists for new interface conditions which are valid for arbitrary flow directions, do not include any fitting parameter and are not restricted to a specific choice of boundary conditions.

In the paper, we propose interface conditions to couple Stokes equations and Darcy's law for flows arbitrary to the interface. These coupling conditions are derived using homogenisation and boundary layer theory. All effective model parameters are computed using geometrical information of the flow system. We validate the developed interface conditions by comparison of the pore-scale resolved model to the macroscale Stokes–Darcy model.

2 Mathematical Models

We consider single-phase incompressible fluid flow at low Reynolds numbers. From the *pore-scale* perspective, the flow in the whole fluid domain (free-flow region $\Omega_{\rm ff}$ and pore space $\Omega^\varepsilon_{\rm pm}$ of the porous medium) is described by the Stokes equations. From the *macroscale* perspective, two different models are applied in the free-flow region $\Omega_{\rm ff}$ and the porous-medium domain $\Omega_{\rm pm}$ together with an appropriate set of interface conditions on the fluid-porous interface Σ (Fig. 1, left).

Pore-Scale Model

The fluid flow in domain $\Omega^\varepsilon = \Omega_{\rm ff} \cup \Omega^\varepsilon_{\rm pm}$ is governed by the non-dimensional steady Stokes equations with the no-slip condition on the boundary of the solid inclusions

$$\begin{aligned} \nabla\cdot\mathbf{u}^\varepsilon &= 0 && \text{in } \Omega^\varepsilon, \\ -\nabla\cdot\mathbf{T}(\mathbf{u}^\varepsilon, p^\varepsilon) &= \mathbf{f} && \text{in } \Omega^\varepsilon, \\ \mathbf{u}^\varepsilon &= \mathbf{0} && \text{on } \partial\Omega^\varepsilon \setminus \partial\Omega\,, \end{aligned} \tag{1}$$

where $\mathbf{u}^\varepsilon$ and p^ε are the non-dimensional pore-scale velocity and pressure, $\mathbf{T}(\mathbf{u}, p) = 2\mathbf{D}(\mathbf{u}) - p\mathbf{I}$ is the stress tensor, $\mathbf{D}(\mathbf{u}) = \frac{1}{2}\left(\nabla\mathbf{u} + (\nabla\mathbf{u})^{\rm T}\right)$ is the rate of strain tensor,

I is the identity tensor, $\mathbf{f}$ is the external force and $\partial\Omega$ is the external boundary of the coupled domain (Fig. 1, left). To obtain a closed formulation of the pore-scale problem (1) boundary conditions on the external boundary $\partial\Omega$ are needed, e.g. (14).

Resolving the detailed pore geometry and solving problem (1) is computationally very expensive for realistic applications. Therefore, macroscale models where the pore-scale information is kept in and reflected in effective parameters are required.

Macroscale Model

The non-dimensional steady Stokes equations describe the flow in the *free-flow domain*

$$\nabla\boldsymbol{\cdot}\mathbf{u}^{\mathrm{ff}} = 0 \quad \text{and} \quad -\nabla\boldsymbol{\cdot}\mathsf{T}(\mathbf{u}^{\mathrm{ff}}, p^{\mathrm{ff}}) = \mathbf{f} \quad \text{in} \;\; \Omega_{\mathrm{ff}}, \tag{2}$$

where $\mathbf{u}^{\mathrm{ff}}$ is the fluid velocity, p^{ff} is the fluid pressure and $\mathbf{f}$ is the external force.

The Darcy equations, also in dimensionless form, characterise the flow through the *porous medium*

$$\nabla\boldsymbol{\cdot}\mathbf{u}^{\mathrm{pm}} = q \quad \text{and} \quad \mathbf{u}^{\mathrm{pm}} = -\mathsf{K}\nabla p^{\mathrm{pm}} \qquad \text{in} \;\; \Omega_{\mathrm{pm}}, \tag{3}$$

where $\mathbf{u}^{\mathrm{pm}}$ is the velocity of the fluid through the porous medium, p^{pm} is the pressure, K is the effective permeability tensor and q is the source term.

The boundary conditions on the external boundary $\partial\Omega$ are

$$\mathbf{u}^{\mathrm{ff}} = \overline{\mathbf{u}}_{\mathrm{ff}} \quad \text{on} \;\; \partial\Omega_{\mathrm{ff}} \setminus \Sigma, \qquad \mathbf{u}^{\mathrm{pm}}\boldsymbol{\cdot}\mathbf{n} = \overline{u}_{\mathrm{pm}} \quad \text{on} \;\; \partial\Omega_{\mathrm{pm}} \setminus \Sigma, \tag{4}$$

where $\mathbf{n}$ is the unit outward normal vector from the domain $\Omega = \Omega_{\mathrm{ff}} \cup \Omega_{\mathrm{pm}}$ on its boundary and the functions $\overline{\mathbf{u}}_{\mathrm{ff}}$ and $\overline{u}_{\mathrm{pm}}$ are given.

To complete the mathematical description of the coupled problem, interface conditions on Σ are required. The most commonly used interface conditions, both for mathematical modelling and numerical simulation, are the *conservation of mass*

$$\mathbf{u}^{\mathrm{ff}}\boldsymbol{\cdot}\mathbf{n}_{\Sigma} = \mathbf{u}^{\mathrm{pm}}\boldsymbol{\cdot}\mathbf{n}_{\Sigma} \qquad \text{on} \;\; \Sigma, \tag{5}$$

the *balance of normal forces*

$$-\,\mathbf{n}_{\Sigma}\boldsymbol{\cdot}\mathsf{T}\left(\mathbf{u}^{\mathrm{ff}}, p^{\mathrm{ff}}\right)\boldsymbol{\cdot}\mathbf{n}_{\Sigma} = p^{\mathrm{pm}} \qquad \text{on} \;\; \Sigma, \tag{6}$$

and the *Beavers–Joseph* condition [2] for the tangential component of velocity

$$(\mathbf{u}^{\mathrm{ff}} - \mathbf{u}^{\mathrm{pm}})\boldsymbol{\cdot}\boldsymbol{\tau}_{\Sigma} + \frac{2\sqrt{\mathsf{K}}}{\alpha}\mathbf{n}_{\Sigma}\boldsymbol{\cdot}\mathsf{D}\left(\mathbf{u}^{\mathrm{ff}}\right)\boldsymbol{\cdot}\boldsymbol{\tau}_{\Sigma} = 0 \qquad \text{on} \;\; \Sigma, \tag{7}$$

where $\mathbf{n}_{\Sigma}$ and $\boldsymbol{\tau}_{\Sigma}$ are the unit normal and tangential vectors at the interface Σ accordingly (Fig. 1, left), and $\alpha > 0$ is the Beavers–Joseph parameter which needs to be determined for each flow scenario.

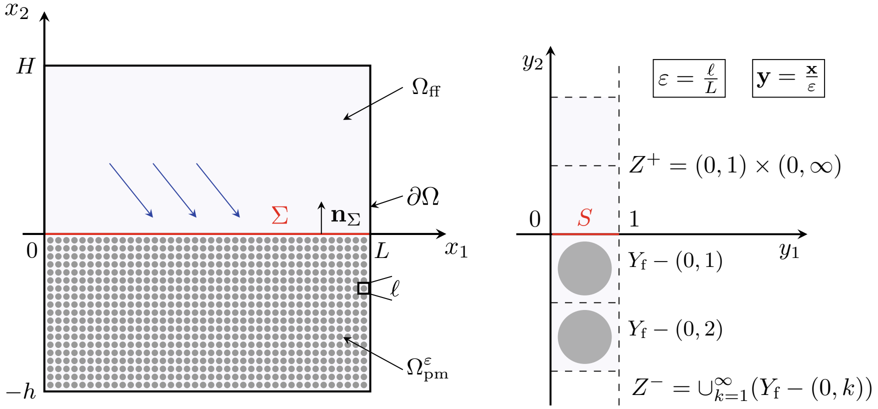

x_2
H
Ω_{ff}
$\partial\Omega$
Σ
$\mathbf{n}_\Sigma$
0
L
x_1
ℓ
$\Omega^\varepsilon_{\mathrm{pm}}$
$-h$
y_2
$\varepsilon = \frac{\ell}{L}$
$\mathbf{y} = \frac{\mathbf{x}}{\varepsilon}$
$Z^+ = (0,1) \times (0,\infty)$
0
S
1
$Y_{\mathrm{f}} - (0,1)$
y_1
$Y_{\mathrm{f}} - (0,2)$
$Z^- = \cup_{k=1}^{\infty}(Y_{\mathrm{f}} - (0,k))$

Fig. 1 Schematic pore-scale geometry (left) and boundary layer stripe $Z^{\mathrm{bl}} = Z^+ \cup S \cup Z^-$ (right)

3 Homogenisation and Boundary Layers

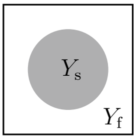
To derive new interface conditions for the macroscale coupled problem, we use the theory of homogenisation and boundary layers [8, 9]. Therefore, we consider porous media with periodic arrangement of solid grains (Fig. 1, left). We define the unit cell $Y = (0, 1)^d$ with respect to the smallest repetitive unit of the pore geometry, where d is the number of space dimensions. In the manuscript, we consider $d = 2$. Each unit cell consists of the solid part Y_{s} and the fluid part $Y_{\mathrm{f}} = Y \setminus Y_{\mathrm{s}}$ (Table 1). We introduce the scale separation parameter $\varepsilon = \ell/L$ which denotes the ratio between the characteristic pore size ℓ to the length L of the domain (Fig. 1). The porous structure is defined through periodic repetition of the scaled unit cell εY.

We follow the classical procedure of homogenisation [8] and study the behaviour of the solutions to the pore-scale problem (1) when $\varepsilon \to 0$. We get the following asymptotic expansions for the velocity $\mathbf{u}^\varepsilon$ and the pressure p^ε in the pore space $\Omega^\varepsilon_{\mathrm{pm}}$:

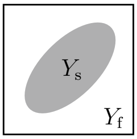
$$\mathbf{u}^\varepsilon(\mathbf{x}) = \varepsilon^2 \mathbf{u}_0(\mathbf{x}, \mathbf{y}) + O(\varepsilon^3), \quad p^\varepsilon(\mathbf{x}) = p_0(\mathbf{x}, \mathbf{y}) + \varepsilon p_1(\mathbf{x}, \mathbf{y}) + O(\varepsilon^2),$$

$$\mathbf{u}_0(\mathbf{x}, \mathbf{y}) = -\sum_{j=1}^{2} \mathbf{w}^j(\mathbf{y}) \frac{\partial p^{\mathrm{pm}}(\mathbf{x})}{\partial x_j}, \quad p_0(\mathbf{x}, \mathbf{y}) = p^{\mathrm{pm}}(\mathbf{x}), \quad p_1(\mathbf{x}, \mathbf{y}) = -\sum_{j=1}^{2} \pi^j(\mathbf{y}) \frac{\partial p^{\mathrm{pm}}(\mathbf{x})}{\partial x_j},$$

Table 1 Boundary layer constants and permeability values for different porous-medium geometries



k_{11}	$1.99 \cdot 10^{-2}$
k_{12}	$1.37 \cdot 10^{-9}$
$M_1^{1,\mathrm{bl}}$	$-4.77 \cdot 10^{-2}$
$M_1^{2,\mathrm{bl}}$	$-5.25 \cdot 10^{-7}$



k_{11}	$1.23 \cdot 10^{-2}$
k_{12}	$2.69 \cdot 10^{-3}$
$M_1^{1,\mathrm{bl}}$	$-3.10 \cdot 10^{-2}$
$M_1^{2,\mathrm{bl}}$	$-3.33 \cdot 10^{-3}$

where $\mathbf{x} \in \Omega^{\varepsilon}_{\text{pm}}$, $\mathbf{y} = \mathbf{x}/\varepsilon$, $\mathbf{u}_0$ and p_1 are 1-periodic in $\mathbf{y}$.

The pair $\{\mathbf{w}^j, \pi^j\}$ are the solutions to the following cell problems for $j = 1, 2$:

$$\begin{gathered}\Delta_{\mathbf{y}} \mathbf{w}^j - \nabla_{\mathbf{y}} \pi^j = -\mathbf{e}_j, \quad \nabla_{\mathbf{y}} \cdot \mathbf{w}^j = 0 \text{ in } Y_{\text{f}}, \quad \int_{Y_{\text{f}}} \pi^j \, \mathrm{d}\mathbf{y} = 0, \\ \mathbf{w}^j = \mathbf{0} \text{ on } \partial Y_{\text{f}} \setminus \partial Y \ \text{ and } \ \{\mathbf{w}^j, \pi^j\} \text{ is 1-periodic in } \mathbf{y},\end{gathered} \tag{8}$$

where $\mathbf{e}_j$ denotes the standard jth unit vector.

As a result of the homogenisation, Darcy's law (3) is obtained in the porous medium Ω_{pm}. The effective permeability tensor is given by $\mathsf{K} = \varepsilon^2 (k_{ij})_{i,j=1,2}$, where

$$k_{ij} = \int_{Y_{\text{f}}} w_i^j \, \mathrm{d}\mathbf{y}, \quad i, j = 1, 2. \tag{9}$$

Derivation of Interface Conditions

To derive effective interface conditions on Σ we need an approximation of the velocity $\mathbf{u}^{\varepsilon}$ and the pressure p^{ε} in the whole flow region Ω^{ε}. Therefore, we start with the following one

$$\begin{aligned} \mathbf{u}^{\varepsilon} &\approx \mathcal{H}(x_2)\, \mathbf{u}^{\text{ff}} + \mathcal{H}(-x_2) \left(-\varepsilon^2 \sum_{j=1}^{2} \mathbf{w}^j \frac{\partial p^{\text{pm}}}{\partial x_j} \right) && \text{in } \Omega^{\varepsilon}, \\ p^{\varepsilon} &\approx \mathcal{H}(x_2)\, p^{\text{ff}} + \mathcal{H}(-x_2) \left(p^{\text{pm}} - \varepsilon \sum_{j=1}^{2} \pi^j \frac{\partial p^{\text{pm}}}{\partial x_j} \right) && \text{in } \Omega^{\varepsilon}, \end{aligned} \tag{10}$$

where $\mathcal{H}$ is the Heaviside function. However, this approximation does not provide continuity of velocity trace across the interface Σ. Therefore, we establish continuity of traces with the help of the boundary layer problems for $j = 1, 2$ similar to [3, 9]:

$$\begin{aligned} -\Delta_{\mathbf{y}}\, \boldsymbol{\beta}^{j,bl} + \nabla_{\mathbf{y}}\, \omega^{j,bl} = \mathbf{0} \quad \text{and} \quad \nabla_{\mathbf{y}} \cdot \boldsymbol{\beta}^{j,bl} &= 0 \quad \text{in } Z^+ \cup Z^-, \\ [\![\boldsymbol{\beta}^{j,bl}]\!]_S = k_{2j} \mathbf{e}_2 - \mathbf{w}^j & \qquad \text{on S}, \\ [\![(\nabla_{\mathbf{y}}\, \boldsymbol{\beta}^{j,bl} - \omega^{j,bl}\, \mathbf{I}) \mathbf{e}_2]\!]_S = -(\nabla_{\mathbf{y}}\, \mathbf{w}^j - \pi^j\, \mathbf{I}) \mathbf{e}_2 & \qquad \text{on S}, \\ \boldsymbol{\beta}^{j,bl} = \mathbf{0} \ \text{ on fluid-solid interface}, \quad \{\boldsymbol{\beta}^{j,bl}, \omega^{j,bl}\} & \text{ is 1-periodic in } y_1. \end{aligned} \tag{11}$$

The boundary layer problem (11) is constructed on the infinite stripe $Z^{bl} = Z^+ \cup S \cup Z^-$ (Fig. 1, right) and we define $[\![\boldsymbol{\beta}]\!]_S = \boldsymbol{\beta}(\cdot, +0) - \boldsymbol{\beta}(\cdot, -0)$. The boundary layer velocity $\boldsymbol{\beta}^{j,bl}$ and pressure $\omega^{j,bl}$ stabilise exponentially towards zero when $y_2 \to -\infty$ and towards the constants $\mathbf{M}^{j,bl}$, $M_{\omega}^{j,bl}$ when $y_2 \to +\infty$, e.g. [3]:

$$\mathbf{M}^{j,bl} = (M_1^{j,bl}, 0) = \left(\int_0^1 \beta_1^{j,bl}(y_1, +0)\, \mathrm{d}y_1, 0 \right), \quad M_{\omega}^{j,bl} = \int_0^1 \omega^{j,bl}(y_1, +0)\, \mathrm{d}y_1.$$

The new, improved approximation of velocity and pressure in Ω^ε is given by

$$
\begin{aligned}
\mathbf{u}^\varepsilon &\approx \mathcal{H}(x_2)\,\mathbf{u}^{\mathrm{ff}} + \mathcal{H}(-x_2)\left(-\varepsilon^2 \sum_{j=1}^{2} \mathbf{w}^j \frac{\partial p^{\mathrm{pm}}}{\partial x_j}\right) + \varepsilon^2 \sum_{j=1}^{2} \boldsymbol{\beta}^{j,bl} \frac{\partial p^{\mathrm{pm}}}{\partial x_j}\bigg|_{\Sigma}, \\
p^\varepsilon &\approx \mathcal{H}(x_2)\,p^{\mathrm{ff}} + \mathcal{H}(-x_2)\left(p^{\mathrm{pm}} - \varepsilon \sum_{j=1}^{2} \pi^j \frac{\partial p^{\mathrm{pm}}}{\partial x_j}\right) + \varepsilon \sum_{j=1}^{2} \omega^{j,bl} \frac{\partial p^{\mathrm{pm}}}{\partial x_j}\bigg|_{\Sigma}.
\end{aligned}
\tag{12}
$$

The next step is to correct the counter-flow effects resulted from adding the boundary layer terms to approximations (10). Thus, we subtract the properly scaled boundary layer constants in the free-flow region Ω_{ff} from approximations (12). To obtain an accurate approximation of the solution to problem (1), additional corrections of the velocity and pressure (12) in terms of boundary layer problems are needed.

Using homogenisation and the discussed boundary layer problems we recovered the conservation of mass condition (5), the balance of normal forces (6) and derived a *new interface condition* for the tangential component of velocity

$$
\mathbf{u}^{\mathrm{ff}} \cdot \boldsymbol{\tau}_\Sigma = \sum_{j=1}^{2} \varepsilon^2 \mathbf{M}^{j,bl} \frac{\partial p^{\mathrm{pm}}}{\partial x_j} \cdot \boldsymbol{\tau}_\Sigma \qquad \text{on } \Sigma. \tag{13}
$$

Equation (13) gives a relation between the tangential free-flow and porous-medium velocity $\mathbf{u}^{\mathrm{ff}} \cdot \boldsymbol{\tau}_\Sigma = \mathbf{u}^{\mathrm{pm}} \cdot \boldsymbol{\tau}_\Sigma + C$ with $C = \sum_{j=1}^{2} \varepsilon^2 (M_1^{j,bl} + k_{1j}) \frac{\partial p^{\mathrm{pm}}}{\partial x_j}$ for $\boldsymbol{\tau}_\Sigma = \mathbf{e}_1$. The conditions (5), (6), (13) are valid for arbitrary flow directions to the interface.

4 Model Validation

To validate the interface conditions (5), (6), (13) we compare macroscale to pore-scale numerical simulations. We consider the free-flow domain $\Omega_{\mathrm{ff}} = (0, 1) \times (0, 0.5)$, the porous medium $\Omega_{\mathrm{pm}} = (0, 1) \times (-0.5, 0)$ and the interface $\Sigma = (0, 1) \times \{0\}$. The porous medium is periodic and we study two geometrical configurations with 20 × 10 equidistantly aligned circular and elliptic solid inclusions (Table 1, Fig. 2a, b). For each porous-medium configuration we solve the pore-scale problem (1) with the following boundary conditions

$$
\mathbf{u}^\varepsilon = (0.5, 0) \quad \text{on } \{x_2 = H\}, \qquad \mathbf{u}^\varepsilon = \mathbf{0} \quad \text{on } \partial\Omega \setminus \{x_2 = H\}. \tag{14}
$$

The computations of problem (1), (14) are performed with FreeFEM++ [7] using the Taylor–Hood (P2/P1) finite elements and a mesh with approx. 500 000 elements.

The corresponding boundary conditions on the external boundary for the macroscale problem (2), (3) in the absence of sources ($\mathbf{f} = \mathbf{0}$, $q = 0$) read

$$\overline{\mathbf{u}}_{\text{ff}} = (0.5, 0) \quad \text{on} \quad \{x_2 = H\}, \qquad \overline{\mathbf{u}}_{\text{ff}} = \mathbf{0} \quad \text{on} \quad \partial\Omega_{\text{ff}} \setminus (\{x_2 = H\} \cup \Sigma),$$
$$\overline{u}_{\text{pm}} = 0 \quad \text{on} \quad \partial\Omega_{\text{pm}} \setminus \Sigma. \tag{15}$$

On the fluid-porous interface Σ we apply both, the traditional interface conditions (5)–(7) and the newly derived conditions (5), (6), (13).

The finite volume method with staggered grids [14] and the grid size $h = 1/100$ is used to discretise the coupled Stokes–Darcy problem. The effective permeability tensor $\mathbf{K}$ is obtained for each geometrical configuration by solving the cell problems (8) and applying formula (9). The boundary layer constants needed in Eq. (13) are computed numerically using a cut-off stripe as proposed in [3]. The cell and the boundary layer problems are solved using FreeFEM++ [7]. The fluid part of the unit cell is partitioned into approx. 35 000 elements, the one of the stripe into approx. 125 000 elements. The boundary layer constants and permeability values for the considered geometrical configurations are presented in Table 1.

The pore-scale velocity fields for circular and elliptic inclusions are presented in Fig. 2a, b. To evaluate the validity of the proposed interface condition (13) we provide cross-sections for the tangential velocity component u_1 for both geometrical configurations in Fig. 2c, d, e.

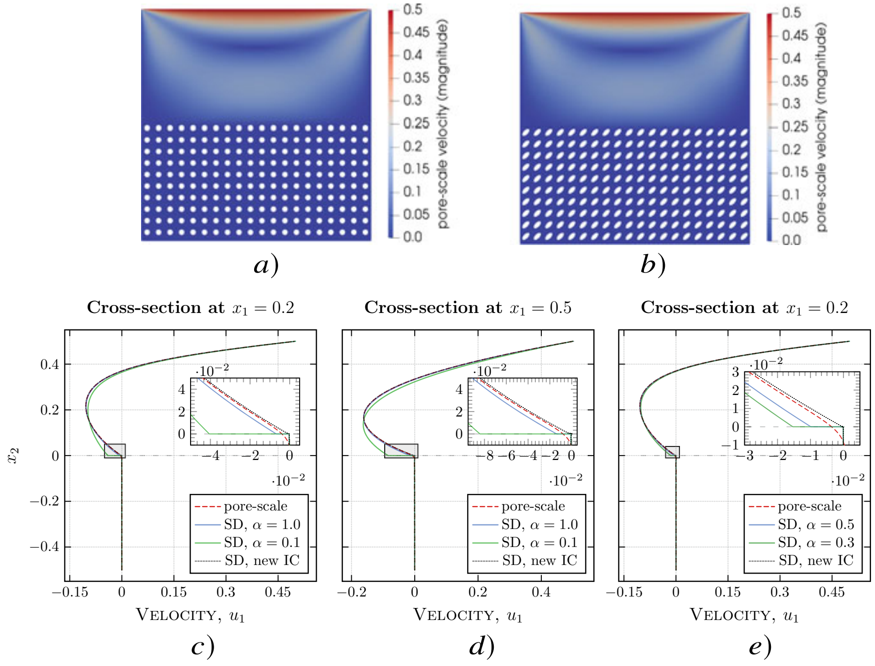


Fig. 2 Pore-scale velocity fields (**a**), (**b**) and comparison of pore-scale to macroscale simulation results for circular (**c**), (**d**) and elliptic (**e**) inclusions

We observe (Fig. 2c, d, e) that the simulation results of the Stokes–Darcy model with the newly derived interface conditions (profile: SD, new IC) are in good agreement with the pore-scale simulations (profile: pore-scale) for all geometries and cross-sections. The classical interface conditions (5)–(7) are very sensitive to the choice of the Beavers–Joseph parameter α (see e.g. profile: SD, $\alpha = 0.1$) and it is not clear which α to choose [6].

5 Conclusion

In this paper, new interface conditions for arbitrary flows to porous media are proposed. These coupling conditions are derived using the theory of homogenisation and boundary layer correctors. The Stokes–Darcy model with new interface conditions is validated numerically against the pore-scale resolved model. The advantage of the proposed interface conditions is that they are rigorously derived, more accurate than the classical ones and no fitting of the Beavers–Joseph parameter α is needed. Computational complexity of the new interface conditions is practically the same as of the classical conditions. The cell problems (8) have to be solved to obtain the effective permeability K for both approaches and the solution of the additional boundary layer problems (11) to determine $\mathbf{M}^{j,bl}$ is computationally cheap.

Acknowledgements The work is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—Project Number 327154368—SFB 1313.

References

1. Angot, P., Goyeau, B., Ochoa-Tapia, J.A.: Asymptotic modeling of transport phenomena at the interface between a fluid and a porous layer: jump conditions. Phys. Rev. E **95**, 063302 (2017)
2. Beavers, G.S., Joseph, D.D.: Boundary conditions at a naturally permeable wall. J. Fluid Mech. **30**, 197–207 (1967)
3. Carraro, T., Goll, C., Marciniak-Czochra, A., Mikelić, A.: Effective interface conditions for the forced infiltration of a viscous fluid into a porous medium using homogenization. Comput. Methods Appl. Mech. Engrg. **292**, 195–220 (2015)
4. Discacciati, M., Gerardo-Giorda, L.: Optimized Schwarz methods for the Stokes–Darcy coupling. IMA J. Numer. Anal. **38**, 1959–1983 (2018)
5. Discacciati, M., Quarteroni, A.: Navier–Stokes/Darcy coupling: modeling, analysis, and numerical approximation. Rev. Mat. Complut. **22**, 315–426 (2009)
6. Eggenweiler, E., Rybak, I.: Unsuitability of the Beavers–Joseph interface condition for filtration problems. J. Fluid Mech. (2020). https://doi.org/10.1017/jfm.2020.194
7. Hecht, F.: New development in FreeFem++. J. Numer. Math. **20**, 251–265 (2012)
8. Hornung, U.: Homogenization and Porous Media. Springer, Berlin (1997)
9. Jäger, W., Mikelić, A.: On the boundary conditions at the contact interface between a porous medium and a free fluid. Ann. Scuola Norm. Sup. Pisa Cl. Sci. **23**, 403–465 (1996)
10. Jäger, W., Mikelić, A.: Modeling effective interface laws for transport phenomena between an unconfined fluid and a porous medium using homogenization. Transp. Porous Media **78**, 489–508 (2009)

11. Lācis, U., Bagheri, S.: A framework for computing effective boundary conditions at the interface between free fluid and a porous medium. J. Fluid Mech. **812**, 866–889 (2017)
12. Rybak, I., Schwarzmeier, C., Eggenweiler, E., Rüde, U.: Validation and calibration of coupled porous-medium and free-flow problems using pore-scale resolved models. Comput. Geosci. (submitted) (arXiv:1906.06884v2) (2019)
13. Saffman, P.G.: On the boundary condition at the surface of a porous medium. Stud. Appl. Math. **50**, 93–101 (1971)
14. Versteeg, H., Malalasekra, W.: An Introduction to Computational Fluid Dynamics: The Finite Volume Method. Prentice Hall (2007)

On the Convergence Rate of the Dirichlet-Neumann Iteration for Coupled Poisson Problems on Unstructured Grids

Morgan Görtz and Philipp Birken

Abstract We consider thermal fluid structure interaction with a partitioned approach, where typically, a finite volume and a finite element code would be coupled. As a model problem, we consider two coupled Poisson problems with heat conductivities λ_1, λ_2 in one dimension on intervals of length l_1 and l_2. Hereby, we consider linear discretizations on arbitrary meshes, such as finite volumes, finite differences, finite elements. For these, we prove that the convergence rate of the Dirichlet-Neumann iteration is given by $\lambda_1 l_2/\lambda_2 l_1$ and is thus independent of discretization and mesh.

Keywords Dirichlet-Neumann iteration · Thermal Fluid-Structure-Interaction · Partitioned Approach · Convergence Rate

1 Introduction

We are concerned with thermal fluid structure interaction, also called conjugate heat transfer. This occurs in many applications, for example gas quenching, which is an industrial heat treatment of metal workpieces [8] or the cooling of rocket nozzles [10, 11]. Here, we follow a partitioned approach [4], where different codes for the sub-problems are reused and the coupling is done by a master program which calls interface functions of the segregated codes. This allows to reuse existing software for each sub-problem, in contrast to a monolithic approach, where a new code is tailored for the coupled equations.

M. Görtz (✉)
Fraunhofer-Chalmers Centre, Chalmers Science Park,
41288 Göteborg, Sweden
e-mail: morgan.gortz@fcc.chalmers.se

P. Birken (✉)
Centre for the Mathematical Sciences, Lund University,
Box 118, 22100 Lund, Sweden
e-mail: philipp.birken@na.lu.se

R. Klöfkorn et al. (eds.), *Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples*, Springer Proceedings in Mathematics & Statistics 323,
https://doi.org/10.1007/978-3-030-43651-3_32

The standard solution method within a partitioned approach is the Dirichlet-Neumann iteration. To satisfy coupling conditions at the interface, the subsolvers are iterated by providing Dirichlet and Neumann data for the other solver in a sequential manner, giving rise to its name. In particular for the interaction of a compressible flow with a structure, the default implementation would be a finite volume method for the fluid and a finite element method for the structure.

The convergence rate for the interaction of a flexible structure with a fluid has been analyzed in [14]. There, the added mass effect is proven to be dependent on the step size for compressible flows and independent for incompressible flows. Furthermore, for incompressible fluids it is known that the ratio of densities of the materials plays an important role [1, 3]. Finally, the Dirichlet-Neumann iteration was reported to be a very fast solver for thermal coupling of the compressible Navier-Stokes equations with a nonlinear heat equation to model steel [2].

A simplified model of this interaction that allows for analysis are two coupled linear heat equations with constant material coefficients that jump across the interface. A semidiscrete Dirichlet-Neumann method using implicit Euler in time was analyzed for domains of different length in [9]. Using Fourier analysis, it was proved that for large Δt, the convergence rate is approximately the quotient of heat conductivities, but that for small Δt, this has to be multiplied by the square root of the quotient of thermal diffusitivites. A fully discrete analysis with equidistant mesh widths and domains of equal size was performed in [12, 13], which shows two important differences to the semidiscrete one. Firstly, the Fourier analysis breaks down for $c = \Delta t/\Delta x^2 < 1$. In the limit $c \to 0$, the fully discrete analysis shows that interestingly, the limit of the convergence rate for a finite volume finite element coupling is zero, whereas for coupled finite element methods, it is the quotient another set of material parameters. Secondly, the convergence rate in the limit $c \to \infty$, differs from the semidiscrete analysis as well: It is now the aspect ratio in the mesh at the interface times the quotient of heat conductivities. Finally, we would like to point out that the numerical results do not show an influence of the size of the domains. This is notable, since this was proven to be the case for waveform variants of this method [5].

Summarizing, this leaves a number of questions open. In this article we take a step back and consider two coupled Laplace problems with a jump in heat conductivities at the interface. As it turns out, the convergence rate of the discrete Dirichlet-Neumann iteration in 1D can be completely analyzed, almost irrespective of discretization employed and on any mesh! While the quotient of heat conductivities remains a crucial number, the ratio of lengths of the domains suddenly plays a role [6].

The problem we analyze is the one dimensional transmission problem. Given are two connected intervals $\Omega_1 = [a, x_\Gamma]$ and $\Omega_2 = [x_\Gamma, b]$. To make the notation easier extendable to the multidimensional case, we also introduce the interface $\Gamma = \{x_\Gamma\}$. The transmission problem in one dimension is given by

$$\lambda_i u_i''(x) = f_i(x), x \in \Omega_i, \tag{1}$$

$$u_1(x) = u_2(x), x \in \Gamma, \tag{2}$$

$$\lambda_1 u_1'(x) = \lambda_2 u_2'(x), x \in \Gamma, \tag{3}$$

$$u_1(a) = 0, u_2(b) = 0,\ i = 1, 2. \tag{4}$$

Here, λ_i are the respective heat conductivities and the $f_i(x)$ ares forcing functions.

2 Dirichlet-Neumann Iteration

The algorithm starts with an initial guess of $u_2(x_\Gamma)$. With this approximation we solve a Dirichlet problem on Ω_1, which will give us a function u_1^1. We then solve the problem on Ω_2 with a Neumann condition on Γ, given by $\frac{\lambda_1}{\lambda_2}(u_1^1)'(x_\Gamma)$. The solution to this problem us called u_2^1. This process continues with u_1^k and u_2^k being the functions for the kth iteration. The continuous Dirichlet-Neumann iteration can thus be written as:

Given $u_2^0(x_\Gamma)$, solve in sequence the following problems:

$$\begin{aligned} \lambda_1(u_1^{k+1})''(x) &= f_1(x),\ x \in [a, x_\Gamma] \\ u_1^{k+1}(a) = u_1(a) \text{ and } u_1^{k+1}(x_\Gamma) &= u_2^k(x_\Gamma) \end{aligned} \tag{5}$$

$$\begin{aligned} \lambda_2(u_2^{k+1})''(x) &= f_2(x),\ x \in [x_\Gamma, b] \\ (u_2^{k+1})'(x_\Gamma) &= \frac{\lambda_1}{\lambda_2}(u_1^k)'(x_\Gamma) \text{ and } u_2^{k+1}(b) = u_2(b). \end{aligned} \tag{6}$$

Each iteration requires us to solve two Poisson equations, one with Dirichlet boundary conditions and one with a Dirichlet and a Neumann boundary condition.

Now we formulate a discrete version on an abstract level. A linear discretization of the Dirichlet problem on Ω_1 can be written as the following linear system:

$$\lambda_1 A^{(1)} \bar{u}_1^{k+1} = b^{(1)} u_1(a) + \bar{f}^{(1)} - \lambda_1 A_\Gamma^{(1)} u_{2_\Gamma}^k, \tag{7}$$

where $\bar{u}_1^{k+1}$ and $\bar{f}^{(1)}$ are the discrete representations of u_1^{k+1} and f_1. Similarly, a linear discretisation of the Neumann problem on Ω_2 has an additional unknown on the interface and can be written as:

$$\lambda_2 \begin{bmatrix} A^{(2)} & A_\Gamma^{(2)} \\ d^{(2)} & d_\Gamma^{(2)} \end{bmatrix} \begin{bmatrix} \bar{u_2}^{k+1} \\ u_\Gamma^{k+1} \end{bmatrix} = \begin{bmatrix} b^{(2)} u(b) + \bar{f}^{(2)} \\ f_\Gamma^{(2)} - f_\Gamma^{(1)} + \lambda_1 d^{(1)} \bar{u}_1^{k+1} + \lambda_1 d_\Gamma^{(1)} u_\Gamma^k, \end{bmatrix} \tag{8}$$

where $\bar{u}_2^{k+1}$ and $\bar{f}^{(2)}$ are the discrete representation of u_2^{k+1} and f_2. Additionally, we assume that the discretization has a nodal value associated with the boundary and call that u_Γ^k. Furthermore,

$$d^{(1)} u_\Gamma^k + d_\Gamma^{(1)} \bar{u}_1^{k+1} \approx (u_1^{k+1})'(x_\Gamma) \tag{9}$$

and

$$d^{(2)}u_\Gamma^{k+1} + d_\Gamma^{(2)}\bar{u}_2^{k+1} \approx (u_2^{k+1})'(x_\Gamma) \tag{10}$$

are discrete approximations of the derivatives at the interface.

Combining these discretisation into one linear system results in a description of one step of the discrete Dirichlet-Neumann iterion. Given u_Γ^k, solve:

$$\begin{bmatrix} \lambda_1 A^{(1)} & 0 & 0 \\ 0 & \lambda_2 A^{(2)} & \lambda_2 A_\Gamma^{(2)} \\ -\lambda_1 d^{(1)} & \lambda_2 d^{(2)} & \lambda_2 d_\Gamma^{(2)} \end{bmatrix} \begin{bmatrix} \bar{u}_1^{k+1} \\ \bar{u}_2^{k+1} \\ u_\Gamma^{k+1} \end{bmatrix} = \begin{bmatrix} b^{(1)}u_1(a) - \lambda_1 A_\Gamma^{(1)} u_\Gamma^k + \bar{f}^{(1)} \\ b^{(2)}u_2(b) + \bar{f}^{(2)} \\ f_\Gamma^{(2)} - f_\Gamma^{(1)} + \lambda_1 d_\Gamma^{(1)} u_\Gamma^k \end{bmatrix}. \tag{11}$$

Here, we consider in several different discretizations. Firstly, a finite volume discretization based on control volumes $C_k = [x_{k-1/2}, x_{k+1/2}]$ on which we have

$$\int_{C_l} (\lambda u'(x))' dx = \lambda u'(x_{k+1/2}) - \lambda u'(x_{k-1/2}).$$

We then use the numerical flux

$$\lambda u'(x_{k+1/2}) \approx F_{k+0.5} = \lambda \frac{u_{k+1} - u_k}{\Delta x_k},$$

which gives a second order approximation, see Fig. 1. The derivative approximations (9) and (10) are then given by integrating over the first cell:

$$\int_a^{x_{1/2}} u'' dx = u'(x_{1/2}) - u'(a) = \int_a^{x_{1/2}} f dx.$$

Approximating $u'(x_{1/2})$ by the numerical flux function gives

$$d^{(1)} = (1, 0, \ldots, 0)^T / \Delta x, \quad d_\Gamma^{(1)} = -1/\Delta x_1.$$

Secondly, we consider standard linear finite elements. There, the derivative approximations are obtained from the weak form by integration by parts and integrating over $[a, a + \Delta x]$. This results in the exact same algebraic expression as above.

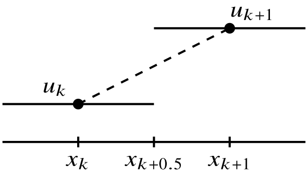


Fig. 1 The numerical flux, $F_{k+0.5}$, is the slope of the dashed line

Finally, there are second order central differences, with the derivative approximations given by a onesided second order difference.

3 Analysis of Convergence Rate

To find the convergence rate of the Dirichlet-Neumann algorithm we first have to determine the iteration matrix. Instead of analyzing the matrix in (11), it is better to reformulate the problem in terms of the interface unknown u_Γ^k only. In the one dimensional case considered here, this reduces the iteration matrix to a scalar, making the computation of the spectral radius trivial. We thus have an iteration

$$u_\Gamma^{k+1} = \Sigma u_\Gamma^k + \alpha, \ \alpha \text{ independent on } k. \tag{12}$$

To get an expression for u_Γ^{k+1}, we start by taking the Schur compliment of (11) with respect to this unknown. With this we get an equation for u_Γ^{k+1}:

$$Su_\Gamma^{k+1} = f_\Gamma^{(2)} - f_\Gamma^{(1)} + \lambda_1 d_\Gamma^{(1)} u_\Gamma^k - (-\lambda_1 d^{(1)})(\lambda_1 A^{(1)})^{-1}(b^{(1)} u_1(a) + \bar{f}^{(1)} - \lambda_1 A_\Gamma^{(1)} u_\Gamma^k)$$

$$-(\lambda_2 d^{(2)})(\lambda_2 A^{(2)})^{-1}(b^{(2)} u_2(b) + \bar{f}^{(2)}),$$

$$S = \lambda_2 d_\Gamma^{(2)} - (\lambda_2 d^{(2)})(\lambda_2 A^{(2)})^{-1}(\lambda_2 A_\Gamma^{(2)}).$$

Next we put everything that does not dependent on k into a new constant α_1.

$$Su_\Gamma^{k+1} = \lambda_1 (d_\Gamma^{(1)} - (d^{(1)})(A^{(1)})^{-1}(A_\Gamma^{(1)}))u_\Gamma^k + \alpha_1$$

$$\Rightarrow u_\Gamma^{k+1} = \frac{\lambda_1}{\lambda_2} \frac{d_\Gamma^{(1)} - (d^{(1)})(A^{(1)})^{-1}(A_\Gamma^{(1)})}{d_\Gamma^{(2)} - (d^{(2)})(A^{(2)})^{-1}(A_\Gamma^{(2)})} u_\Gamma^k + \alpha,$$

where α is again independent of k. From this we get the iteration matrix

$$\Sigma = \frac{\lambda_1}{\lambda_2} \frac{d_\Gamma^{(1)} - d^{(1)}(A^{(1)})^{-1} A_\Gamma^{(1)}}{d_\Gamma^{(2)} - d^{(2)}(A^{(2)})^{-1} A_\Gamma^{(2)}}. \tag{13}$$

The convergence rate is then $\mu = |\Sigma|$.

First, we present a proof for the asymptotic convergence rate under the weak assumption that we have convergent discretizations and additional assumptions on the discretization at the boundary.

Theorem 1 (Aymptotic convergence rate) *Consider linear discretizations for the problems* (5)–(6) *with* n_1*, respectively* n_2 *unknowns. Let these be convergent in spaces*

$\mathcal{V}_1$ and $\mathcal{V}_2$. Let furthermore the boundary conditions at the interface be enforced strongly and the derivative approximations (9)–(10) *be convergent. If these discretisations are used in the Dirichlet-Neumann algorithm,* (11)*, then the asymptotic convergence rate is*

$$\lim_{n_1,n_2\to\infty} \mu = \left|\frac{\lambda_1 l_2}{\lambda_2 l_1}\right|.$$

Proof The convergence rate is given by (13). We first analyze the nominator:

$$d_\Gamma^{(1)} - d^{(1)}(A^{(1)})^{-1}A_\Gamma^{(1)} = d_\Gamma^{(1)} + d^{(1)}\bar{y},$$

with $A^{(1)}\bar{y} = -A_\Gamma^{(1)}$. By (7), $A^{(1)}\bar{y} = -A_\Gamma^{(1)}$, is a discretisation of:

$$y''(x) = 0,\ \ y(a) = 0 \text{ and } y(x_\Gamma) = 1,$$

which has the solution $y(x) = (x-a)/l_1$ with derivative $y'(x) = 1/l_1$. Since the discretization is convergent, the discrete approximation y^h converges to y. We further note that

$$d_\Gamma^{(1)} + d^{(1)}\bar{y} = d_\Gamma^{(1)} y^h(x_\Gamma) + d^{(1)}\bar{y} \tag{14}$$

is a convergent discrete approximation of $y'(x_\Gamma)$. Thus

$$\lim_{n_1\to\infty} d_\Gamma^{(1)} y^h(x_\Gamma) + d^{(1)}\bar{y} = 1/l_1.$$

Next we analyze the denominator:

$$d_\Gamma^{(2)} - d^{(2)}(A^{(2)})^{-1}A_\Gamma^{(2)} = d_\Gamma^{(2)} + d^{(2)}\bar{z},$$

with $A^{(2)}\bar{z} = -A_\Gamma^{(2)}$. We discretize the following differential equation with the Neumann discretisation, see (8):

$$z''(x) = 0,\ z'(x_\Gamma) = -\frac{1}{l_2},\ z(b) = 0,$$

where the exact solution is $z(x) = \frac{b-x}{l_2}$ with $z'(x) = 1/l_2$. The corresponding system defining the coefficients $\bar{z}$ of this discretisation is:

$$\begin{bmatrix} A^{(2)} & A_\Gamma^{(2)} \\ d^{(2)} & d_\Gamma^{(2)} \end{bmatrix} \begin{bmatrix} \bar{z} \\ z^h(x_\Gamma) \end{bmatrix} = \begin{bmatrix} 0 \\ -\frac{1}{l_2} \end{bmatrix},$$

where we see that the first equation is the definition of $\bar{z}$. Because this is a convergent discretization and the last equation is a convergent approximation of $z'(x_\Gamma)$:

$$\lim_{n_2 \to \infty} d_\Gamma^{(2)} + d^{(2)} \bar{g} = -\frac{1}{l_2}. \tag{15}$$

Combining (13), (14) and (15) finishes the proof. □

The assumptions in this theorem hold for any of the discretizations mentioned. Note in particular that no assumptions are made on the derivatives of the numerical solution, but only on the derivative approximations used within the Dirichlet-Neumann iteration. The finite volume discretization as described does not use a reconstruction procedure and thus the corresponding solution has a zero derivative at the boundary. This is not a problem for the theorem and infact, the derivative approximation in the algorithm comes from the flux.

For specific discretizations, it is possible to prove that the asymptotic rate is attained all the time.

Corollary 1 *If the discretisations in Theorem 1 approximate a function in $\mathcal{P}_1$ exactly, then the convergence rate of the discrete Dirichlet-Neumann iteration is*

$$\mu = \left| \frac{\lambda_1 l_2}{\lambda_2 l_1} \right|.$$

Proof Follow the proof of Theorem 1 and remove all the limits by using the extra condition. □

This is the case for the linear finite elements, the second order central differences and for the finite volume discretization with an additional reconstruction procedure.

4 Numerical Results

We tested four different strategies when choosing the grids x^Is and y^Is for each test performed. The first strategy is the equidistant distribution. The second strategy is using pseudo-random uniform distribution, the third has grid points concentrated around the interface, x_Γ, and the fourth has grid points concentrated around the boundary points a and b. The points for the second and third strategy are generated by starting with an equidistant mesh on [0, 1]. These points are then put into the functions:

$$f_l(x) = \frac{e^x - 1}{e - 1} \text{ and } f_u(x) = f_l^{-1}(x) = \ln(x(e-1)+1). \tag{16}$$

This creates new sets that are either concentrated around 0 (f_l) or 1 (f_u). Next the values in the sets are sorted, scaled, and offset to be between the wanted boundary values. For the third strategy we used f_u for the x_i^Is and f_l for the y_i^Is, the fourth is vice versa.

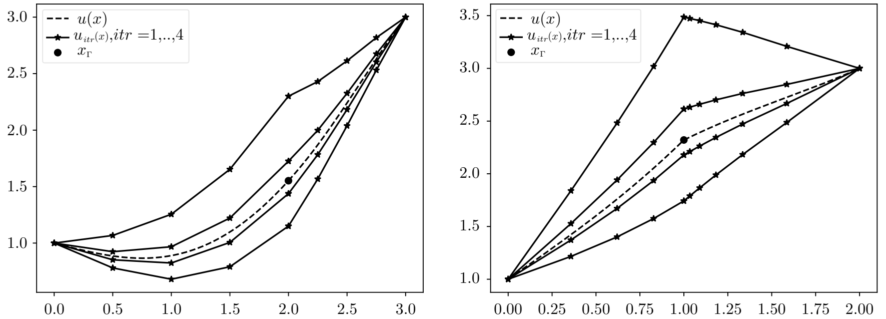


Fig. 2 Exact solution and discrete solution after 1, ..., 4 iterations. Left: FDM-FVM coupling on equidistant mesh on the domain [0, 2], [2, 3]. Right: FEM-FVM coupling with $\lambda_1 = 1, \lambda_2 = 2$ and mesh following (16) on the domain [0, 1], [1, 2]

Three tests were performed using an implementation in Python [7]. The first test varied λ_1, λ_2, the second $l_1,\ l_2$, and the third the resolutions n_1 and n_2. In all tests we set $\lambda_1 = \lambda_2 = 1,\ l_1 = l_2 = 1$ and $n_1 = n_2 = 100$ unless stated otherwise. Furthermore, we choose $f_1(x) = \sin x$ and $f_2(x) = \cos(3x)$. The numerical and exact solution is visualised in Fig. 2.

For each test we computed the observed convergence rate as:

$$\max_{k=0,\ldots,9} \left| \frac{e_k}{e_{k-1}} \right|,$$

where $e_k = |u^k(x_\Gamma) - u(x_\Gamma)|$. These tests were performed for all possible combinations involving the three discretizations

- linear finite elements
- second order central differences
- finite volumes with a central flux.

In all cases, the difference between the convergence rate and $\frac{\lambda_1 l_2}{\lambda_2 l_1}$ was negligible.

References

1. Badia, S., Nobile, F., Vergara, C.: Fluid-structure partitioned procedures based on Robin transmission conditions. J. Comp. Phys. **227**, 7027–7051 (2008)
2. Birken, P., Gleim, T., Kuhl, D., Meister, A.: Fast solvers for unsteady thermal fluid structure interaction. Int. J. Num. Meth. Fluids **79**, 16–29 (2015)
3. Causin, P., Gerbeau, J.F., Nobile, F.: Added-mass effect in the design of partitioned algorithms for fluid-structure problems. Comput. Methods Appl. Mech. Eng. **194**, 4506–4527 (2005)
4. Farhat, C.: CFD-based nonlinear computational aeroelasticity. In: Stein, E., de Borst, R., Hughes, T.J.R. (eds.) Encyclopedia of Computational Mechanics, vol. 3: Fluids, ch. 13, pp. 459–480. Wiley (2004)

5. Gander, M.J., Kwok, F., Mandal, B.C.: Dirichlet-Neumann and Neumann-Neumann waveform relaxation algorithms for parabolic problems. ETNA **45**, 424–456 (2016)
6. Görtz, M.: Convergence rate of the Dirichlet-Neumann algorithm for coupled Poisson equations. Master thesis, Lund University (2019)
7. Görtz, M.: Analysis-of-the-1D-Dirichlet-Neumann-algorithm, GitHub (2020). https://github.com/morgan-gortz/Analysis-of-the-1D-Dirichlet-Neumann-Algorithm
8. Heck, U., Fritsching, U., Bauckhage, K.: Fluid flow and heat transfer in gas jet quenching of a cylinder. Int. J. Numer. Methods Heat Fluid Flow **11**, 36–49 (2001)
9. Henshaw, W.D., Chand, K.K.: A composite grid solver for conjugate heat transfer in fluid-structure systems. J. Comp. Phys. **228**, 3708–3741 (2009)
10. Kowollik, D., Tini, V., Reese, S., Haupt, M.: 3D fluid-structure interaction analysis of a typical liquid rocket engine cycle based on a novel viscoplastic damage model. Int. J. Num. Meth. Eng. **94**, 1165–1190 (2013)
11. Kowollik, D.S.C., Horst, P., Haupt, M.C.: Fluid-structure interaction analysis applied to thermal barrier coated cooled rocket thrust chambers with subsequent local investigation of delamination phenomena. Prog. Propuls. Phys. **4**, 617–636 (2013)
12. Monge, A.: Partitioned methods for time-dependent thermal fluid-structure interaction. Ph.D. thesis, Lund University, 2018
13. Monge, A., Birken, P.: On the convergence rate of the Dirichlet-Neumann iteration for unsteady thermal fluid-structure interaction. Comp. Mech. **62**, 525–541 (2018)
14. van Brummelen, E.H.: Added mass effects of compressible and incompressible flows in fluid-structure interaction. J. Appl. Mech. **76**, 021206 (2009)

Optimized Overlapping DDFV Schwarz Algorithms

Martin J. Gander, Laurence Halpern, Florence Hubert, and Stella Krell

Abstract We introduce an overlapping optimized Schwarz methods in the DDFV framework for an anisotropic diffusion equation, and we show that a discrete and bounded domain convergence analysis is important to get best performance for strong anisotropy.

Keywords DDFV schemes · Anisotropic diffusion · Domain decomposition method

1 Introduction

We are interested in parallel solvers for the anisotropic diffusion problem

$$\mathcal{L}(u) := -\operatorname{div}(A\nabla u) + \eta u = f \text{ in } \Omega, \quad u = 0 \quad \text{on } \partial\Omega, \tag{1}$$

$$\text{with } (x, y) \in \Omega \mapsto A(x, y) = \begin{pmatrix} A_{xx} & A_{xy} \\ A_{xy} & A_{yy} \end{pmatrix}, \quad \eta > 0, \tag{2}$$

M. J. Gander
University of Geneva, 2-4 Rue du Lièvre, CP 64, 1211 Genève, Switzerland
e-mail: martin.gander@unige.ch

L. Halpern
Université PARIS 13, LAGA, 93430 Villetaneuse, France
e-mail: halpern@math.univ-paris13.fr

F. Hubert
Aix-Marseille Université, CNRS, Centrale Marseille, I2M UMR 7373,
39 rue F. Joliot Curie, 13453 Marseille Cedex 13, France
e-mail: florence.hubert@univ-amu.fr

S. Krell (✉)
Université de Nice, Parc Valrose, 28 Avenue Valrose,
06108 Nice Cedex 2, France
e-mail: krell@unice.fr

R. Klöfkorn et al. (eds.), *Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples*, Springer Proceedings in Mathematics & Statistics 323,
https://doi.org/10.1007/978-3-030-43651-3_33

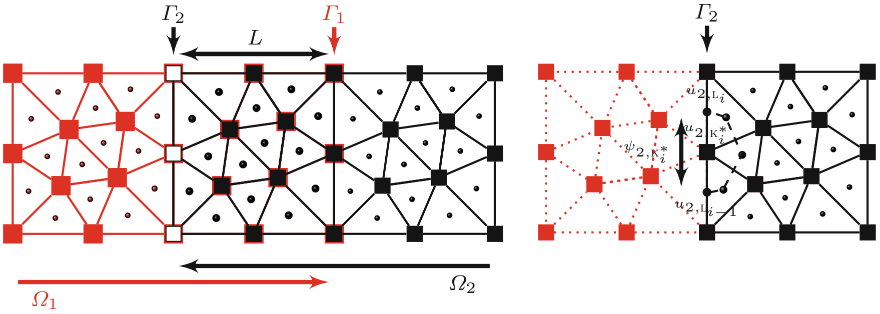


Fig. 1 Left: example of overlapping meshes, primal meshes $\mathfrak{M}_j$ are shown. Right: detailed notation near the interface Γ_2

where A is a uniformly symmetric positive definite matrix. Non-overlapping optimized Schwarz methods have been developed for (1) discretized with Discrete Duality Finite Volume (DDFV) schemes [6], because these techniques are especially well suited for anisotropic diffusion [7], [3], [1]. Since overlap in general greatly enhances the performance of Schwarz algorithms, we introduce and test here a new, optimized overlapping DDFV Schwarz algorithm, and we show that a discrete and bounded domain convergence analysis is important to get best performance for strong anisotropy.

2 Optimized Overlapping Schwarz Algorithm

For simplicity, we describe the algorithm for two rectangular subdomains with overlap, $\Omega = \Omega_1 \cup \Omega_2$, $\Omega_1 \cap \Omega_2 \neq \emptyset$, with interfaces $\Gamma_j = \partial\Omega_j \setminus \partial\Omega_j \cap \partial\Omega$, see Fig. 1. A general parallel[1] Schwarz method on these two subdomains is given by solving for $n = 1, 2, \ldots$ the subdomain problems

$$\begin{gathered}\mathcal{L}u_j^n = f \text{ in } \Omega_j, \quad u_j^n = 0 \text{ on } \partial\Omega_j \cap \partial\Omega, \\ \mathcal{B}_1 u_1^n = \mathcal{B}_1 u_2^{n-1} \text{ on } \Gamma_1, \quad \mathcal{B}_2 u_2^n = \mathcal{B}_2 u_1^{n-1} \text{ on } \Gamma_2.\end{gathered}$$

If we choose for the transmission operators $\mathcal{B}_j$ the identity, we obtain the classical Schwarz method. If we choose $\mathcal{B}_j := A_{xx}\partial_{n_j} + P(\partial_y)$ with $P(\partial_y) = p - qA_{yy}\partial_{yy}$, we obtain the so called optimized Schwarz methods [4], with Robin ($q = 0$) or Ventcell ($q \neq 0$) transmission conditions and overlap L,

$$\begin{aligned}A_{xx}\partial_x u_1^n(L, y) + Pu_1^n(L, y) &= A_{xx}\partial_x u_2^{n-1}(L, y) + Pu_2^{n-1}(L, y), \\ -A_{xx}\partial_x u_2^n\,(0, y) + Pu_2^n(0, y) &= -A_{xx}\partial_x u_1^{n-1}(0, y) + Pu_1^{n-1}(0, y).\end{aligned} \tag{3}$$

[1] Or alternating if $\mathcal{B}_2 u_1^n$ is transmitted on Γ_2, leaving the rest unchanged.

The convergence, discretization and optimization of such algorithms in the nonoverlapping case were studied in [6]; we study here for the first time the overlapping case.

3 DDFV Discretization

We now describe the DDFV Schwarz algorithm for overlapping subdomains and decompositions using the notation from [2], see Fig. 2.

The meshes: for $j = 1, 2$, the primal mesh $\mathfrak{M}_j$ is a set of disjoint open polygonal control volumes $\mathrm{K} \subset \Omega_j$ such that $\cup\overline{\mathrm{K}} = \overline{\Omega_j}$. We denote by $\partial\mathfrak{M}_j$ the set of edges of the control volumes in $\mathfrak{M}_j$ included in $\partial\Omega_j$, and by $\partial\mathfrak{M}_{\Gamma_j}$ the subset of $\partial\Omega_j$ of edges of primal boundary cells related to the interface $\Gamma_j = \partial\Omega_j \cap \Omega_i$ (*i.e.* in what follows, $i = 2$ if $j = 1$, and $i = 1$ if $j = 2$). We assume that each edge of Γ_j corresponds to an edge of $\mathfrak{M}_i$. We use the same notation for the dual mesh, $\mathfrak{M}_j^*$, $\partial\mathfrak{M}_j^*$ and $\partial\mathfrak{M}_{\Gamma_j}^*$. We define the diamond cells $\mathrm{D}_{\sigma,\sigma^*}$ as the quadrangles whose diagonals are a primal edge $\sigma = \mathrm{K}|\mathrm{L} = (x_{\mathrm{K}^*}, x_{\mathrm{L}^*})$ and a corresponding dual edge $\sigma^* = \mathrm{K}^*|\mathrm{L}^* = (x_{\mathrm{K}}, x_{\mathrm{L}})$. The set of diamond cells is called the diamond mesh, denoted by $\mathfrak{D}_j$.

For any V in $\mathfrak{M}_j \cup \partial\mathfrak{M}_j$ or $\mathfrak{M}_j^* \cup \partial\mathfrak{M}_j^*$, we denote by m_V its Lebesgue measure, by $\mathcal{E}_V$ the set of its edges, and $\mathfrak{D}_V := \{\mathrm{D}_{\sigma,\sigma^*} \in \mathfrak{D}_j,\ \ \sigma \in \mathcal{E}_V\}$. For $\mathrm{D} = \mathrm{D}_{\sigma,\sigma^*}$ with vertices $(x_{\mathrm{K}}, x_{\mathrm{K}^*}, x_{\mathrm{L}}, x_{\mathrm{L}^*})$, we denote by x_{D} the center of D, that is the intersection of the primal edge σ and the dual edge σ^*, by m_{D} its measure, by m_σ the length of σ, by m_{σ^*} the length of σ^*, by $m_{\sigma_{\mathrm{K}^*}}$ the length of $\partial\mathrm{K}^* \cap \Omega_j$, by $m_{\sigma_{\mathrm{L}}}$ the length of $\mathrm{D} \cap \partial\Omega_j$, and by $m_{\sigma_{\mathrm{K}}}$ the length of $[x_{\mathrm{K}}, x_{\mathrm{D}}]$. $\mathbf{n}_{\sigma\mathrm{K}}$ is the unit vector normal to σ oriented from x_{K} to x_{L}, and $\mathbf{n}_{\sigma^*\mathrm{K}^*}$ is the unit vector normal to σ^* oriented from x_{K^*} to x_{L^*}.

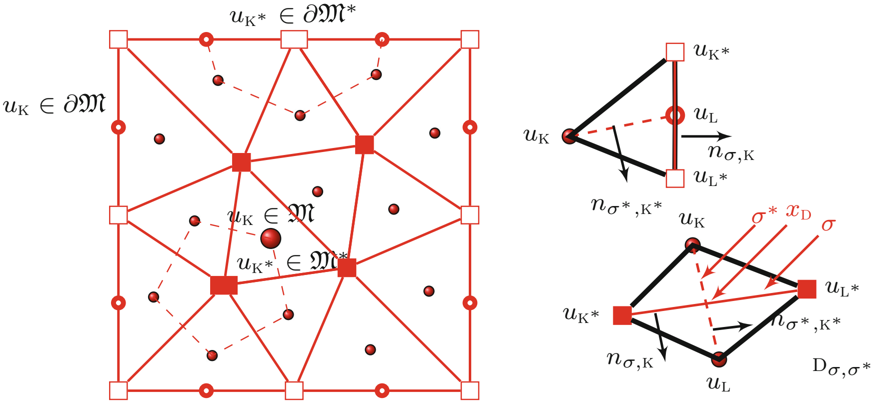

$u_{\mathrm{K}^*} \in \partial\mathfrak{M}^*$
$u_{\mathrm{K}} \in \partial\mathfrak{M}$
$u_{\mathrm{K}} \in \mathfrak{M}$
$u_{\mathrm{K}^*} \in \mathfrak{M}^*$
u_{K^*}
u_{L}
u_{K}
$n_{\sigma,\mathrm{K}}$
u_{L^*}
$n_{\sigma^*,\mathrm{K}^*}$
u_{K}
σ^* x_{D}
σ
u_{L^*}
u_{K^*}
$n_{\sigma^*,\mathrm{K}^*}$
$n_{\sigma,\mathrm{K}}$
u_{L}
$\mathrm{D}_{\sigma,\sigma^*}$

Fig. 2 Left: primal mesh and some dual cells. Right: diamond cell $\mathrm{D}_{\sigma,\sigma^*}$

For $\sigma = \partial\mathfrak{M}_{\Gamma_j}$, there exists $\widetilde{\mathrm{D}} \in \mathfrak{D}_i$ whose vertices are $x_{\mathrm{K}_i}, x_{\mathrm{K}^*}, x_{\mathrm{L}}, x_{\mathrm{L}^*}$ with $x_{\mathrm{K}_i} \in \Omega_i$. We denote by the half-diamond D_i the triangle whose vertices are $x_{\mathrm{K}_i}, x_{\mathrm{K}^*}, x_{\mathrm{L}^*}$ and by the half-diamond D_j the triangle whose vertices are $x_{\mathrm{K}_j}, x_{\mathrm{K}^*}, x_{\mathrm{L}^*}$, where $\mathrm{K}_j \in \mathfrak{M}_j$ such that $\sigma \in \partial\mathrm{K}_j$. Then let $\mathrm{D} = \mathrm{D}_i \cup \mathrm{D}_j$ and $\mathfrak{D}_{\Gamma_j}$ be the set of these diamonds.

The unknowns: the DDFV method associates to all primal control volumes $\mathrm{K} \in \mathfrak{M}_j \cup \partial\mathfrak{M}_j$ an unknown value $u_{j,\mathrm{K}}$, and to all dual control volumes $\mathrm{K}^* \in \mathfrak{M}_j^* \cup \partial\mathfrak{M}_j^*$ an unknown value u_{j,K^*}. To handle the transmission condition on Γ_j, we also require a flux unknown ψ_{j,K^*} per boundary dual cell $\mathrm{K} \in \partial\mathfrak{M}^*_{\Gamma_j}$. We denote the approximate solution on the mesh $\mathcal{T}_j$ by $u_{\mathcal{T}_j} = ((u_{j,\mathrm{K}})_{\mathrm{K}\in(\mathfrak{M}_j\cup\partial\mathfrak{M}_j)}, (u_{j,\mathrm{K}^*})_{\mathrm{K}^*\in(\mathfrak{M}_j^*\cup\partial\mathfrak{M}_j^*)}, (\psi_{j,\mathrm{K}^*})_{\mathrm{K}^*\in\partial\mathfrak{M}^*_{\Gamma_j}}) \in \mathbb{R}^{\mathcal{T}_j}$. When f is a continuous function, we define for all control volumes $\mathrm{C} \in \mathcal{T}_j$, $f_{\mathcal{T}_j} = (f_{\mathrm{C}})$ by $f_{\mathrm{C}} := f(x_{\mathrm{C}})$.

Operators. DDFV schemes can be described by two operators: a discrete gradient $\nabla^{\mathfrak{D}_j}$ and a discrete divergence $\mathrm{div}^{\mathcal{T}_j}$, which are dual to each other, see [2]. Let $\nabla^{\mathfrak{D}_j}$: $u_{\mathcal{T}_j} \in \mathbb{R}^{\mathcal{T}_j} \mapsto \left(\nabla^{\mathrm{D}} u_{\mathcal{T}_j}\right)_{\mathrm{D}\in\mathfrak{D}_j} \in (\mathbb{R}^2)^{\mathfrak{D}_j}$ and $\mathrm{div}^{\mathcal{T}_j} : \xi_{\mathfrak{D}_j} = (\xi_{\mathrm{D}})_{\mathrm{D}\in\mathfrak{D}_j} \mapsto \mathrm{div}^{\mathcal{T}_j}\xi_{\mathfrak{D}_j} \in \mathbb{R}^{\mathcal{T}_j}$ be defined as

$$\nabla^{\mathrm{D}} u_{\mathcal{T}_j} := \frac{1}{2m_{\mathrm{D}}}\left((u_{\mathrm{L}} - u_{\mathrm{K}})m_\sigma \mathbf{n}_{\sigma\mathrm{K}} + (u_{\mathrm{L}^*} - u_{\mathrm{K}^*})m_{\sigma^*}\mathbf{n}_{\sigma^*\mathrm{K}^*}\right), \quad \forall \mathrm{D} \in \mathfrak{D}_j,$$

$$\mathrm{div}^{\mathrm{K}}\xi_{\mathfrak{D}_j} := \frac{1}{m_{\mathrm{K}}}\sum_{\mathrm{D}\in\mathfrak{D}_{\mathrm{K}}} m_\sigma(\xi_{\mathrm{D}}, \mathbf{n}_{\sigma\mathrm{K}}),\ \forall \mathrm{K} \in \mathfrak{M}_j,\ \text{and}\ \mathrm{div}^{\mathrm{K}}\xi_{\mathfrak{D}_j} = 0, \forall \mathrm{K} \in \partial\mathfrak{M}_j,$$

$$\mathrm{div}^{\mathrm{K}^*}\xi_{\mathfrak{D}_j} := \frac{1}{m_{\mathrm{K}^*}}\sum_{\mathrm{D}\in\mathfrak{D}_{\mathrm{K}^*}} m_{\sigma^*}(\xi_{\mathrm{D}}, \mathbf{n}_{\sigma^*\mathrm{K}^*}),\ \forall \mathrm{K}^* \in \mathfrak{M}_j^* \cup \partial\mathfrak{M}_j^*.$$

DDFV scheme on Ω_j for Ventcell boundary conditions on Γ_j.

For $u_{\mathcal{T}_j} \in \mathbb{R}^{\mathcal{T}_j}$, $f_{\mathcal{T}_j} \in \mathbb{R}^{\mathcal{T}_j}$ and $h_{\mathcal{T}_j} \in \mathbb{R}^{\partial\mathfrak{M}_{\Gamma_j}\cup\partial\mathfrak{M}^*_{\Gamma_j}}$, we denote by $\mathcal{L}^{\mathcal{T}_j}_{\Omega_j}$ $(u_{\mathcal{T}_j}, f_{\mathcal{T}_j}, h_{\mathcal{T}_j}) = 0$ the linear system

$$-\mathrm{div}^{\mathrm{K}}\left(A_{\mathfrak{D}}\nabla^{\mathfrak{D}}u_{\mathcal{T}_j}\right) + \eta_{\mathrm{K}}u_{j,\mathrm{K}} = f_{\mathrm{K}},\ \ \forall\,\mathrm{K} \in \mathfrak{M}_j, \tag{4}$$

$$-\mathrm{div}^{\mathrm{K}^*}\left(A_{\mathfrak{D}}\nabla^{\mathfrak{D}}u_{\mathcal{T}_j}\right) + \eta_{\mathrm{K}^*}u_{j,\mathrm{K}^*} = f_{\mathrm{K}^*},\ \ \forall\,\mathrm{K}^* \in \mathfrak{M}_j^*, \tag{5}$$

$$-\sum_{\mathrm{D}\in\mathfrak{D}_{\mathrm{K}^*}}\frac{m_{\sigma^*}}{m_{\mathrm{K}^*}}\left(A_{\mathrm{D}}\nabla^{\mathrm{D}}u_{\mathcal{T}_j}, \mathbf{n}_{\sigma^*\mathrm{K}^*}\right) - \frac{m_{\sigma_{\mathrm{K}^*}}}{m_{\mathrm{K}^*}}\psi_{j,\mathrm{K}^*} + \eta_{\mathrm{K}^*}u_{j,\mathrm{K}^*} = f_{\mathrm{K}^*}, \forall\mathrm{K}^* \in \partial\mathfrak{M}^*_{\Gamma_j}, \tag{6}$$

$$\left(A_{\mathrm{D}}\nabla^{\mathrm{D}}u_{\mathcal{T}_j}, \mathbf{n}_{\sigma\mathrm{L}}\right) + \Lambda_{\mathrm{L}}^{\partial\mathfrak{M}_{\Gamma_j}}(u_{\partial\mathfrak{M}_{\Gamma_j}}) = h_{j,\mathrm{L}},\ \ \forall\,\mathrm{L} \in \partial\mathfrak{M}_{\Gamma_j}, \tag{7}$$

$$\psi_{j,\mathrm{K}^*} + \Lambda_{\mathrm{K}^*}^{\partial\mathfrak{M}^*_{\Gamma_j}}(u_{\partial\mathfrak{M}^*_{\Gamma_j}}) = h_{j,\mathrm{K}^*},\ \ \forall\,\mathrm{K}^* \in \partial\mathfrak{M}^*_{\Gamma_j}, \tag{8}$$

$$u_{j,\mathrm{K}} = 0,\ \ \forall\,\mathrm{K} \in \partial\mathfrak{M}_j \cap \partial\Omega, \qquad u_{j,\mathrm{K}^*} = 0,\ \ \forall\,\mathrm{K}^* \in \partial\mathfrak{M}_j^* \cap \partial\Omega, \tag{9}$$

The transmission operators $\Lambda^{\partial\mathfrak{M}_{\Gamma_j}}$ and $\Lambda^{\partial\mathfrak{M}^*_{\Gamma_j}}$ are defined by

$$\Lambda_{\mathrm{L}_s}^{\partial\mathfrak{M}_{\Gamma_j}}(u_{\partial\mathfrak{M}_{\Gamma_j}}) := p u_{j,\mathrm{L}_s} - A_{yy}\frac{q}{m_{\sigma_s}}\left(\frac{u_{j,\mathrm{L}_{s+1}} - u_{j,\mathrm{L}_s}}{m_{\sigma_{\mathrm{K}^*_{s+1}}}} - \frac{u_{j,\mathrm{L}_s} - u_{j,\mathrm{L}_{s-1}}}{m_{\sigma_{\mathrm{K}^*_s}}}\right),$$

for $s = 1, \dots, N_j$, with $u_{j,\mathrm{L}_0} = u_{j,\mathrm{L}_{N_j+1}} = 0$, and for $s = 2, \dots, N_j$ by

$$\Lambda_{\mathrm{K}_s^*}^{\partial\mathfrak{M}_{\Gamma_j}^*}(u_{\partial\mathfrak{M}_{\Gamma_j}^*}) := p u_{j,\mathrm{K}_s^*} - A_{yy}\frac{q}{m_{\sigma_{\mathrm{K}^*s}}}\left(\frac{u_{j,\mathrm{K}_{s+1}^*} - u_{j,\mathrm{K}_s^*}}{m_{\sigma_s}} - \frac{u_{j,\mathrm{K}_s^*} - u_{j,\mathrm{K}_{s-1}^*}}{m_{\sigma_{s-1}}}\right).$$

Note that the edges $\sigma_1, \dots, \sigma_{N_j}$ have been sorted such that $\sigma_s \cap \sigma_{s+1} \neq \emptyset$, and $x_{\mathrm{K}_s^*}, x_{\mathrm{K}_{s+1}^*}$ are the vertices of σ_s, where $x_{\mathrm{K}_s^*} = \sigma_s \cap \sigma_{s-1}$. Note also that $u_{j,\mathrm{K}_1^*} = u_{j,\mathrm{K}_{N_j+1}^*} = 0$ because of the homogeneous boundary condition on $\partial\Omega$. Equations (4)–(6) correspond to approximations of the equation after integration on $\mathfrak{M}_j$, $\mathfrak{M}_j^*$ and $\partial\mathfrak{M}_j^*$; equations (7) and (8) stem from the transmission condition on $\partial\mathfrak{M}_{\Gamma_j}$ and $\partial\mathfrak{M}_{\Gamma_j}^*$; equation (9) corresponds to the Dirichlet boundary condition on $\partial\Omega$. One can show that this discrete formulation is well posed, see [6, Theorem 3.1].

DDFV Schwarz algorithm. The DDFV optimized Schwarz algorithm performs for an arbitrary initial guess $h_{\mathcal{T}_j}^0 \in \mathbb{R}^{\partial\mathfrak{M}_{\Gamma_j} \cup \partial\mathfrak{M}_{\Gamma_j}^*}$, $(j, i) = (1, 2)$ or $(j, i) = (2, 1)$ and $l = 1, 2, \dots$ the following steps:

- Compute the solutions $u_{\mathcal{T}_j}^{l+1} \in \mathbb{R}^{\mathcal{T}_j}$ of $\mathcal{L}_{\Omega_j}^{\mathcal{T}_j}(u_{\mathcal{T}_j}^{l+1}, f_{\mathcal{T}_j}, h_{\mathcal{T}_j}^l) = 0$.
- Then, we define $P_{\partial\mathfrak{M}_{\Gamma_j}^*}(u_{\mathcal{T}_i}^{l+1})$ on the vertices of the interface Γ_j

  as follows: for $x_{\mathrm{K}_j^*} \in \partial\mathfrak{M}_{\Gamma_j}^*$, there exists a unique vertex $x_{\mathrm{K}^*} \in \mathfrak{M}_i^*$ such that $x_{\mathrm{K}_j^*} = x_{\mathrm{K}^*}$, we set $u_{\mathrm{K}_j^*}^{l+1} = u_{\mathrm{K}^*}^{l+1}$.
- Then, we define $P_{\partial\mathfrak{M}_{\Gamma_j}}(u_{\mathcal{T}_i}^{l+1})$ on the interface Γ_j as follows.

  For $\mathrm{D} \in \mathfrak{D}_{\Gamma_j}$, there exist two half-diamonds D_i and D_j such that $\mathrm{D} = \mathrm{D}_i \cup \mathrm{D}_j$. We define $P_\sigma(u_{\mathcal{T}_i}^{l+1}) = u_\mathrm{L}$ such that

$$\left(A_{\mathrm{D}_i}\nabla^{\mathrm{D}_i}u_{\mathcal{T}_i}^{l+1}, \mathbf{n}_{\sigma\mathrm{L}_i}\right) = \left(A_{\mathrm{D}_j}\nabla^{\mathrm{D}_j}u_{\mathcal{T}_j}^{l+1}, \mathbf{n}_{\sigma\mathrm{L}_j}\right).$$

- We evaluate the flux unknowns. For $x_{\mathrm{K}_j^*} \in \partial\mathfrak{M}_{\Gamma_j}^*$, there exists a unique vertex $x_{\mathrm{K}^*} \in \mathfrak{M}_i^*$ such that $x_{\mathrm{K}_j^*} = x_{\mathrm{K}^*}$. We set $\mathrm{K}_i^* = \mathrm{K}^* \cap (\Omega_i \setminus \Omega_j)$

$$\frac{m_{\sigma_{\mathrm{K}^*}}}{m_{\mathrm{K}_i^*}}\psi_{i,\mathrm{K}^*}^{l+1} = -\sum_{\mathrm{D}\in\mathfrak{D}_{\mathrm{K}_i^*}}\frac{m_{\sigma^*}}{m_{\mathrm{K}_i^*}}\left(A_\mathrm{D}\nabla^\mathrm{D}u_{\mathcal{T}_i}^{l+1}, \mathbf{n}_{\sigma^*\mathrm{K}_i^*}\right) + \eta_{\mathrm{K}^*}u_{i,\mathrm{K}_i^*}^{l+1} - f_{\mathrm{K}_i^*}.$$

- We evaluate the new interface values $h_{\mathcal{T}_j}^{l+1}$ by

$$h_{j,\mathrm{L}}^{l+1} = -\left(A_\mathrm{D}\nabla^\mathrm{D}u_{\mathcal{T}_i}^{l+1}, \mathbf{n}_{\sigma\mathrm{L}_i}\right) + \Lambda_\mathrm{L}^{\partial\mathfrak{M}_{\Gamma_j}}(P_{\partial\mathfrak{M}_{\Gamma_j}}(u_{\mathcal{T}_i}^{l+1})),\ \forall\mathrm{L} \in \partial\mathfrak{M}_{\Gamma_j}, \tag{10a}$$

$$h_{j,\mathrm{K}^*}^{l+1} = -\psi_{i,\mathrm{K}^*}^{l+1} + \Lambda_{\mathrm{K}^*}^{\partial\mathfrak{M}_{j,\Gamma}^*}(P_{\partial\mathfrak{M}_{\Gamma_j}^*}(u_{\mathcal{T}_i}^{l+1})),\ \forall\mathrm{K}^* \in \partial\mathfrak{M}_{\Gamma_j}^*. \tag{10b}$$

4 Convergence Factors

We now give the convergence factors of the DDFV Schwarz algorithm for a rectangular two subdomain decomposition, $\Omega := (-a_1, a_2) \times (0, b)$, $\Omega_1 := (-a_1, L) \times (0, b)$ and $\Omega_2 := (0, a_2) \times (0, b)$, $L \geq 0$ being the overlap size.

Continuous case. The error $e_j^n := u - u_j^n$ satisfies homogeneous Dirichlet boundary conditions, and can thus be expanded in a Fourier sine series, $e_j^n(x, y) = \sum_{k \in E} \hat{e}_j^n(x, k) \sin ky$ with the set $E := \frac{\pi \mathbb{N}^+}{b}$. A direct computation with

$$r(k) := \frac{\sqrt{\eta A_{xx} + k^2 \det A}}{A_{xx}} \tag{11}$$

leads to the continuous convergence factors for classical and optimized Schwarz:

$$\rho_{c,a}^{cla} = \frac{\sinh(r(a_2 - L))}{\sinh(r(L + a_1))} \frac{\sinh(r a_1)}{\sinh(r a_2)}, \tag{12}$$

$$\rho_{c,a} = \rho_{c,a}^{cla} \frac{P - A_{xx} r \coth((a_2 - L)r)}{P + A_{xx} r \coth((a_1 + L)r)} \frac{P - A_{xx} r \coth(a_1 r)}{P + A_{xx} r \coth(a_2 r)}. \tag{13}$$

Discrete case on a Cartesian mesh. Let the mesh sizes be (h_x, h_y), and $M_j = N_j h_x$, $L = M h_x$. When $A_{xy} = 0$, the equations for the primal and dual unknowns decouple into two finite difference schemes of order 2: the primal unknowns are solutions of a cell-centered (CC) scheme, and the dual unknowns are solutions of a vertex centered (VC) scheme. Using again Fourier analysis, with the notation

$$\begin{aligned} &\alpha(k) := \tfrac{4A_{yy}}{{h_y}^2} \sin^2(\tfrac{k h_y}{2}), \quad \mu(k) := \tfrac{h_x^2}{A_{xx}}(\alpha(k) + \eta), \\ &\lambda(k) := 1 + \tfrac{\mu(k)}{2} - \sqrt{\mu(k) + \tfrac{\mu(k)^2}{4}} \in (0, 1), \ \tilde{r}(k) := -\ln \lambda(k) > 0, \end{aligned} \tag{14}$$

the discrete convergence factor for classical Schwarz for both CC and VC is

$$\rho_{d,M}^{cla} = \frac{\sinh((M_2 - M)\tilde{r})}{\sinh((M_1 + M)\tilde{r})} \frac{\sinh(M_1 \tilde{r})}{\sinh(M_2 \tilde{r})}. \tag{15}$$

For optimized Schwarz, with $P(k) := p + q\alpha(k)$, we get for VC and CC

$$\begin{aligned} \rho_{dvc,M} &= \rho_{d,M}^{cla} \frac{P - \frac{A_{xx}}{h_x} \sinh \tilde{r} \coth((M_2 - M)\tilde{r})}{P + \frac{A_{xx}}{h_x} \sinh \tilde{r} \coth((M_1 + M)\tilde{r})} \frac{P - \frac{A_{xx}}{h_x} \sinh \tilde{r} \coth(M_1 \tilde{r})}{P + \frac{A_{xx}}{h_x} \sinh \tilde{r} \coth(M_2 \tilde{r})}, \\ \rho_{dcc,M} &= \rho_{d,M}^{cla} \frac{P - 2\frac{A_{xx}}{h_x} \tanh \frac{\tilde{r}}{2} \coth((M_2 - M)\tilde{r})}{P + 2\frac{A_{xx}}{h_x} \tanh \frac{\tilde{r}}{2} \coth((M_1 + M)\tilde{r})} \frac{P - 2\frac{A_{xx}}{h_x} \tanh \frac{\tilde{r}}{2} \coth(M_1 \tilde{r})}{P + 2\frac{A_{xx}}{h_x} \tanh \frac{\tilde{r}}{2} \coth(M_2 \tilde{r})}. \end{aligned} \tag{16}$$

We also obtain the classical unbounded domain convergence factors $\rho_{c,\infty}^{cla}$, $\rho_{c,\infty}$, $\rho_{d,\infty}^{cla}$ $\rho_{dvc,\infty}$, and $\rho_{dcc,\infty}$ from the bounded ones in (12), (13), (15) and (16) by passing to the limit as a_1, a_2 and M_1, M_2 go to infinity, which greatly simplifies the expressions, see [5, Sect. 5] for the continuous case. Since the convergence speed of the methods is bounded by the largest contraction factor over all k, we further introduce the corresponding upper case quantity $R := \sup_{k \in E} |\rho|$, and add a superscript R^*, p^* and q^* to denote the quantities obtained when R has been minimized using p and q.

5 Importance of a Bounded Domain Discrete Analysis

For isotropic diffusion, bounded and unbounded domain analyses both at the continuous and discrete level give very similar optimized parameters (p^*, q^*) as in the non-overlapping case [6]. We show now that for anisotropic diffusion the discrete bounded domain analysis gives much more accurate predictions $(p^*_{dvc,M}, q^*_{dvc,M})$, and $(p^*_{dcc,M}, q^*_{dcc,M})$. We choose $A_{xx} = 16$, $A_{yy} = 1$ and $A_{xy} = 0$, and decompose the domain $\Omega := (-1, 1) \times (0, 1)$ into two subdomains $\Omega_1 := (-1, L) \times (0, 1)$ and $\Omega_2 := (0, 1) \times (0, 1)$. We choose for the mesh sizes $h_x = h_y = h$ and for the overlap $L = h$. We also compute the numerically best working transmission parameters $(p^*_{dvc,num}, q^*_{dvc,num})$, and $(p^*_{dcc,num}, q^*_{dcc,num})$ by minimizing the numerical error remaining after $n = 50$ alternating Schwarz iterations with random initial guess solving directly the homogeneous error equations, and the corresponding numerical convergence factor $R^*_{num} := (||e^n||/||e^1||)^{\frac{1}{n-1}}$ where $||e^n||$ denotes the L^2 norm of the error e^n on the interface of the subdomains. We see in Table 1 in the top part that for VC the optimized parameters (p^*, q^*) are quite different for the bounded and unbounded analysis, and there is also a difference between discrete and continuous analysis. The asymptotic growth rate is however the same for all different analysis types, just the constant differs. The theoretically optimized convergence factors R^* in the bottom part of Table 1 on the left, and the numerically measured convergence factors R using the theoretically optimized parameters in the bottom part of Table 1

Table 1 $A_{xx} = 16$, $A_{yy} = 1$, Ventcell coefficients, vertex centered (VC)

		Theoretical and best numerical parameters										
h		$p^*_{c,\infty}$	$p^*_{c,a}$	$p^*_{dvc,\infty}$	$p^*_{dvc,M}$	$p^*_{vc,num}$		$q^*_{c,\infty}$	$q^*_{c,a}$	$q^*_{dvc,\infty}$	$q^*_{dvc,M}$	$q^*_{vc,num}$
2^{-3}		13.3040	20.6673	12.1818	19.7453	19.7200		0.2137	0.1661	0.2623	0.1995	0.1997
2^{-4}		15.8876	22.9419	14.7928	21.5210	21.4130		0.1391	0.1132	0.1688	0.1393	0.1395
2^{-5}		18.4998	26.0734	17.5707	24.6458	24.5234		0.0936	0.0769	0.1089	0.0921	0.0926
2^{-6}		21.2112	29.6330	20.6142	28.4895	28.2013		0.0647	0.0531	0.0707	0.0602	0.0604

		Theoretical convergence factors					Numerically measured convergence factors				
h		$R^*_{c,\infty}$	$R^*_{c,a}$	$R^*_{dvc,\infty}$	$R^*_{dvc,M}$		$R_{c,\infty}$	$R_{c,a}$	$R_{dvc,\infty}$	$R_{dvc,M}$	$R^*_{vc,num}$
2^{-3}		0.0049	0.0012	0.0027	0.0003		0.0154	0.0013	0.0205	0.0003	0.0003
2^{-4}		0.0161	0.0079	0.0111	0.0043		0.0183	0.0079	0.0138	0.0042	0.0041
2^{-5}		0.0347	0.0223	0.0280	0.0159		0.0338	0.0222	0.0274	0.0159	0.0156
2^{-6}		0.0596	0.0440	0.0541	0.0371		0.0561	0.0439	0.0526	0.0370	0.0360

Table 2 $A_{xx} = 16$, $A_{yy} = 1$, Ventcell coefficients, cell centered (CC)

		Theoretical and best numerical parameters										
h		$p^*_{c,\infty}$	$p^*_{c,a}$	$p^*_{dcc,\infty}$	$p^*_{dcc,M}$	$p^*_{cc,num}$		$q^*_{c,\infty}$	$q^*_{c,a}$	$q^*_{dcc,\infty}$	$q^*_{dcc,M}$	$q^*_{cc,num}$
2^{-3}		13.6259	21.0138	12.4676	19.9705	19.9264		0.2015	0.1568	0.2441	0.1837	0.1840
2^{-4}		16.0014	23.1167	15.0389	21.8064	21.7329		0.1363	0.1105	0.1601	0.1315	0.1316
2^{-5}		18.5257	26.1345	17.8211	24.9634	24.8306		0.0932	0.0763	0.1041	0.0878	0.0882
2^{-6}		21.2157	29.6388	20.8856	28.5642	28.5642		0.0648	0.0531	0.0678	0.0577	0.0579

		Theoretical convergence factors					Numerically measured convergence factors				
h		$R^*_{c,\infty}$	$R^*_{c,a}$	$R^*_{dcc,\infty}$	$R^*_{dcc,M}$		$R_{c,\infty}$	$R_{c,a}$	$R_{dcc,\infty}$	$R_{dcc,M}$	$R^*_{cc,num}$
2^{-3}		0.0058	0.0016	0.0032	0.0005		0.0137	0.0017	0.0193	0.0005	0.0005
2^{-4}		0.0167	0.0084	0.0121	0.0049		0.0182	0.0084	0.0141	0.0049	0.0048
2^{-5}		0.0349	0.0226	0.0297	0.0172		0.0338	0.0226	0.0293	0.0172	0.0169
2^{-6}		0.0596	0.0440	0.0566	0.0392		0.0563	0.0439	0.0549	0.0391	0.0381

Table 3 $A_{xx} = 16$, $A_{yy} = 1$, DDFV with VC and CC coefficients from the discrete bounded domain analysis, and best working ones for DDFV

h	$p^*_{dvc,M}$	$q^*_{dvc,M}$	$R_{ddfv,num}$	$p^*_{dcc,M}$	$q^*_{dcc,M}$	$R_{ddfv,num}$	$p^*_{ddfv,num}$	$q^*_{ddfv,num}$	$R^*_{ddfv,num}$
2^{-3}	19.7453	0.1995	0.0369	19.9705	0.1837	0.0314	20.6898	0.1836	0.0242
2^{-4}	21.50210	0.1393	0.0780	21.8064	0.1315	0.0713	22.1314	0.1324	0.0676
2^{-5}	24.64583	0.0921	0.1326	24.8306	0.0878	0.1278	24.8414	0.0891	0.1258
2^{-6}	28.4895	0.0602	0.1906	28.5642	0.0577	0.1909	28.2904	0.0593	0.1881

on the right clearly show that the best results are obtained for the discrete bounded domain analysis technique, very close to the numerically optimized $R^*_{vc,num}$. The results in Table 2 for CC are similar, only the q^* are a bit smaller, and convergence is slightly slower for CC than for VC. Table 3 shows the results for DDFV which computes both VC and CC interlaced simultaneously. We see that both the optimized parameters from the VC and CC discrete and bounded domain analysis give very good performance, the CC ones just being a little better.

To conclude, we presented a first step for the design and analysis of optimized overlapping DDFV Schwarz methods for anisotropic diffusion. Using the fact that for rectangular meshes the primal and dual unknowns decouple, we computed two convergence factors, whose maximum represents an upper bound on the convergence of the DDFV Schwarz method. Our analysis will allow us to study anisotropic meshes, and we will also investigate the influence of non-matching meshes in the different subdomains. Finally, a theoretical optimization of the coupled convergence factors in the parameters is needed to get closed formulas for the optimized parameters.

References

1. Boyer, F., Hubert, F.: Finite volume method for 2D linear and nonlinear elliptic problems with discontinuities. SIAM J. Numer. Anal. **46**, (2008)
2. Boyer, F., Hubert, F., Krell, S.: Non-overlapping Schwarz algorithm for solving 2D m-DDFV schemes. IMA Jour. Num. Anal. **30**, (2009)
3. Domelevo, K., Omnes, P.: A finite volume method for the Laplace equation on almost arbitrary two-dimensional grids. M2AN Math. Model. Numer. Anal. **39**(6), 1203–1249 (2005)
4. Gander, M.J.: Optimized Schwarz method. SIAM J. Numer. Anal. **44**(2), 699–731 (2006)
5. Gander, M.J., Dubois, O.: Optimized Schwarz methods for a diffusion problem with discontinuous coefficient. Numer. Algorithms **69**(1), 109–144 (2015)
6. Gander, M.J., Halpern, L., Hubert, F., Krell, S.: Optimized Schwarz methods for anisotropic diffusion with discrete duality finite volume discretizations (2018), (Submitted). URL https://hal.archives-ouvertes.fr/hal-01782357.
7. Hermeline, F.: Approximation of diffusion operators with discontinuous tensor coefficients on distorted meshes. Comput. Methods Appl. Mech. Engrg. **192**(16–18), 1939–1959 (2003)

Model Adaptation of Balance Laws Based on A Posteriori Error Estimates and Surrogate Fluxes

Jan Giesselmann and Hrishikesh Joshi

Abstract In this proceeding, we present model adaptation for hyperbolic balance laws based on a posteriori error estimates. The model adaptation is carried out by decomposing the computational domain and choosing to solve either the full system or a simpler reduced system. The decision is made based on error estimates constructed employing the relative entropy framework which allows us to bound the difference between the numerical solution to the reduced system and the exact solution to the full system. Furthermore, the use of surrogate fluxes in the simple model constructed by machine learning is proposed to further reduce the computational expenses.

Keywords Model adaptation · Hyperbolic balance laws · A posteriori error estimates · Surrogate fluxes · Machine learning

1 Introduction

A motivating example for the model adaptation strategy presented here is modelling of chemically reacting flows. They can be modelled using Euler equations with source terms, whereby we solve for mass balance of each constituent, conservation of total momentum and total energy. In this description, the source terms in the equations of mass balance describe reactions between the constituents. For details on the modelling of chemically reacting flows and their analysis the interested reader can refer to [1–3]. As the constituent species react, the system is driven towards chemical equillibrium and as a result the source terms vanish. From whence, it is no longer neccessary to solve the full system of equations. The full system of equations is from

J. Giesselmann · H. Joshi (✉)
Numerical Analysis and Scientific Computing, Department of Mathematics, Technical University of Darmstadt, Darmstadt, Germany
e-mail: joshi@mathematik.tu-darmstadt.de

J. Giesselmann
e-mail: giesselmann@mathematik.tu-darmstadt.de

R. Klöfkorn et al. (eds.), *Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples*, Springer Proceedings in Mathematics & Statistics 323,
https://doi.org/10.1007/978-3-030-43651-3_34

hereon referred to as the *complex* system. In such instances, when the system is *close to* chemical equillibrium, the governing equations can be simplified by projecting the system on the equillibrium manifold without introducing significant errors, i.e the set of states for which the reaction terms vanish. This gives rise to a reduced *simple* system. Model adaptation can be carried out by exploiting this structure. The idea is to decompose the computational domain and then solve the complex system wherever neccessary and the simple system everywhere else. It is important to note that to calculate the flux function of the simple system, a mapping is needed from the state space of the simple system to the equillibrium manifold in the state space of the complex system. We refer to this mapping as the *Maxwellian*, which is given by the solution of a non-linear system of equations. In the case of chemically reacting flows, pressure is calculated as part of this system.

We propose model adaptation consisting of a twofold approach. Firstly, we construct a posteriori error estimates, i.e. bounds for the difference between the numerical solution of the simple system and the exact solution of the complex system based on the relative entropy framework. Secondly, in order to further reduce the computational expenses we propose constructing an approximate *Maxwellian* by employing machine learning.

The structure of this paper is as follows: in Sect. 2, the framework is described abstractly. In Sect. 3, the a posteriori error estimates are presented. Lastly, in Sect. 4, convergence and construction of the approximate *Maxwellian* is discussed.

2 Abstract Form

2.1 Balance Laws

The complex system of partial differential equations in an abstract form is given by

$$\partial_t \mathbf{U} + \sum_\alpha \partial_{x_\alpha} \mathbf{F}_\alpha(\mathbf{U}) = \frac{1}{\varepsilon}\mathbf{R}(\mathbf{U}), \mathbf{U} : \mathbb{R}^d \times \mathbb{R}^+ \to \mathbb{R}^N, \tag{1}$$

where $\varepsilon > 0$, $\mathbf{R}, \mathbf{F} : \mathbb{R}^N \to \mathbb{R}^N$.
Employing some projection matrix $\mathbb{P} : \mathbb{R}^N \to \mathbb{R}^n$ such that $\mathbb{P}\mathbf{R}(\mathbf{U}) = 0$ and $\mathbf{u} := \mathbb{P}\mathbf{U}$, we get

$$\partial_t \mathbf{u} + \sum_\alpha \partial_{x_\alpha} \mathbb{P}\mathbf{F}_\alpha(\mathbf{U}) = 0, \quad \mathbf{u} : \mathbb{R}^d \times \mathbb{R}^+ \to \mathbb{R}^n. \tag{2}$$

The *Maxwellian*, i.e. the map from the state space of the simple system to the equillibrium manifold in the state space of the complex system is given by

$$\mathbf{R}(M(\mathbf{u})) = 0, \ \mathbb{P}M(\mathbf{u}) = \mathbf{u}. \tag{3}$$

In the limit $\varepsilon \to 0$, system (2) reduces to the simple system given by

$$\partial_t \mathbf{u} + \sum_\alpha \partial_{x_\alpha} \mathbb{P}\mathbf{F}_\alpha(M(\mathbf{u})) = 0. \tag{4}$$

3 A Posteriori Error Analysis

To carry out model adaptation we need to decide which model to solve where. We do this based on the bounds for the difference between the numerical solution of the simple system and the exact solution of the complex system. It is well-known that solutions to hyperbolic partial differential equations develop discontinuities in finite time, and hence one looks for entropy admissible weak solutions [4]. It is also known that entropy admissible weak solutions for systems are not unique [5]. However, based on the relative entropy framework it can be shown that weak-strong uniqueness holds when a Lipschitz continuous solution exists. In a similar vein, it can be shown that solutions of the complex system converge to that of the simple system for vanishing ε as long as a Lipschitz solution to (4) exists. For more details about the relative entropy framework, the reader is referred to [6].

In this section, the posteriori error estimates are presented. For the sake of simplicity of presenting the a posteriori analysis we restrict ourselves to one spatial dimension. First the entropic structure is outlined followed by the construction of the error estimates.

3.1 Relative Entropy Framework

The complex system is equipped with a strictly convex entropy-entropy flux pair $(H(\mathbf{U}), Q(\mathbf{U}))$ which satisfies

$$\mathrm{D}\, H(\mathbf{U})\, \mathrm{D}\, \mathbf{F}(\mathbf{U}) = \mathrm{D}\, Q(\mathbf{U}). \tag{5}$$

Furthermore, smooth solutions of (1) satisfy

$$\partial_t H(\mathbf{U}) + \partial_x Q(\mathbf{U}) = \frac{1}{\varepsilon} \frac{\partial H(\mathbf{U})}{\partial \mathbf{U}} \cdot \mathbf{R}(\mathbf{U}) \leq 0. \tag{6}$$

As in [6], we assume that for the simple system, the *Maxwellian* induces an entropy-entropy flux pair via $\eta(\mathbf{u}) := H(M(\mathbf{u})),\ q(\mathbf{u}) := Q(M(\mathbf{u}))$. Thus smooth solutions satisfy

$$\partial_t \eta(\mathbf{u}) + \partial_x q(\mathbf{u}) = 0. \tag{7}$$

Remark 1 The assumption that the *Maxwellian* composed with the entropy and entropy flux of the complex system induces an entropy-entropy flux pair (η, q) on the state space of the simple system is one of the fundamental ingredients in the convergence analysis [6] and which we employ in our numerical analysis. This condition is natural since the state space of the simple system is a submanifold (parameterized by the Maxwellian) of constrained equilibrium in the state space of the complex system.

Definition 1 The relative entropy and relative entropy flux between states $\mathbf{U}, \mathbf{V} \in \mathbb{R}^N$ are defined as

$$H(\mathbf{U}|\mathbf{V}) = H(\mathbf{U}) - H(\mathbf{V}) - \frac{\partial H}{\partial \mathbf{U}}(\mathbf{V})(\mathbf{U} - \mathbf{V}), \tag{8}$$

$$Q(\mathbf{U}|\mathbf{V}) = Q(\mathbf{U}) - Q(\mathbf{V}) - \frac{\partial H}{\partial \mathbf{U}}(\mathbf{V})(\mathbf{F}(\mathbf{U}) - \mathbf{F}(\mathbf{V})). \tag{9}$$

Furthermore, strict convexity of H implies that for some $c \geq 0$

$$H(\mathbf{U}|\mathbf{V}) \geq c\,|\mathbf{U} - \mathbf{V}|^2. \tag{10}$$

3.2 Reconstruction

The relative entropy framework requires one of the solutions to be Lipschitz continuous. As the exact solution can be discontinous, we introduce a Lipschitz reconstruction of the numerical solution. For more details on reconstruction of Discontinuous Galerkin solutions, the reader is referred to [7, 8].

Let $\mathbf{U}$ be the exact solution to (1), let $\mathbf{U}_h$ be some numerical solution to (1), let $\widehat{\mathbf{U}}_h$ be its reconstruction, furthermore let $\mathbf{u}_h$ be some numerical solution to (4) and let $\widehat{\mathbf{u}}_h$ be its reconstruction. Employing triangle inequality, we can bound the error between the numerical solution to the simple system and the exact solution to the complex system as

$$||\mathbf{U} - M(\mathbf{u}_h)|| \leq ||\mathbf{U} - M(\widehat{\mathbf{u}}_h)|| + ||M(\widehat{\mathbf{u}}_h) - M(\mathbf{u}_h)||. \tag{11}$$

The second term is explicitly computable and the first will be bounded by the error estimates.

3.3 Error Estimates

3.3.1 Computational Domain Decomposition

Let the computational domain be $\mathbb{R}$, assuming we solve the simple system on Ω_s and solve the complex system on Ω_c where, $\Omega_c \cup \Omega_s = \mathbb{R}$ and for simplicity $\Omega_c \cap \Omega_s = \{x_q\} \notin int(\Omega_c),\ int(\Omega_s)$. The Lipschitz reconstruction of the numerical solution satisfies a perturbed system of partial differential equations.
The reconstruction of the numerical solution of the complex system $\widehat{\mathbf{U}}_h$ satisfies

$$\partial_t \widehat{\mathbf{U}}_h + \partial_x \mathbf{F}(\widehat{\mathbf{U}}_h) - \frac{1}{\varepsilon}\mathbf{R}(\widehat{\mathbf{U}}_h) =: r_c,\ \widehat{\mathbf{U}}_h : \Omega_c \times \mathbb{R}^+ \to \mathbb{R}^N,$$

where r_c is the discretization residual in the complex system. Similarly, the reconstruction of the numerical solution of the simple system $\widehat{\mathbf{u}}_h$ satisfies

$$\partial_t \widehat{\mathbf{u}}_h + \partial_x \mathbb{P}\mathbf{F}(M(\widehat{\mathbf{u}}_h)) =: r_s,\ \widehat{\mathbf{u}}_h : \Omega_s \times \mathbb{R}^+ \to \mathbb{R}^n,$$

where r_s is the discretization residual in the simple system.

Theorem 1 *Let $\mathbf{U} : \mathbb{R} \times \mathbb{R}^+ \to \mathbb{R}^N$ be the exact solution to* (1)*. Let $\widehat{\mathbf{U}}_h$ be the Lipschitz reconstruction of the numerical solution of* (1) *on Ω_c and let $\widehat{\mathbf{u}}_h$ be the Lipschitz reconstruction of the numerical solution of* (4) *on Ω_s. We assume that for some $\nu = \nu(M)$*

$$-\left(\frac{\partial H}{\partial \mathbf{U}}(\mathbf{U}) - \frac{\partial H}{\partial \mathbf{U}}\left(M\left(\mathbb{P}\mathbf{U}\right)\right)\right) \cdot \left(\mathbf{R}(\mathbf{U}) - \mathbf{R}\left(M\left(\mathbb{P}\mathbf{U}\right)\right)\right) \geq \nu \left|\mathbf{U} - M(\mathbb{P}\mathbf{U})\right|^2 \tag{12}$$

for $\mathbf{U} \in \mathbb{R}^N$. Furthermore, let for any $\mathbf{U}, \mathbf{V} \in \mathbb{R}^N$

$$-\left(\frac{\partial H}{\partial \mathbf{U}}(\mathbf{U}) - \frac{\partial H}{\partial \mathbf{U}}\left(\mathbf{V}\right)\right) \cdot \left(\mathbf{R}(\mathbf{U}) - \mathbf{R}(\mathbf{V})\right) \geq 0. \tag{13}$$

Then, we have

$$\begin{aligned} &\int_{\Omega_c} \left|\mathbf{U} - \widehat{\mathbf{U}}_h\right|^2 \mathrm{d}x \,\bigg|_t + \int_{\Omega_s} \left|\mathbf{U} - M(\widehat{\mathbf{u}}_h)\right|^2 \mathrm{d}x \,\bigg|_t \\ &\quad \leq \left(I + D_c + D_s + \mathcal{M}_s + C_Q\right) \exp\left(\frac{\max(C_c, C_s + 1 + |\mathbb{P}|)}{c} t\right), \end{aligned} \tag{14}$$

where

$$I = \int_{\Omega_s} H(\mathbf{U}|M(\widehat{\mathbf{u}}_h))\bigg|_{x,t=0} \mathrm{d}x + \int_{\Omega_c} H(\mathbf{U}|\widehat{\mathbf{U}}_h)\bigg|_{x,t=0} \mathrm{d}x,$$

$$
\begin{aligned}
D_c &= \int_0^t \int_{\Omega_c} \left|\nabla^2_{\mathbf{U}} H(\widehat{\mathbf{U}}_h) r_c\right|^2 \mathrm{d}x\,\mathrm{d}\tau,\\
D_s &= \int_0^t \int_{\Omega_s} \left|\nabla^2_{\mathbf{u}} \eta(\widehat{\mathbf{u}}_h) r_s\right|^2 \mathrm{d}x\,\mathrm{d}\tau,\\
\mathcal{M}_s &= \frac{\varepsilon}{\nu} \int_0^t \int_{\Omega_s} |\partial_x\left(\nabla_{\mathbf{u}}\eta(\widehat{\mathbf{u}}_h)\right) \cdot \mathbb{P}\nabla_{\mathbf{U}}\mathbf{F}(M(\widehat{\mathbf{u}}_h))|^2\,\mathrm{d}x\,\mathrm{d}\tau,\\
C_Q &= \int_0^t Q(\mathbf{U}|M(\widehat{\mathbf{u}}_h))\Big|_{x_q,\tau} \mathrm{d}\tau - \int_0^t Q(\mathbf{U}|M(\widehat{\mathbf{u}}_h))\Big|_{x_q,\tau} \mathrm{d}\tau,\\
C_c &= \left|\left|\partial_x\left(\nabla_{\mathbf{U}} H(\widehat{\mathbf{U}}_h)\right)\nabla^2_{\mathbf{U}}\mathbf{F}(\widehat{\mathbf{U}}_h)\right|\right|_\infty + \left|\left|\frac{1}{\varepsilon}\mathbf{R}(\widehat{\mathbf{U}}_h)\nabla^3_{\mathbf{U}} H(\widehat{\mathbf{U}}_h)\right|\right|_\infty,\\
C_s &= \left|\left|\partial_x\left(\nabla_{\mathbf{u}}(\eta(\widehat{\mathbf{u}}_h))\right)\nabla^2_{\mathbf{u}}\left(\mathbb{P}\mathbf{F}(\widehat{\mathbf{u}}_h)\right)\right|\right|_\infty.
\end{aligned}
$$

One can note that the D terms are the discretization errors and the term $\mathcal{M}_s$ is the modelling error in the simple system. More details and analysis of the estimates will be provided in [9].

4 Approximate *Maxwellian*

As discussed previously, the *Maxwellian* needs to be employed to calculate the flux in the simple system. Calculating the *Maxwellian* can be expensive as a non-linear system of equations needs to be solved. Hence to further reduce the computational expenses when solving the simple system an approximate *Maxwellian* $\tilde{M}$ can be devised whereby we solve the following system of equations

$$\partial_t\tilde{\mathbf{u}} + \partial_x\mathbb{P}\mathbf{F}(\tilde{M}(\tilde{\mathbf{u}})) = 0, \quad \tilde{\mathbf{u}} : \mathbb{R}\times\mathbb{R}^+ \to \mathbb{R}^n. \tag{15}$$

To this end, error analysis of the system using an approximate *Maxwellian* is presented in Sect. 4.1, followed by a discussion about the construction of an approximate *Maxwellian* in Sect. 4.2.

4.1 Convergence Analysis

The relative entropy framework requires an entropy-entropy flux pair such that the compatibility condition (5) is satisfied. Hence, it is desirable for the approximate *Maxwellian* to be defined so that for some $\tilde{Q}$ the following holds

$$\mathrm{D}\,H(\tilde{M}(\tilde{\mathbf{u}}))\,\mathrm{D}\,\mathbf{F}(\tilde{M}(\tilde{\mathbf{u}})) = \mathrm{D}\,\tilde{Q}(\tilde{M}(\tilde{\mathbf{u}}). \tag{16}$$

But it is unclear how to define such a $\tilde{Q}$. Hence, the compatibility condition can be approximately satisfied by employing Q such that

$$\mathrm{D}\, H(\tilde{M}(\tilde{\mathbf{u}}))\, \mathrm{D}\,\mathbf{F}(\tilde{M}(\tilde{\mathbf{u}})) - \mathrm{D}\, Q(\tilde{M}(\tilde{\mathbf{u}})) =: \delta_\eta(\tilde{\mathbf{u}}). \tag{17}$$

Here δ_η is the mismatch in the compatibility condition when an approximate Maxwellian and the definition of original entropy flux is used.
Furthermore, let $\tilde{\eta}(\tilde{\mathbf{u}}) := H(\tilde{M}(\tilde{\mathbf{u}})),\ \tilde{q}(\tilde{\mathbf{u}}) := Q(\tilde{M}(\tilde{\mathbf{u}}))$, then any Lipschitz continuous $\tilde{\mathbf{u}}$ of (15) satisfies

$$\partial_t \tilde{\eta}(\tilde{\mathbf{u}}) + \partial_x \tilde{q}(\tilde{\mathbf{u}}) = -\delta_\eta(\tilde{\mathbf{u}})\partial_x \tilde{\mathbf{u}}. \tag{18}$$

Theorem 2 *Let* $\mathbf{u} : \mathbb{R} \times \mathbb{R}^+ \to \mathbb{R}^n$ *be the exact solution to* (4) *on* Ω *and let* $\tilde{u}$ *be a Lipschitz continuous solution of* (15) *on* Ω*. Then, we have*

$$\int_\Omega \left| M(\mathbf{u}) - \tilde{M}(\tilde{\mathbf{u}}) \right|^2 \mathrm{d}x \,\Bigg|_t \le \left(\tilde{I} + \tilde{\mathcal{M}} \right) \exp\left(\frac{\tilde{C}}{c} t \right), \tag{19}$$

where c is a non-negative constant as in (10) *and*

$$\begin{aligned} \tilde{I} &= \textstyle\int_\Omega H(\tilde{M}(\tilde{\mathbf{u}})|M(\mathbf{u})) \,\Bigg|_{x,t=0} \mathrm{d}x, \\ \tilde{C} &= \left|\left| \partial_x \left(\nabla_{\mathbf{u}}(\eta(\tilde{\mathbf{u}})) \right) \nabla^2_{\mathbf{u}} \left(\mathbb{P}\mathbf{F}(\tilde{\mathbf{u}}) \right) \right|\right|_\infty, \\ \tilde{\mathcal{M}} &= \textstyle\int_0^t \int_\Omega \left| \delta_\eta(\tilde{\mathbf{u}}) \partial_x \tilde{\mathbf{u}} \right| \mathrm{d}x\, \mathrm{d}\tau. \end{aligned}$$

Moreover, if $\left|\left| M(\mathbf{u}) - \tilde{M}(\tilde{\mathbf{u}}) \right|\right|_\infty \le \varepsilon_{\tilde{M}}$ and $\left|\left| \mathrm{D}\, M(\mathbf{u}) - \mathrm{D}\, \tilde{M}(\tilde{\mathbf{u}}) \right|\right|_\infty \le \varepsilon_{\mathrm{D}\tilde{M}}$ then $\left|\left|\delta_\eta\right|\right|_\infty \le C\varepsilon_{\tilde{M}}\varepsilon_{\mathrm{D}\tilde{M}}$, where C is bounded by the maximum of the absolute values of eigenvalues of $\mathrm{D}\,\mathbf{F}$. Furthermore, as δ_η vanishes $\tilde{M}$ converges to M and the solution of system using the approximate *Maxwellian* (15) converges to the solution of system using the exact *Maxwellian* (4).

4.2 Construction of an Approximate Maxwellian

The approximate *Maxwellian* should be constructed such that the resulting numerical solution is as close as possible to the numerical solution if the original system would have been employed. For this, the quantity δ_η needs to be minimized. An approximate *Maxwellian* should be constructed by balancing the computational resources needed to evaluate it and its distance to the exact *Maxwellian*.

Previous research investigates the stability properties due to a change in the flux based on standard Riemann semigroup in [10]. But in our case, the approximate

Maxwellian should be devised by accounting for the structure of δ_η as discussed in Sect. 4.1. Furthermore, in the case of chemically reacting flows, physical constraints such as positivity (e.g. density, total energy, temperature), compliance with the pressure law and positive entropy dissipation should be satisfied. Moreover, according to the definition the source terms should vanish on the equillibrium manifold in the state space of the complex system. When $\tilde{M}$ is constructed extra attention has to be paid to ensure this structure is preserved. If $\tilde{M}$ does not exactly satisfy these constraints it will lead to added contributions to δ_η. The distance of the approximate *Maxwellian* to the exact *Maxwellian* can be kept small if the mismatch in the constraints is bounded and small or if $\tilde{M}$ exactly satisfies the constraints.

One approach to construct $\tilde{M}$ is to employ machine learning techniques such as neural networks. Neural networks are trained on some training data sets, and then are used as black boxes for new data. Generally the neural nets are constructed by minimizing some loss function on training data sets such as the L_2 error between the outputs produced by the neural networks and the expected exact values. As a result even though the neural nets are constructed on training data sets which satisfy the constraints, the outputs produced by the neural networks on the new data do not necessarily satisfy them. Hence we need to employ machine learning techniques that account for the constraints on the data sets.

In [11] machine learning techniques were developed where natural symmetries of systems such as rotational, translational invariance are preserved. In [12] constraint aware neural networks were developed termed as Constraint Resolving layer method (CRes), in which an extra constraint layer is added to produce a constraint compliant neural network. This requires the constraint to be re-written in a form like that in the implicit function theorem. A second approach is to add penalty terms to the loss function, which is then minimized. In this approach the constraints may not be exactly satisfied.

In the instance of chemically reacting flows the CRes method can be employed, whereby a series of constraint layers are implemented in the neural network for each of the constraints. With this approach a $\tilde{M}$ can be constructed which is mass, energy conservative and compliant with the pressure law. A sequence of constraint layers can be employed as the outputs can be calculated sequentially. But vanishing source terms on the equillibrium manifold cannot be enforced as the constraint cannot be written in the required format. But this precise information is provided to us by the error estimates. Furthermore, a constraint can be added as a penalty term to the loss function to minimize the distance of $\tilde{\mathbf{U}}$ from the equillibrium manifold. With such an approach an approximate *Maxwellian* can be constructed that is close to the exact one and also satisfies the desirable constraints.

The strategy described scales down the computational resources needed by carrying out model adaptation and employing machine learning to construct an approximate mapping.

Remark 2 In assessing the computational efficiency of the model adaptive scheme compared to solving the complex system everywhere one needs to compare the costs and savings. The costs consists of computing the reconstruction (local, involves matrix vector multiplication) and computing the residual (evaluating non-linear functions). The savings result from being able to avoid stiff source terms in areas where the simple system is solved. Thus, in these areas explicit schemes can be used instead of implicit schemes, such that no non-linear function needs to be solved in time stepping. Note that the non-linear function whose roots need to be found in each time step is the same as the one evaluated in computing residuals. By how much the gains outweigh the costs depends on the complexity of this non-linear function, e.g. for chemical reacting flows it depends on the number and the complexity of the chemical reactions.

More details and numerical results and computational resource analysis will be presented in [9].

Acknowledgements This research is supported by the German Research Foundation (DFG) grant GI1131/1-1: Dynamical, spatially heterogeneous model adaptation in compressible flows.

References

1. Bothe, D., Dreyer, W.: Continuum thermodynamics of chemically reacting fluid mixtures. Acta Mech. **226**, 1757–1805 (2015)
2. Hantke, M., Müller, S.: Analysis and simulation of a new multi-component two-phase flow model with phase transitions and chemical reactions. Quart. Appl. Math. **76**, 253–287 (2018)
3. Müller, I., Müller, W.H.: Fundamentals of thermodynamics and applications, 1st edn. Springer, Berlin (2009)
4. Dafermos, C.M.: Hyperbolic Conservation Laws in Continuum Physics, 3rd edn. Grundlehren der Mathematischen Wissenschaften, vol. 325. Springer, Berlin (2010)
5. De Lellis, C., Székelyhidi, L.: On admissibility criteria for weak solutions of the Euler equations. Arch. Ration. Mech. Anal. **95**, 225–260 (2010)
6. Tzavaras, A.: Relative entropy in hyperbolic relaxation. Commun. Math. Sci **3**, 119–132 (2005)
7. Giesselmann, J., Makridakis, C., Pryer, T.: A posteriori analysis of discontinuous Galerkin schemes for systems of hyperbolic conservation laws. SIAM J. Numer. Anal. **53**, 1280–1303 (2015)
8. Giesselmann, J., Pryer, T.: A posteriori analysis for dynamic model adaptation problems in convection dominated problems. Math. Models Methods Appl. Sci. **27**, 2381–2423 (2017)
9. Giesselmann, J., Joshi, H.: A posteriori error analysis for model adaptation of hyperbolic systems with relaxation, in preparation
10. Bianchini, S., Colombo, R.M.: On the stability of the standard Reimann semigroup. Proc. Amer. Math. Soc. **130**, 1961–1973 (2002)
11. Zhang, L., Han, J., Wang, H., Saidi, W.A., Car, R., Weinan, E.: End-to-end symmetry preserving inter-atomic potential energy model for finite and extended systems. NIPS. **31** (2018)
12. Magiera, J., Ray, D., Hesthaven, J.S., Rohde, C.: Constraint-aware neural networks for Riemann problems. J. Comput. Phys. **409** (2020)

Robust Newton Solver Based on Variable Switch for a Finite Volume Discretization of Richards Equation

Sabrina Bassetto, Clément Cancès, Guillaume Enchéry, and Quang Huy Tran

Abstract We propose an efficient nonlinear solver for the resolution of the Richards equation. It is based on variable switching and is easily implemented thanks to a fictitious variable allowing to describe both the saturation and the pressure. Numerical experiments show that our method enables to use Newton's method with large time steps, reasonable number of iterations and in regions where the pressure-saturation relationship is given by a graph.

Keywords Degenerate parabolic equation · Nonlinear solver · Variable switch

MSC (2010) 65M08 · 65N08 · 35Q30

1 Finite Volume Approximation of the Richards Equation

The Richards equation is often used to model unsaturated flows in a porous medium $\Omega \subset \mathbb{R}^d$ $(1 \le d \le 3)$. The fluid occupying the pore space is described by the pressure $p \in \mathbb{R}$ of the water phase and the water saturation $s \in [0, 1]$, which represents the volume ratio of water in the pore space. The conservation law for the water volume then writes

S. Bassetto (✉) · G. Enchéry · Q. H. Tran
IFPEN, 1 & 4, avenue du Bois Préau, 92852 Rueil-Malmaison Cedex, France
e-mail: sabrina.bassetto@ifpen.fr

G. Enchéry
e-mail: guillaume.enchery@ifpen.fr

Q. H. Tran
e-mail: quang-huy.tran@ifpen.fr

C. Cancès
Inria, University of Lille, CNRS, UMR 8524-Laboratoire Paul Painlevé,
59000 Lille, France
e-mail: clement.cances@inria.fr

R. Klöfkorn et al. (eds.), *Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples*, Springer Proceedings in Mathematics & Statistics 323,
https://doi.org/10.1007/978-3-030-43651-3_35

$$\partial_t(\phi\, s) - \operatorname{div}\left(\frac{\lambda}{\mu} k_r(s)(\nabla p - \varrho \mathbf{g})\right) = 0 \quad \text{in } \Omega \times \mathbb{R}_+, \tag{1}$$

where $\phi \in (0, 1)$ is the porosity of Ω, λ its intrinsic permeability, μ the water viscosity, ϱ the water density and $\mathbf{g}$ the gravitational acceleration. The relative permeability function $k_r : [0, 1] \to \mathbb{R}^+$ is continuous and nondecreasing, and we denote by $s_{\text{rw}} = \max\{s \mid k_r(s) = 0\}$ the residual water saturation. The saturation s and pressure p are linked pointwise by the relation

$$s = \mathscr{S}(p) \quad \text{in } \Omega \times \mathbb{R}_+, \tag{2}$$

where $\mathscr{S} : \mathbb{R} \to [0, 1]$ is nondecreasing and satisfies $\mathscr{S}(p) = 1 - s_{\text{rn}}$ if $p \geq p_b$, s_{rn} denoting the residual saturation of air, p_b the entry pressure and $\mathscr{S}(p) \to s_{\text{rw}}$ as $p \to -\infty$. We assume that $\mathscr{S}$ is C^1 and convex on $(-\infty, p_{\text{s}})$, and C^1 and concave on $(p_{\text{s}}, +\infty)$ for some $p_{\text{s}} \leq 0$. We denote by $s_{\text{s}} = \mathscr{S}(p_{\text{s}})$. The above assumptions on k_r and $\mathscr{S}$ are satisfied by the classical Brooks-Corey and van Genuchten-Mualem models respectively given by

$$k_{r\,\text{BC}}(s) = s_{\text{eff}}^{3+\frac{2}{n}}, \qquad \mathscr{S}_{\text{BC}}(p) = \begin{cases} s_{\text{rw}} + (1 - s_{\text{rn}} - s_{\text{rw}}) \left(\frac{p}{p_b}\right)^{-n} & \text{if } p \leq p_b, \\ 1 - s_{\text{rn}} & \text{if } p > p_b, \end{cases} \tag{3}$$

$$k_{r\,\text{vGM}}(s) = s_{\text{eff}}^{\frac{1}{2}}\{1 - [1 - s_{\text{eff}}^{\frac{1}{m}}]^m\}^2, \quad \mathscr{S}_{\text{vGM}}(p) = \begin{cases} s_{\text{rw}} + \frac{1 - s_{\text{rn}} - s_{\text{rw}}}{\left[1 + \left|\frac{\alpha}{\rho g} p\right|^n\right]^m} & \text{if } p \leq p_b, \\ 1 - s_{\text{rn}} & \text{if } p > p_b, \end{cases} \tag{4}$$

where $s_{\text{eff}} = \frac{s - s_{\text{rw}}}{1 - s_{\text{rn}} - s_{\text{rw}}}$ and, for the van Genuchten-Mualem model, $m = 1 - \frac{1}{n}$ and $p_b = 0\ Pa$. Dirichlet boundary conditions are imposed on a part Γ^D of $\partial\Omega$, while inflow Neumann boundary conditions are imposed on the complement $\Gamma^N = \partial\Omega \setminus \Gamma^D$:

$$p = p_D \ \text{ on } \Gamma^D \times \mathbb{R}_+, \qquad -\frac{\lambda}{\mu} k_r(s)(\nabla p - \varrho \mathbf{g}) \cdot \mathbf{n} = q_N \ \text{ on } \Gamma^N \times \mathbb{R}_+, \tag{5}$$

with $q_N \leq 0$. Finally, the system is closed by prescribing an initial saturation profile

$$s(\cdot, 0) = s^0 \quad \text{in } \Omega, \quad \text{with} \quad 0 \leq s^0 \leq 1. \tag{6}$$

We refer to [2] for further details on the modeling and to [1] for the well-posedness of the problem.

The problem (1), (2), (5), and (6) is discretized by means of a finite-volume scheme: an upstream mobility is used for convection and a two-point flux approximation (TPFA) for the capillary diffusion. Let $(\mathscr{T}, \mathscr{E}, (\mathbf{x}_K)_{K \in \mathscr{T}})$ be a finite volume mesh of Ω fulfilling the classical orthogonality condition required for the consistency of TPFA. Since this notion is classical, we remain sloppy here on the definition and refer to [6, Definition 9.1] for details. Let us just mention that $\mathscr{T}$ denotes the set of the cells, while the set of the edges $\mathscr{E}$ is partitioned into the set of the internal edges $\mathscr{E}_{\text{int}} = \{\sigma \in \mathscr{E} \mid \sigma = K|L = \partial K \cap \partial L\}$, the set of the Dirichlet boundary edges $\mathscr{E}_{\text{ext}}^D = \{\sigma \in \mathscr{E} \mid \sigma \subset \Gamma^D\}$, and the set of the Neumann boundary edges $\mathscr{E}_{\text{ext}}^N = \{\sigma \in \mathscr{E} \mid \sigma \subset \Gamma^N\}$. We denote by $\mathscr{E}_K = \{\sigma \in \mathscr{E} \mid \sigma \subset \partial K\}$. For the time discretization, we allow for non-uniform time steps $\tau_n = t^n - t^{n-1}$, $n \geq 1$. At initial time $t = 0$, s^0 is discretized into $s_K^0 = \frac{1}{|K|}\int_K s^0$. For $\sigma \in \mathscr{E}_{\text{int}} \cup \mathscr{E}_{\text{ext}}^D$, $\sigma \in \mathscr{E}_K$, we define the mirror value $u_{K,\sigma}^n$ of u_K^n across σ by $u_{K,\sigma}^n = u_L^n$ if $\sigma = K|L \in \mathscr{E}_{\text{int}}$ and $u_{K,\sigma}^n = u_\sigma^n = \frac{1}{\tau_n|\sigma|}\int_\sigma \int_{t^{n-1}}^{t^n} u^D$ if $\sigma \in \mathscr{E}_{\text{ext}}^D$. The conservation of the water phase is discretized into

$$\phi_K \frac{s_K^n - s_K^{n-1}}{\tau_n}|K| + \sum_{\sigma \in \mathscr{E}_K} F_{K\sigma}^n = 0, \qquad K \in \mathscr{T},\ n \geq 1. \tag{7}$$

The expression of the fluxes relies on a unique upwinding for capillary diffusion and for gravitationally induced convection, that is

$$F_{K\sigma}^n = \begin{cases} A_\sigma \left\{ \frac{k_{r\,\sigma,up}^{\,n}}{\mu}\left[\,(p_K^n - p_{K,\sigma}^n) \;+\; \rho g\,(z_K - z_{K,\sigma})\,\right]\right\} & \text{if } \sigma \in \mathscr{E}_{\text{int}} \cup \mathscr{E}_{\text{ext}}^D,\\ \frac{1}{\tau_n}\int_{t^{n-1}}^{t^n}\int_\sigma q_N & \text{if } \sigma \in \mathscr{E}_{\text{ext}}^N, \end{cases} \tag{8}$$

where

$$k_{r\,\sigma,up}^{\,n} = \begin{cases} k_r(s_K^n) & \text{for } (p_K^n - p_{K,\sigma}^n) \;+\; \rho g\,(z_K - z_{K,\sigma}) \geq 0,\\ k_r(s_{K,\sigma}^n) & \text{otherwise}, \end{cases} \tag{9}$$

$$A_\sigma = \begin{cases} m_\sigma \frac{\lambda_K\lambda_L}{\lambda_L d_{K,\sigma} + \lambda_K d_{L,\sigma}} & \text{if } \sigma = K \cap L,\\ m_\sigma \frac{\lambda_K}{d_{K,\sigma}} & \text{if } \sigma \in \mathscr{E}_{\text{ext}}^D, \end{cases} \tag{10}$$

with $\lambda_K = \lambda(\mathbf{x}_K)$, $d_{K,\sigma} = |\mathbf{x}_K - \mathbf{x}_L|$ if $\sigma = K|L \in \mathscr{E}_{\text{int}}$, $d_{K,\sigma} = \text{dist}(\mathbf{x}_K, \sigma)$ if $\sigma \in \mathscr{E}_{\text{ext}}^D$ and m_σ is the Lebesgue measure of the edge σ. The discrete water saturation and pressure are related cellwise by the relation

$$s_K^n = \mathscr{S}(p_K^n), \qquad K \in \mathscr{T},\ n \geq 1. \tag{11}$$

The scheme (7)–(11) admits a unique discrete solution $\left(s_K^n, p_K^n\right)_{K\in\mathscr{T}}$ for all $n \geq 1$ and converges as the mesh size and the time step tend to 0 (this will be proved in a forthcoming work). In this contribution, we rather focus on the practical resolution of the nonlinear system (7)–(11) via an iterative method. For our works, we choose to use Newton's method. Notice that the physical models presented above, both feature two difficulties for Newton's method: the function $\mathscr{S}_{\text{BC}}$ is Lipschitz continuous but not C^1 and the mobility function $k_{r\,\text{vGM}}$ is singular at $s = 1 - s_{\text{rn}}$ where the derivative blows up.

2 Fictitious Variable and Newton's Method

A natural approach to solve the nonlinear system (7)–(11) is to choose $(p_K)_{K\in\mathscr{T}}$ as a primary unknown and to solve the corresponding nonlinear system thanks to Newton's method (or alternatively some modified Picard's method, see e.g. [9]). However, the choice of the pressure as the primary variable is known to be inefficient for dry soils $s \ll 1$ where they are outperformed by schemes using s as primary variable. On the other hand, the knowledge of the saturation is not sufficient to describe the pressure in saturated regions where $s = 1$. This motivated the introduction of schemes based on variable switching between s and p, see [5, 7]. Our approach is based on [3] and can be seen as a reformulation of the variable switch which makes its implementation much easier. Unlike in [3], we do not use the Kirchhoff transform which cannot be easily computed for the van Genuchten-Mualem model. The idea is to choose a parametrization of the graph $\{p, \mathscr{S}(p)\}$, i.e. to choose two functions $\mathfrak{s} : I \to [s_{\mathrm{rw}}, 1 - s_{\mathrm{rn}}]$ and $\mathfrak{p} : I \to \mathbb{R}$ such that $\mathfrak{s}(u) = \mathscr{S}(\mathfrak{p}(u))$ for all $u \in I \subset \mathbb{R}$. Such a parametrization is not unique: one can for instance choose $I = \mathbb{R}$, $\mathfrak{p} = \mathrm{Id}$ and $\mathfrak{s} = \mathscr{S}$, or $\mathfrak{p} = (\mathrm{Id} + \mathscr{S})^{-1}$ and $\mathfrak{s} = (\mathrm{Id} + \mathscr{S}^{-1})^{-1}$ so that $\mathfrak{s}'(u) + \mathfrak{p}'(u) = 1$ for all $u \in \mathbb{R}$. Here, we rather set $I = (s_{\mathrm{rw}}, +\infty)$ and

$$\mathfrak{s}(u) = \begin{cases} u & \text{if } u \le u_{\mathrm{s}}, \\ \mathscr{S}\left(p_{\mathrm{s}} + \dfrac{u - u_{\mathrm{s}}}{\mathscr{S}'(p_{\mathrm{s}}^{-})}\right) & \text{if } u \ge u_{\mathrm{s}}, \end{cases} \qquad \mathfrak{p}(u) = \begin{cases} \mathscr{S}^{-1}(u) & \text{if } u \le u_{\mathrm{s}}, \\ p_{\mathrm{s}} + \dfrac{u - u_{\mathrm{s}}}{\mathscr{S}'(p_{\mathrm{s}}^{-})} & \text{if } u \ge u_{\mathrm{s}}. \end{cases} \tag{12}$$

where $\mathscr{S}'(p_{\mathrm{s}}^{-})$ denotes the limit of $\mathscr{S}'(p)$ as p tends to p_{s} from below. Since $(p_{\mathrm{s}}, u_{\mathrm{s}})$ is the inflexion point of $\mathscr{S}$, both $\mathfrak{s}$ and $\mathfrak{p}$ are C^1 and concave, and even C^2 if $\mathscr{S}$ is given by (4). Moreover, for all $p \in \mathbb{R}$, there exists a unique $u \in (s_{\mathrm{rw}}, +\infty)$ such that $(p, \mathscr{S}(p)) = (\mathfrak{p}(u), \mathfrak{s}(u))$.

Choosing u as a primary variable in the scheme (7)–(11) amounts to search for $\mathbf{u}^n = \left(u_K^n\right)_{K\in\mathscr{T}}$ such that $s_K^n = \mathfrak{s}(u_K^n)$ and $p_K^n = \mathfrak{p}(u_K^n)$ for all $K \in \mathscr{T}$. Equation (11) is then automatically satisfied. The resulting system $\mathscr{F}_n(\mathbf{u}^n) = \mathbf{0}$ made of $N_{\mathscr{T}} = \mathrm{Card}(\mathscr{T})$ nonlinear equations admits a unique solution $\mathbf{u}^n$ since it is fully equivalent to (7)–(11). However, the nonlinear change of variable to pass from $\mathbf{p}^n = \left(p_K^n\right)_{K\in\mathscr{T}}$ to $\mathbf{u}^n$ as primary variable strongly impacts the nonlinear solver. Our approach is based on Newton's method, that is detailed in Algorithm 1 and that include the following procedures in order to handle difficulties which are inherent to the chosen petro-physical models.

- *check*() and *update*()
  The law of the relative permeability k_r, in the van Genuchten-Mualem case (4), has very large derivative values, which can be equal to ∞ for $s \to 1$. In order to overcome this difficulty, we approximate $k_{r\,\mathrm{vGM}}$, during Newton's iterations, for $s \in N = [s_{\mathrm{lim}}, 1]$, with a polynomial $\tilde{k}_{r\,\mathrm{vGM}}(s)$ of second degree which satisfies the following conditions: $k_r(s_{\mathrm{lim}}) = \tilde{k}_{r\,\mathrm{vGM}}(s_{\mathrm{lim}})$, $k_r'(s_{\mathrm{lim}}) = \tilde{k}_{r\,\mathrm{vGM}}'(s_{\mathrm{lim}})$, $k_r''(s_{\mathrm{lim}}) = \tilde{k}_{r\,\mathrm{vGM}}''(s_{\mathrm{lim}})$. The idea is to progressively increase the value of s_{lim} in order to

recover the real law at convergence. The function $check()$ verifies the error we commit in the approximation. If this error is smaller than a fixed tolerance, namely $|k_{r\text{vGM}}(1) - \tilde{k}_{r\text{vGM}}(1)| < \varepsilon_{k_r^{\text{vGM}}}$, it returns true, false otherwise. At each Newton's iteration, we increase the value of s_{lim} thanks to the function $update()$. The increment speed depends on the norm of the residual. Let us call $\delta_{s_{w,max}} = 1 - s_{\text{rn}} - s_{\text{lim}}$. If $\left\|\mathscr{F}_n(\mathbf{u}^{n,i})\right\|_\infty > \varepsilon_{\mathscr{F}_{\text{vGM}}}$ we set $\delta_{s_{w,max}} = \delta_{s_{w,max}} \cdot \omega$ and $\delta_{s_{w,max}} = \delta^2_{s_{w,max}}$ otherwise, with $\omega < 1$.

- $truncation()$
  Since $\mathscr{F}_n$ is not necessarily C^1 ($\mathscr{S}_{\text{BC}}$ is not C^1 in the Brooks and Corey case), following [8, 10], the Newton increment is truncated near the inflection point s_{s}, as described in Algorithm 2.
- $decreaseDeltaTime()$ and $increaseDeltaTime()$
  In our numerical tests, we increase the time step in such a way that $\Delta t^{n+1} = \min(\Delta t_{max}, \alpha^+_{\Delta t} \cdot \Delta t^n)$ and decrease it in such a way that $\Delta t^{n+1} = \max(\alpha^-_{\Delta t} \cdot \Delta t^n, \Delta t_{min})$ with $\alpha^+_{\Delta t} > 1$ and $\alpha^-_{\Delta t} < 1$. If Δt_{min} is reached, the simulation stops.

```
Initialization:
$i = 0$;
$\mathbf{u}^{n,0} = \mathbf{u}^{n-1}$;

while
$\left[\left(\left\|\mathscr{F}_n(\mathbf{u}^{n,i})\right\|_\infty \geq \varepsilon \wedge i \leq i_{\max}\right) \vee \neg\, check()\right]$
do
    solve $\mathbb{J}(\mathbf{u}^{n,i-1})\delta^{n,i} + \mathscr{F}_n(\mathbf{u}^{n,i-1}) = \mathbf{0}$ ;
    for $K \in \mathscr{T}$ do
        $truncation()$;
        $u_K^{n,i} = \max(s_{\text{rw}}, u_K^{n,i-1} + \delta_K^{n,i})$;
    end
    $i = i + 1$;
    $update()$;
end
if $i > i_{\max}$ then
    $decreaseDeltaTime()$;
    restart while loop ;
else
    $u^n = u^{n,i}$;
    $n = n + 1$;
    $increaseDeltaTime()$;
end
```

Algorithm 1: Practical resolution of the system $\mathscr{F}_n(\mathbf{u}^n) = \mathbf{0}$, where $\mathbb{J}$ is the Jacobian matrix.

```
for $K \in \mathscr{T}$ do
    if $s_{\text{s}} - \delta_K^{n,i} < u_K^{n,i-1} \leq s_{\text{s}}$ then
        $\delta_K^{n,i} = s_{\text{s}} - u_K^{n,i-1} + \varepsilon_\delta$;
    else if $s_{\text{s}} \leq u_K^{n,i-1} < s_{\text{s}} - \delta_K^{n,i}$
    then
        $\delta_K^{n,i} = s_{\text{s}} - u_K^{n,i-1} - \varepsilon_\delta$;
    end
end
```

Algorithm 2: Detail of the function $truncation()$, where $\varepsilon_\delta \ll 1$.

3 Numerical Results

For the numerical validation of our scheme, we consider two tests inspired from those proposed in [4]. These two tests make use of the classical Brooks and Corey and Van Genuchten-Mualem models. For the simulations we take the following parameters: $\varepsilon = 10^{-12}$, $i_{\max} = 500$, $\varepsilon_{k_r^{\mathrm{vGM}}} = 10^{-3}$, $\varepsilon_{\mathscr{F}_{\mathrm{vGM}}} = 10^{-9}$, $\varepsilon_{\delta} = 10^{-6}$, $\alpha^{-}_{\Delta t} = 0.5$, $\omega = 0.07$. As in [4], our aim is here just to improve the robustness of the Newton's algorithm when used with the TPFA scheme. Therefore, our study here focuses on the corresponding nonlinear system even if more accurate schemes could be used to better take into account the heterogeneities, in particular the ones related to the capillary pressures.

3.1 Test 1 with the Brooks and Corey model

In this test we simulate a vertical drainage through a layered soil $\Omega = [0 \text{ m}, \ 2 \text{ m}]$ from initially saturated conditions during a time interval $[0, T]$ with $T = 105 \cdot 10^4$ s. At the initial time the pressure varies with respect to the height of the column, that is $p^0(z) = -\varrho g(z-2)$, where $\varrho = 10^3 \text{ kg} \cdot \text{m}^{-3}$ and $g = 9.81 \text{ ms}^{-2}$. During the simulation, we impose a Dirichlet boundary condition $p_D = 0$ Pa on the bottom of the column and a no-flow boundary condition on the top. The soil is made of two rock types: RT1 for $0 \text{ m} < z < 0.6$ m and $1.2 \text{ m} < z < 2$ m, and RT2 for $0.6 \text{ m} < z < 1.2$ m. Their hydraulic properties are given in Table 1. Simulations are performed on a mesh with 1000 cells and an initial time step $\Delta t_{ini} = 2000\ s$ which increases after the first time iteration up to $\Delta t_{\max} = 2 \cdot \Delta t_{ini}$ using $\alpha^{+}_{\Delta t} = 1.2$. The truncation procedure, detailed in Algorithm 2, is activated during Newton's iterations. Table 2 gives the average number of iterations of the nonlinear solvers used here and in [4] along with the number of time steps.
Note that a coarser mesh has been used in [4] for this test. Solutions obtained at the final time are shown in Fig. 1. In some areas, pressures are higher than the entry pressure and the saturation-pressure relationship is there no more a function. The problem can still be solved thanks to the use of the parametrization technique. Figure 2 shows the evolution of the average Newton's convergence rate given, for a time step

Table 1 Hydraulic properties for Test 1

	$1 - s_{\mathrm{rn}}$	s_{rw}	p_b[Pa]	n	$\lambda[\mathrm{m}^2]$	ϕ
RT1	1.0	0.2	$-3.4301 \cdot 10^3$	1.5	10^{-11}	0.35
RT2	1.0	0.1	$-1.4708 \cdot 10^3$	3.0	10^{-9}	0.35

Table 2 Performances of the nonlinear (nl) solvers for Test 1

	♯ Total nl iterations	♯ Time iterations
Our method	1118	265
Method proposed in [4] (coarser mesh)	4469 (inner iterations)	300

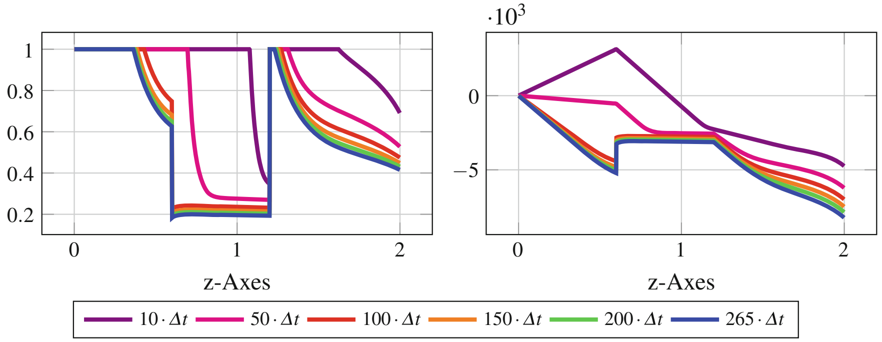


Fig. 1 Evolution in time of the saturation on the left and of the pressure on the right

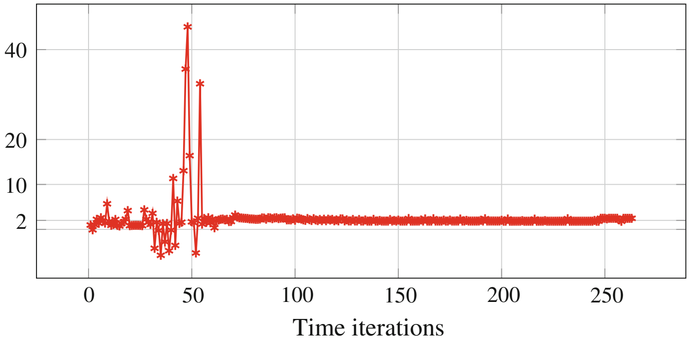


Fig. 2 Test 1: evolution of the average Newton's convergence rate during time iterations

n, by $CV^n_{rate} = \frac{1}{N^n_{iter}} \sum_{i=1}^{N^n_{iter}} \frac{\log_{10} \left\| \mathscr{F}_n(\mathbf{u}^{n,i}) \right\|_\infty}{\log_{10} \left\| \mathscr{F}_n(\mathbf{u}^{n,i-1}) \right\|_\infty}$. The rate is on the whole equal to 2. Negative rates are due to residual norms which are greater than one at some iterations.

3.2 Test 2 with the Van Genuchten-Mualem model

In this test, starting from an initially very dry layered domain, $\Omega = [0\text{ m}, 1\text{ m}] \times [-3\text{ m}, 0\text{ m}]$, made of sand and clay, water flows from the top of the structure as

Table 3 Hydraulic properties for Test 2

	RT1 (Sand)	RT2 (Clay)
$1 - s_{\text{rn}}$	1.0	1.0
s_{rw}	0.0782	0.2262
n	2.239	1.3954
λ [m^2]	$6.3812 \cdot 10^{-12}$	$1.5461 \cdot 10^{-13}$
α [m^{-1}]	2.8	1.04
s_{lim}	0.985	0.985
ϕ	0.3658	0.4686

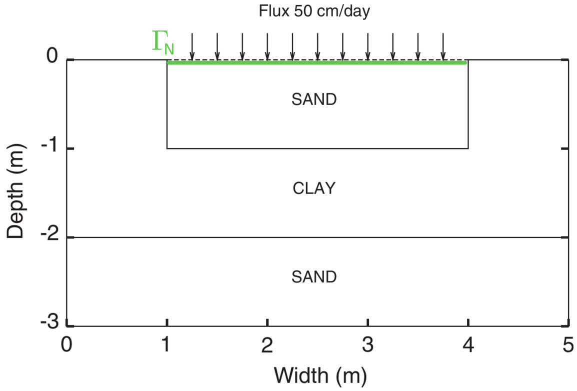


Fig. 3 Configuration of the domain for Test 2

shown in Fig. 3. The hydraulic properties of the rock types are given in Table 3. The initial pressure is set to $-47.088 \cdot 10^5$ Pa. A no-flow boundary condition is applied everywhere except on the top sand surface where the water flux rate is equal to 0.5 m/day. The simulation is performed on a mesh composed of a 100 × 60 cells during a time interval $[0, T]$ with T equal to one day. We use an initial time step $\Delta t_{ini} = 25 \cdot 10^2$ s which increases after the first time iteration up to $\Delta t_{max} = 3 \cdot \Delta t_{ini}$ using $\alpha^+_{\Delta t} = 2$.

During this simulation, the relative permeability is approximated following the strategy which has been previously described and activating the $check()$ and $update()$ procedures. The truncation method is not required here because $\mathscr{S}_{\text{vGM}}$ is C^2. Table 4 gives the average number of iterations of the nonlinear solvers used here and in [4] along with the number of time steps. Solutions obtained at the final time are shown in Fig. 4. Figure 5 shows the evolution of the average Newton's convergence rate which is slightly bigger than 1. The loss of the quadratic convergence may be due to the low regularity of the laws and to the use of the approximation $\tilde{k}_{r vGM}$.

Table 4 Performances of the nonlinear (nl) solvers for Test 2

	♯ Total nl iterations	♯ Time iterations
Our method	151	13
Method proposed in [4]	482 (inner iterations)	24

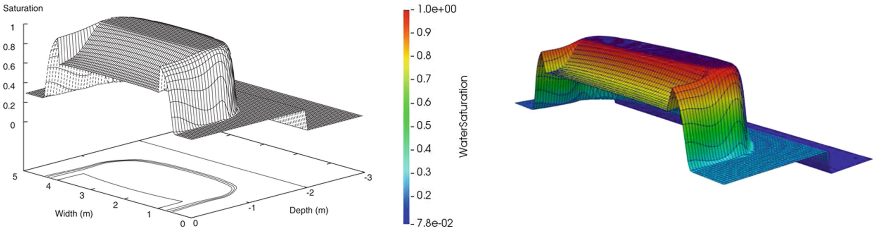
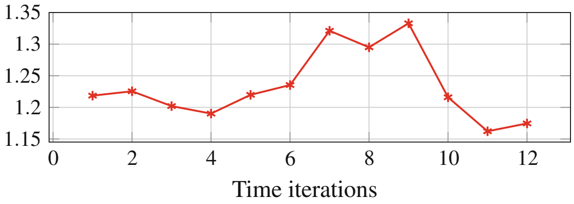


Fig. 4 At the final time for Test 2: s obtained in [4] (left) and with our solution (right)

Fig. 5 Test 2: evolution of the average Newton's convergence rate during time iterations



References

1. Alt, H.W., Luckhaus, S.: Quasilinear elliptic-parabolic differential equations. Math. Z. **183**(3), 311–341 (1983)
2. Bear, J., Bachmat, Y.: Introduction to Modeling of Transport Phenomena in Porous Media. Kluwer Academic Publishers, Dordrecht, The Netherlands (1990)
3. Brenner, K., Cancès, C.: Improving Newton's method performance by parametrization: the case of the Richards equation. SIAM J. Numer. Anal. **55**(4), 1760–1785 (2017)
4. Casulli, V., Zanolli, P.: A nested Newton-type algorithm for finite volume methods solving Richards' equation in mixed form. SIAM J. Sci. Comput. **32**(4), 2255–2273 (2010)
5. Diersch, H.J.G., Perrochet, P.: On the primary variable switching technique for simulating unsaturated-saturated flows. Adv. Water Resour. **23**(3), 271–301 (1999)
6. Eymard, R., Gallouët, T., Herbin, R.: Finite volume methods. In: Ciarlet, P.G., et al. (ed.s) In Handbook of Numerical Analysis, pp. 713–1020. North-Holland, Amsterdam (2000)
7. Forsyth, P.A., Wu, Y.S., Pruess, K.: Robust numerical methods for saturated-unsaturated flow with dry initial conditions in heterogeneous media. Adv. Water Resour. **18**, 25–38 (1995)
8. Jenny, P., Tchelepi, H.A., Lee, S.H.: Unconditionally convergent nonlinear solver for hyperbolic conservation laws with S-shaped flux functions. J. Comput. Phys. **228**(20), 7497–7512 (2009)
9. List, F., Radu, F.A.: A study on iterative methods for solving Richards equation. Comput. Geosci. **20**(2), 341–353 (2016)
10. Wang, X., Tchelepi, H.A.: Trust-region based solver for nonlinear transport in heterogeneous porous media. J. Comput. Phys. **253**, 114–137 (2013)

Acceleration of Newton's Method Using Nonlinear Jacobi Preconditioning

Konstantin Brenner

Abstract For mildly nonlinear systems, involving concave diagonal nonlinearities, semi-global monotone convergence of Newton's method is guarantied provided that the Jacobian of the system is an M-matrix. However, regardless this convergence result, the efficiency of Newton's method becomes poor for stiff nonlinearities. We propose a nonlinear preconditioning procedure inspired by the Jacobi method and resulting in a new system of equations, which can be solved by Newton's method much more efficiently. The obtained preconditioned method is shown to exhibit semi-global convergence.

Keywords Mildly nonlinear systems · Newton's method · Nonlinear preconditioning · Monotone convergence

MSC (2010) 58C15, 65H10, 65H20, 65M22

1 Introduction

Let N be a positive integer, we consider the problem of finding $u \in \left(\mathbb{R}^+\right)^N$ satisfying

$$f(u) + Au = b, \tag{1}$$

where A belongs to the set of real $N \times N$ matrices, denoted in the following by $\mathbb{M}(N)$, $b \in \left(\mathbb{R}^+\right)^N$ and the mapping f is defined by

$$f : u \mapsto (f_1(u_1), \dots f_N(u_N))^T$$

K. Brenner (✉)
Université Côte d'Azur, Inria Team Coffee, CNRS, Laboratoire J.A. Dieudonné,
Parc Valrose, Nice 06108 Nice cedex 02, France
e-mail: konstantin.brenner@univ-cotedazur.fr

R. Klöfkorn et al. (eds.), *Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples*, Springer Proceedings in Mathematics & Statistics 323,
https://doi.org/10.1007/978-3-030-43651-3_36

with f_i strictly increasing continuous functions from $\mathbb{R}^+$ to $\mathbb{R}^+$ satisfying $f_i(0) = 0$. More precisely we will assume the following:

(A_1) For $1 \le i \le N$, f_i is strictly increasing, concave and belongs to C^1 on $(0, +\infty)$.
(A_2) For any $u > 0$ the matrix $f'(u) + A$ is an M-matrix in the sense of the definition below.
(A_3) The matrix A has zero diagonal elements.

Definition 1 We say that A is an M-matrix if A is invertible, $A^{-1} \ge 0$, and $a_{i,j} \le 0$ for $i, j = 1, \dots, N$ with $i \ne j$.

We remark that the derivatives of f_i are potentially unbounded at the origin.

The system (1) can be found in numerical modeling of flow and transport processes. In particular it arises from the discretization of the nonlinear evolutionary PDEs of the form

$$\partial_t \beta(u) + \operatorname{div}(\mathbf{v}u - \lambda \nabla u) = \gamma(u), \tag{2}$$

where $\mathbf{v}$ is some given velocity field. Applying the backward Euler scheme and some space discretization method to (2) one typically get the discrete problem of the form

$$\frac{\beta(u_h^n) - \beta(u_h^{n-1})}{\Delta t} + M^{-1} S u_h^n = \gamma(u_h^n) + \sigma_h^n, \tag{3}$$

where $u_h^n, u_h^{n-1} \in \mathbb{R}^N$ are the vectors of the discrete unknowns associated with two sequential time steps, while M and S are respectively the mass and the stiffness matrices, and the vector σ_h^n contains boundary data.

To fix the ideas let's assume that the Dirichlet boundary conditions are imposed. Several space discretization methods provide (possibly under some geometrical condition on the mesh) that the matrix $M^{-1}S$ is an M-matrix. In the presence of diffusion (that is $\lambda > 0$), the examples of such *monotone discretization* schemes is the standard finite volume method with two-point flux approximation and P_1 finite element method with mass lumping under the Delaunay condition on the underlying mesh (see [3]). Let us mention that the monotone discretizations are not only beneficial to the nonlinear solver (as it is going to be discussed in this paper), but also allow to preserve the local maximum principle on the discrete level, thus avoiding any spurious oscillations of the discrete solution. Let D denote the diagonal of $M^{-1}S$ and let $A = \Delta t \left(M^{-1}S - D\right)$. Setting

$$f(u) = \beta(u) + \Delta t \,(Du - \gamma(u))$$

the system (3) can be written as (1).

Given the assumption (A_1) on the mapping f, and thus on the nonlinearities $\beta(u)$ and $\gamma(u)$, several physical models are relevant. Such models are for example the porous media equation [6], models of transport in porous media with adsorption (using e.g. the Freundlich isotherm [1]), the Richards' equation [2, 5] or the Dupuit-Forchheimer equation [1] (provided that convection is discretized using an explicit scheme). Let us further remark that the analysis and the algorithms presented in this paper can be extended to the Hele-Shaw or Stefan like problems where $\beta(u)$ is no longer a function, but rather a monotone graph of the form $f(u) = \zeta H(u) + \tilde{f}$, where $\tilde{f}$ is a function satisfying the assumption (A_1), ζ is a positive real number and $H(u)$ denotes the multivalued Heaviside graph. In [2] this type of nonlinearity has been addressed trough the parametrization of f, that is a couple of the functions $\tau \to (\overline{u}(\tau), \overline{v}(\tau))$ with $\overline{v}(\tau) \in f(\overline{u}(\tau))$ for all τ. The problem has been then rewritten in terms of the new variable τ.

Due to its quadratic convergence, Newton's method is a very popular tool that can be used to solve the systems (1) numerically; moreover under assumptions (A_1) and (A_2) one can show that Newton's method converges monotonically toward any strictly positive solution $u_\star$ as soon as the initial guess u_0 satisfies $0 < u_0 \leq u_\star$. This *semi-global* convergence result is based on the concavity of the underlying functional and the non-negativity of the inverse of it's Jacobian; it is in fact a straightforward adaptation of the convergence results from [4] (see also Proposition 1 below) to the concave setting.

Despite an available convergence result, the numerical evidences presented in [2] suggest that the efficiency of Newton's method applied to (1) can be very poor especially for stiff problems with $f'(0) = +\infty$. To give an example let $\gamma(u) = 0$ and $\beta(u) = u^{\frac{1}{m}}$, $m \geq 1$ (this choice corresponds to the porous media equation [6]). It is demonstrated in the numerical Sect. 3 that the convergence of Newton's method is slow; moreover the number of Newton's iterations required to solve the system grows with m. The numerical experiment also demonstrates that the efficiency of Newton's method can be greatly improved by a simple change of the variable $v = \beta(u)$. Let us note that for Richards-like parabolic-elliptic problems with $\beta'(u) = 0$ for $u \geq u_s > 0$ the similar change-of-variable trick can be performed using the variable switching technique as suggested in [2]. Compared to the initial formulation of (1) the drawback of the change-of-variable approaches is that the concavity of the problem is lost, and therefore the monotone convergence is no longer guarantied.

In this article we reformulate the system (1) in a way that accelerates convergence of Newton's method while preserving concavity of the problem. More precisely we replace the system (1) by a different one having the same solution set but is easier to solve using Newton's method. Since the modified system is similar to the one obtained in Jacobi method, we refer to our approach as to Jacobi preconditioned Newton's method.

The mapping f is diagonal, strictly increasing and continuous and therefore admits an inverse denoted $g = f^{-1}$. We consider the following left-preconditioned and right-preconditioned problems

$$F_l(u) := u - g(b - Au) = 0 \tag{4}$$

or

$$F_r(u) := u + Ag(u) - b = 0. \tag{5}$$

Under the assumption (A_1) the function g is increasing and convex, and therefore $F_\star(u)$, $\star = l, r$ remains concave; moreover the derivative of g is finite for all $u \in (\mathbb{R}^+)^N$ and it can be shown that $F'_\star(u)$ exists and is an M-matrix for all $u \in (\mathbb{R}^+)^N$. This implies monotone convergence of Newton's method applied to (4) and (5) for any initial guess u_0 satisfying $F_\star(u_0) \leq 0$. The numerical experiment shows (see Sect. 3) that performance of the preconditioned methods turns out to be superior compare to the original formulation of (1), or alternatively to the change-of-variable approaches.

The reminder of the article is organized as follows. In Sect. 2, starting with convergence result from [4], we prove monotone convergence of Newton's method applied to the problem (1) in its original formulation and applied to the preconditioned problems (4) and (5). Section 3 is deduced to the numerical experiment.

2 Main Results

Let us first present the adaptation of the convergence result 13.3.4 from [4] to the case of concave mappings.

Theorem 1 (Convergence of Newton's method) *Let F be a continuous G-differentiable concave mapping from $(\mathbb{R}^+)^N$ to $\mathbb{R}^N$ and let $F'(u)$ be an M-matrix for all $u \in (\mathbb{R}^+)^N$. Assume in addition that there exist $u_\star \in (\mathbb{R}^+)^N$ satisfying $F(u_\star) = 0$ and $u_0 \in (\mathbb{R}^+)^N$ such that $F(u_0) \leq 0$. Then the sequence*

$$u_{n+1} = u_n - F'(u_n)^{-1}F(u_n), \qquad n \geq 0 \tag{6}$$

is well defined, satisfies

$$u_n \leq u_{n+1} \leq u_\star, \qquad F(u_n) \leq 0$$

and is convergent. If in addition there exists an invertible $P \in \mathbb{M}(N)$ such that $F'(u_n)^{-1} \geq P \geq 0$ for all $n \geq 0$, then the sequence u_n converges to $u_\star$.

Let us denote $F_u(u) = f(u) + Au - b$, based on the assumptions (A_1) and (A_2), it can be shown that the solution of (1) exists and is unique; in addition under the same assumptions it follows from Theorem 1 that Newton's method applied to (1) converges monotonically provided that $u_\star > 0$ and $F_u(u_0) \leq 0$. More precisely the following proposition holds.

Proposition 1 (Convergence of the original formulation) *Assume that $b > 0$, then there exists the unique solution $u_\star$ to (1) satisfying $u_\star > 0$; moreover there exists u_0 such that $F_u(u_0) \leq 0$ and Newton's iterates (6) are well defined and monotonically converge to $u_\star$.* □

We remark that if $f'(0) = +\infty$ the assumption $b > 0$ can not be avoided, therefore Newton's method can not be applied to $F_u(u) = 0$ unless the solution is strictly positive. In contrast the preconditioned methods can be applied without any additional restrictions on f' or on the sign of the solution. Convergence of the preconditioned methods is summarized by the following proposition, which relies on the assumption (A_3) ensuring the concavity of $F_\star$ and the M-matrix property of $F'_\star$, $\star = l, r$.

Proposition 2 (Convergence of the preconditioned methods) *The mappings F_l and F_r satisfy the assumptions of Theorem 1 with $u_0 = 0$; moreover for all $u \in (\mathbb{R}^+)^N$ the matrix $F'_\star(u)$, $\star = l, r$ is such that $F'_\star(u) \le I \le F'_\star(u)^{-1}$.* □

3 Numerical Experiment

Let us consider the porous medium equation (see [6])

$$\partial_t \beta(u) - \partial^2_{xx} u = 0 \tag{7}$$

on $(0, 1) \times (0, T)$. The nonlinearity in the accumulation term is given by $\beta(u) = u^{1/m}$ with $m > 1$. We consider the Neumann boundary conditions

$$\partial_x u(0, t) = -q, \quad \partial_x u(1, t) = 0, \quad \text{for all } t \in (0, T)$$

with $q > 0$, and the constant initial condition $u(x, 0) = u_0 > 0$. The value of u_0 is going to be chosen close to zero leading to "an almost traveling wave solution". For $m = 10$, $q = 10^4$, $T = 1.2\ 10^{-2}$ and $N_T = 100$ the approximate profile of $\beta(u)$ at different time steps is exhibited at the right side of Fig. 2.

Equation (7) is discretized using the standard implicit in time finite volume method. Let the positive integers N and N_T denote the number of cells and the total number of time steps, let $h = \dfrac{1}{N}$ be the cell size and $\Delta t = \frac{T}{N_T}$ be the size of the time step. For all cells i and time steps n the discretized version of (7) reads

$$\beta(u_i^n) + \frac{\Delta t}{h^2} \sum_{j \in \mathcal{N}_i} (u_i^n - u_j^n) = \beta(u_i^{n-1}) + \frac{\Delta t}{h} q\ \delta_{i,1}, \tag{8}$$

where $\delta_{i,1}$ stands for the Kronecker symbol and where $\mathcal{N}_i$ denotes the set of the neighbors of the cell i. Let L denote the tridiagonal matrix associated to the discretization of the diffusion operator and D be it's diagonal. We denote by b_n the right-hand-side of (8). The system (8) results in the following problem, which has to be solved for each time step

$$(\beta(u) + Du) + (L - D)\, u = b_n. \tag{9}$$

It is easy to show that $f(u) = \beta(u) + Du$ and $A = L - D$ satisfy the assumptions (A_1)–(A_3).

The objective of the numerical experiment is to evaluate the efficiency of Newton's method (NM) applied to left and right-preconditioned problems

$$F_l^n(u) := u - g\,(b_n - Au) = 0 \tag{10}$$

and

$$F_r^n(u) := u + Ag(u) - b_n = 0. \tag{11}$$

Those preconditioned methods are compared, in terms of the performance, with three more standard approaches specified below.
$u-$**formulation**: NM applied to (9) in the original form

$$F_u^n(u) := \beta(u) + Lu - b_n = 0 \tag{12}$$

In view of Proposition 2 this method is monotonically convergent provided that the initial guess satisfy $F(u_0) \le 0$.
$v-$**formulation**: The problem (9) is reformulated with respect to the variable v with $u = \beta^{-1}(v)$ and NM is applied to

$$F_v^n(v) := v + L\beta^{-1}(v) - b_n = 0 \tag{13}$$

$\tau-$**formulation**: Following [2] we introduce the function pair $\tau \to (\overline{u}(\tau), \overline{v}(\tau))$ such that for all τ it holds $\overline{v}(\tau) = \beta(\overline{u}(\tau))$ and $\max\big(\overline{u}'(\tau), \overline{v}'(\tau)\big) = 1$. Then NM is applied to

$$F_\tau^n(\tau) := \overline{v}(\tau) + L\overline{u}(\tau) - b_n = 0. \tag{14}$$

At each time step n and for each of the formulations (10)–(14) the sequence of the approximate solutions $\left(\xi_k^n\right)_k$ (where ξ denotes an appropriate primary variable) is computed using Newton's method until the stopping criterion $\|F_\star^n(\xi_k^n)\|_\infty < \varepsilon$ is satisfied for some small predefined tolerance ε. As the initial guess we use the value of the variable obtained at the previous time step (this value will differ between the formulations). This choice of the initial guess is motivated by the following observation.

Remark 1 Under the given initial and boundary conditions the solution of (7) satisfies $\partial_t u \ge 0$. This property is reproduced by the discrete solution u^n resulting from $u-$formulation and the preconditioned methods. For $\star = u, r, l$, let u^n denote an approximate solution of $F_\star^n(u) = 0$, then one can show that $F_\star^n(u^{n-1}) \le 0$, and therefore $u_0^n = u^{n-1}$ provides the appropriate choice of the initial guess.

In the following we present the results of the numerical experiment. The test case is configurated as follows: in order to allow for the use of $u-$formulation we chose strictly positive initial condition $\beta(u_0) = 10^{-10}$, we set $q = 10^4$, $T = 1.2\ 10^{-2}$, $N_T = 100$ and we let the parameter m take values in the set $\{4, 8, 16, 32\}$. For a given value of m, the tolerance ε and a specific solution method $\star$, we denote by $\left(u^{n,\star}_{m,\varepsilon}\right)_{n\in\{1,\dots,N_T\}} \in \mathbb{R}^N$ the approximate solution of (9).

The methodology of the study is similar to [2], that is for each value of m we compute, using $\tau-$formulation and the tolerance $\varepsilon_{ref} = 10^{-10}$, the reference solution denoted by $\left(u^n_{m,ref}\right)_{n\in\{1,\dots,N_T\}}$. Then, for each solution method (10)–(14) and for the tolerance values of $\varepsilon \in \{10^{-1}, 10^{-2}, 10^{-4}, 10^{-6}, 10^{-8}\}$, we perform the computations measuring the total number of Newton's iteration, required CPU time and the relative deviation from the reference solution measured in terms of the conservative variable $\beta(u^\star_{m,\varepsilon})$

$$err^\star_{m,\varepsilon} = \frac{\|\beta(u^{n,\star}_{m,\varepsilon}) - \beta(u^n_{m,ref})\|_{L^\infty(0,T;L^1(0,1))}}{\|\beta(u^n_{m,ref})\|_{L^\infty(0,T;L^1(0,1))}}.$$

Performance comparison. The first set of tests is performed using the fixed mesh size parameter $N = 100$. In accordance with the results reported in [2], Fig. 1 witness the qualitative differences in the performance of u, v and τ-formulations. Compared to the original u-formulation, the formulation using v as the primary variable is few time faster, it also performs slightly better then τ-formulation for the moderate values of m. However, in contrast with τ-formulation, none of the formulations u or v is robust with respect to the variation of m. Finally, Fig. 2 shows a relatively similar behavior of τ-formulation and the preconditioned methods, with the latter ones requiring a slightly fewer number of iterations.

Computational overhead due to local problem solution. It can be observed on Figs. 1 and 2 that preconditioned Newton's methods require less iterations then the other formulations. However, each iteration of the preconditioned method requires to evaluate the function g, and therefore to solve the set of the scalar nonlinear equations. Those computations, performed again using Newton's method, result in a certain computational overhead which has to be accounted for. To access the overall

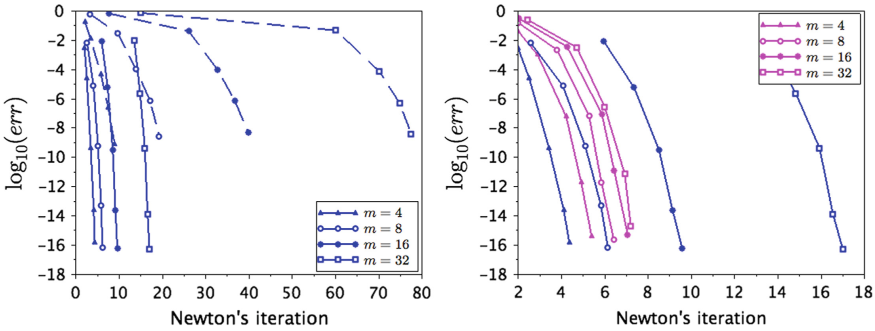


Fig. 1 Relative error $err^\star_{m,\varepsilon}$ as the function of the average number of Newton's iterations per time step. Left: for v-formulation (solid blue) and u-formulation (dashed blue). Right: for v-formulation (blue) and τ-formulation (magenta)

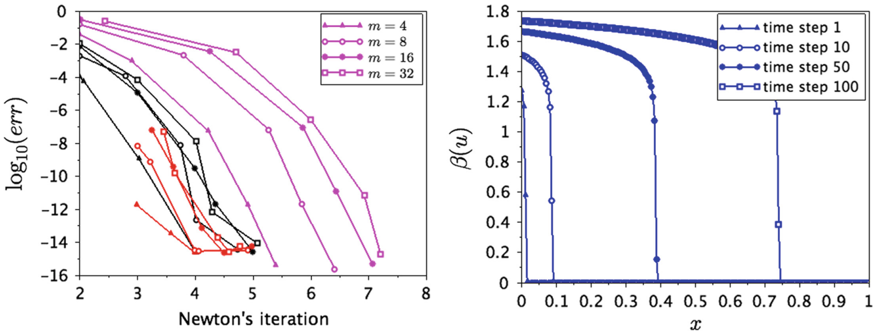


Fig. 2 Left: relative error $err^{\star}_{m,\varepsilon}$ as the function of the average number of Newton's iterations per time step for τ-formulation (magenta), left-preconditioned (black) and right-preconditioned (red) Newton's method (magenta). Right: Approximate solution at different time steps

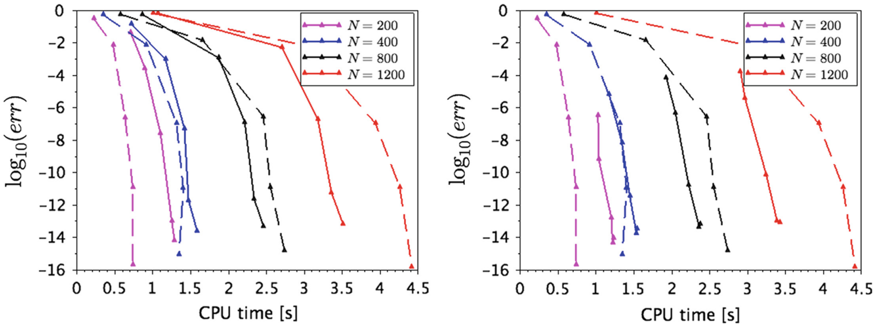


Fig. 3 Relative error $err^{v,\star}_{m,\varepsilon}$ as the function of CPU time for different grid sizes. Left: left-preconditioned NM (solid lines) and τ-formulation (dashed lines). Right: right-preconditioned NM (solid lines) and τ-formulation (dashed lines)

computational effort required by the preconditioned methods we present the analysis in terms of the CPU time. Figure 3 shows, for different values of the mesh size parameter $N \in \{200, 400, 800, 1200\}$, the comparison of the left (respectively right) preconditioned NM with the method based on τ-formulation. In can be observed that for the small problems ($N \lesssim 400$) τ-formulation outperforms the preconditioned NM due to the computational overhead related to the latter ones. In turn, for larger problems the preconditioned methods became advantages due to a smaller number of the linear problem solves.

References

1. Bear, J., Verruijt, A.: Modeling groundwater flow and pollution. Reidel (1987)
2. Brenner, K., Cancès, C.: Improving newton's method performance by parametrization: the case of Richards equation. SIAM J. Numer. Anal. (2017)
3. Eymard, R., Gallouët, T., Herbin, R.: Finite volume methods. In: Ciarlet, P.G. (ed.) et al. Handbook of Numerical Analysis, pp. 713–1020. North-Holland, Amsterdam (2000)
4. Ortega, J.M., Rheinboldt, W.C.: Iterative Solutions of Nonlinear Equations in Several Variables. Academic (1970)
5. Van Duijn, C.J., Peletier, L.A.: Nonstationary filtration in partially saturated porous media. Arch. Rat. Mech. Anal. **78**(2), 173–198 (1982)
6. Vázquez, J.L.: The Porous Medium Equation—Mathematical Theory. The Clarendon Press, Oxford University Press (2007)

A Finite Volume Method for a Convection-Diffusion Equation Involving a Joule Term

Caterina Calgaro and Emmanuel Creusé

Abstract This work is devoted to a Finite Volume method to approximate the solution of a convection-diffusion equation involving a Joule term. We propose a way to discretize this so-called "Joule effect" term in a consistent way with the non linear diffusion one, in order to ensure some maximum principle properties on the solution. We then investigate the numerical behavior of the scheme on two original benchmarks.

Keywords Finite volume scheme · Joule term · Maximum principle

MSC (2010) 65M08 · 65M60

1 Introduction

In this work, we are interested in a convection-diffusion equation involving a Joule term, given by:

$$\partial_t u + \nabla \cdot (u\, \mathbf{v}) + 2\lambda\, |\nabla u|^2 - \lambda \nabla \cdot (u\, \nabla u) = f \quad \text{in } \Omega \times]0, T[, \tag{1}$$

$$u(\mathbf{x}, 0) = u_0(\mathbf{x}) \quad \text{in } \Omega, \tag{2}$$

where Ω is a polygonal open bounded subset of $\mathbb{R}^2$, $T \in \mathbb{R}_*^+$, $\lambda \in \mathbb{R}_*^+$, $\mathbf{v}$ is a divergence-free velocity field, and f the right-hand-side. The system (1)–(2) is completed with boundary conditions, given by:

C. Calgaro (✉)
Inria-Laboratoire Paul Painlevé, Univ. Lille, CNRS, UMR 8524, 59000 Lille, France
e-mail: caterina.calgaro@univ-lille.fr

E. Creusé
Univ. Poly. Hauts-de-France, EA 4015, LAMAV, FR CNRS 2956,
59313 Valenciennes, France
e-mail: emmanuel.creuse@uphf.fr

R. Klöfkorn et al. (eds.), *Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples*, Springer Proceedings in Mathematics & Statistics 323,
https://doi.org/10.1007/978-3-030-43651-3_37

$$
\begin{aligned}
u(\mathbf{x},t) &= u_D(\mathbf{x},t) && \forall\, \mathbf{x} \in \Gamma_D,\ \forall t \in]0,T[,\\
\nabla u(\mathbf{x},t)\cdot \mathbf{n} &= 0 && \forall\, \mathbf{x} \in \Gamma_N,\ \forall t \in]0,T[,
\end{aligned}
\qquad (3)
$$

where u_D is a given function corresponding to non homogeneous Dirichlet boundary conditions, $\overline{\Gamma_D} \cup \overline{\Gamma_N} = \overline{\Gamma} = \overline{\partial \Omega}$ and $\Gamma_D \cap \Gamma_N = \emptyset$.

System (1)–(3) (with $f = 0$) can be derived in the context of low-Mach modeling. Taking into account the compressible Navier-Stokes system where an asymptotic development of the pressure with respect to the Mach number is done, we start by considering the mass conservation equation

$$\partial_t \rho + \nabla \cdot (\rho\, \mathbf{V}) = 0,$$

where $\rho(\mathbf{x},t)$ is the density and $\mathbf{V}(\mathbf{x},t)$ the velocity field of the fluid. In the case of the ideal gas law $P_0 = R\,\rho\, u$, where $u(\mathbf{x},t)$, $P_0 > 0$ and $R > 0$ stand respectively for the temperature, the constant thermodynamic pressure and the ideal gas constant, a solenoidal velocity field $\mathbf{v}(\mathbf{x},t)$ can be introduced. It is shown in [3] that the change of variable $\mathbf{v} = \mathbf{V} - \lambda\, \nabla u$ leads to equation (1), where $\lambda > 0$ is a fixed constant which depends on the constant heat conductivity $k > 0$ in the nonstandard constraint

$$P_0 \nabla \cdot \mathbf{V} = R\, \nabla \cdot (k \nabla u)$$

introduced in the low-Mach model. In [3] a particular dynamic viscosity is also introduced, defined by $\mu(u) = -\lambda \ln u$, in order to remove the $\mathcal{O}(\lambda^2)$ terms in the momentum equation. With this choice, $\mu(u)$ is strictly positive if and only if $u \in (0,1)$. However, in this work we assume only that there exist two real numbers m and M such that $0 < m \le u_0(\mathbf{x}) \le M < +\infty$ a.e. $\mathbf{x} \in \Omega$.

From the theoretical point of view, several results have been obtained for the system (1)–(3). For instance, the local-in-time existence of strong solutions has been established in [3] in the framework of a coupling with the Navier-Stokes system. In particular, a maximum-principle has been derived (see [6], Theorem 5.1). This formulation is also related to others obtained in the context of the so-called ghost effect system, where a thermal stress term is added to the right-hand-side of the momentum equation, for which some results on the existence and uniqueness of solutions are available [9, 10].

From the numerical point of view, in the context of Finite Volume schemes, an important question to be addressed consists in the way to discretize the Joule term $|\nabla u|^2$ arising in (1) in each control volume, in a consistent way with the non linear diffusion one. It has to be done in order to ensure some properties on the numerical solution, such as some maximum principles which hold at the continuous level. Several possibilities have already been investigated in the context of the electrical conductivity (see for example [1, 4, 5]) but, to our knowledge, never for the model (1)–(3).

In this work, we present a finite volume scheme for the discretization of (1)–(3) which has been initially introduced in [2]. The aim of the present contribution is to give some results in the case $f \neq 0$, and to investigate the efficiency of the derived scheme (as well as a variant one) on two original benchmarks. More precisely, we first illustrate our theoretical results on a discontinuous solution submitted to a solenoidal convective velocity field. Then, we consider a regular analytical solution for which the right-hand-side is positive, in order to investigate the lower bound preserving property of the numerical solution as well as the convergence process.

2 Finite Volume Scheme

2.1 Notations

As usual, the discretization in space is based on a triangulation $\mathcal{T}$ of the domain $\Omega \subset \mathbb{R}^2$, a family $\mathcal{E}$ of edges and a set $\mathcal{P} = (\mathbf{x}_K)_{K\in\mathcal{T}}$ of points of Ω defining an admissible mesh in the sense of Definition 3.1 in [7]. We recall that the admissibility of $\mathcal{T}$ implies that the straight line between two neighboring centers of cells $\mathbf{x}_K$ and $\mathbf{x}_L$ is orthogonal to the edge $\sigma \in \mathcal{E}$ such that $\overline{\sigma} = \overline{K} \cap \overline{L}$ (and which is noted $\sigma = K|L$) in a point $\mathbf{x}_\sigma$.

The set of interior (resp. boundary) edges is denoted by $\mathcal{E}^{\text{int}} = \{\sigma \in \mathcal{E}\,;\ \sigma \not\subset \Gamma\}$ (resp. $\mathcal{E}^{\text{ext}} = \{\sigma \in \mathcal{E}\,;\ \sigma \subset \Gamma\}$). Among the outer edges, there are $\mathcal{E}^N = \{\sigma \in \mathcal{E}\,;\ \sigma \subset \Gamma_N\}$ and $\mathcal{E}^D = \{\sigma \in \mathcal{E}\,;\ \sigma \subset \Gamma_D\}$. For all $K \in \mathcal{T}$, we denote by $\mathcal{E}_K = \left\{\sigma \in \mathcal{E}\,;\ \sigma \subset \overline{K}\right\}$ the edges of K, $\mathcal{E}_K^{\text{int}} = \mathcal{E}^{\text{int}} \cap \mathcal{E}_K$, $\mathcal{E}_K^{\text{ext}} = \mathcal{E}^{\text{ext}} \cap \mathcal{E}_K$, $\mathcal{E}_K^N = \mathcal{E}_K^N \cap \mathcal{E}_K$ and $\mathcal{E}_K^D = \mathcal{E}_K^D \cap \mathcal{E}_K$.

The measure of $K \in \mathcal{T}$ is denoted by m_K and the length of σ by m_σ. For $\sigma \in \mathcal{E}^{\text{int}}$ such that $\sigma = K|L$, d_σ denotes the distance between $\mathbf{x}_K$ and $\mathbf{x}_L$ and $d_{K,\sigma}$ the distance between $\mathbf{x}_K$ and σ. For $\sigma \in \mathcal{E}_K^{\text{ext}}$, we note d_σ the distance between $\mathbf{x}_K$ and σ. For $\sigma \in \mathcal{E}$, the transmissibility coefficient is given by $\tau_\sigma = \dfrac{m_\sigma}{d_\sigma}$. Finally, for $\sigma \in \mathcal{E}_K$, we denote by $\mathbf{n}_{K,\sigma}$ the exterior unit normal vector to σ. The size of the mesh is given by:

$$h = \max_{K\in\mathcal{T}} \text{diam}(K).$$

We define a partition of the time interval $(0, T)$ such that $0 = t^0 < \cdots < t^n < \cdots < t^N = T$ $(N \in \mathbb{N}^*)$, and we denote $\Delta t_n = t^{n+1} - t^n$ for $0 \leq n \leq N-1$. Here, u_K^n denotes the value of the numerical solution $u_h^n = (u_K^n)_{K\in\mathcal{T}}$ in K and in the time interval $[t^n, t^{n+1}]$.

2.2 The Finite Volume Scheme 1

The Finite Volume scheme for the discretization of (1)–(3) is given by integrating (1) on a control volume K to obtain:

$$m_K \frac{u_K^{n+1} - u_K^n}{\Delta t_n} + \sum_{\sigma \in \mathcal{E}_K} v_{K,\sigma}^n \, u{K}_{\sigma,+}^{n+1} + 2\lambda \, m_K \, \mathcal{J}_K(u_h^{n+1}) + \lambda \sum_{\sigma \in \mathcal{E}_K} F_{K,\sigma}^{n+1} = m_K \, f_K^{n+1} \quad \forall K \in \mathcal{T}, \tag{4}$$

with

$$v_{K,\sigma}^n = \int_\sigma \mathbf{v}(\mathbf{x}, t^n) \cdot \mathbf{n}_{K,\sigma} \, \mathrm{d}\gamma(\mathbf{x}) \quad \text{and} \quad f_K^{n+1} = \frac{1}{m_K} \int_K f(\mathbf{x}, t^{n+1}) \, \mathrm{d}\mathbf{x}.$$

Here, $u{K}_{\sigma,+}^{n+1}$ is defined for $\sigma \in \mathcal{E}_K$ by:

$$u{K}_{\sigma,+}^{n+1} = \begin{cases} u_K^{n+1} & \text{if } v_{K,\sigma}^n \geq 0, \\ u_{K,\sigma}^{n+1} & \text{otherwise,} \end{cases}$$

with

$$u_{K,\sigma}^{n+1} = \begin{cases} u_L^{n+1} & \text{for } \sigma \in \mathcal{E}_K^{\text{int}} \text{ such that } \sigma = K|L, \\ u_D^{n+1}(\mathbf{x}_\sigma) & \text{for } \sigma \in \mathcal{E}_K^D, \\ u_K^{n+1} & \text{for } \sigma \in \mathcal{E}_K^N. \end{cases}$$

The numerical flux $F_{K,\sigma}^{n+1}$ is an approximation of the exact flux of the non linear diffusion term through the edge σ, given classically by:

$$F_{K,\sigma}^{n+1} = \frac{\tau_\sigma}{2} \left((u_K^{n+1})^2 - (u_{K,\sigma}^{n+1})^2 \right). \tag{5}$$

Concerning the Joule term, $\mathcal{J}_K(u_h^{n+1})$ is an approximation of $\frac{1}{m_K} \int_K |\nabla u(\mathbf{x}, t^{n+1})|^2 \mathrm{d}\mathbf{x}$. A first idea may be to consider a scheme also centered for this term, in order to increase the accuracy of the approximation, by following the piecewise discrete gradient introduced in [8], for instance. Nevertheless, we have shown in [2] that with this choice, a discrete maximum principle is only verified under very restrictive conditions on the initial data. A more effective approach is to consider an upwind discretization of the Joule term defined by:

$$\mathcal{J}_K(u_h^{n+1}) = \frac{1}{m_K} \sum_{\sigma \in \mathcal{E}_K} \tau_\sigma \left((u_K^{n+1} - u_{K,\sigma}^{n+1})^+ \right)^2, \tag{6}$$

where we use the notation $a^+ = \max(0, a)$. Let us note that the discretization is implicit in time, in order to allow some maximum principle results without any restriction on the time step. We give now the main result:

Theorem 1 *We assume that*

$$0 < m \leq u_K^0 \leq M \quad \forall K \in \mathcal{T} \text{ and } 0 < m \leq u_D^n(\boldsymbol{x}_\sigma) \leq M, \quad \forall \sigma \in \mathcal{E}^D, \, \forall n = 1, \cdots, N.$$

We suppose moreover that:

$$f_K^n \geq 0 \quad \forall K \in \mathcal{T},\ \forall n = 1, \dots, N.$$

Then the scheme (4) admits at least one solution that satisfies:

$$m \leq u_K^n \leq M + T\,||f||_{L^\infty(\Omega\times[0,T])}, \quad \forall K \in \mathcal{T},\ \forall n = 1, \dots, N. \tag{7}$$

Proof The proof consists in an extension of the proof of Theorem 3.1 in [2] established in the case $f \equiv 0$, noticing in particular that $((u_K^{n+1} - u_{K,\sigma}^{n+1})^+)^2 = 0$ if $u_K^{n+1} \leq u_{K,\sigma}^{n+1}$. First, we prove that for any solution of (4), estimation (7) holds. Then, we use a topological degree argument to establish the existence of the solution.

Remark 1 In the case $f_K^n \leq 0$, a similar result can also be obtained. Concerning the upper bound, it can easily be proved that any solution of (4) satisfies $u_K^n \leq M$, $\forall K \in \mathcal{T}$, $\forall n = 1, \dots, N$. Concerning the lower bound, the time T has to be chosen sufficiently small to ensure that $u_K^n > 0$, $\forall K \in \mathcal{T}$, $\forall n = 1, \dots, N$ and to be able to prove the existence of the solution.

2.3 A Variant: Scheme 2

Here, we propose a variant of the previous numerical scheme. First, we would like to suggest a centered treatment of the Joule term instead of (6). It is worth noticing that $|\nabla u|^2 = \nabla \cdot (u\,\nabla u) - u\,\Delta u$, then we propose the following definition:

$$\mathcal{J}_K(u_h^{n+1}) = \frac{1}{m_K}\sum_{\sigma\in\mathcal{E}_K} \tau_\sigma\, \overline{u}_\sigma^{n+1}\left(u_{K,\sigma}^{n+1} - u_K^{n+1}\right) - \frac{1}{m_K} u_K^{n+1} \sum_{\sigma\in\mathcal{E}_K} \tau_\sigma\, \left(u_{K,\sigma}^{n+1} - u_K^{n+1}\right),$$

where $\overline{u}_\sigma^{n+1}$ is an approximation of u^{n+1} at $\mathbf{x}_\sigma$ defined by: $\overline{u}_\sigma^{n+1} = \dfrac{u_K^{n+1} + u_{K,\sigma}^{n+1}}{2}$. Finally we obtain:

$$\mathcal{J}_K(u_h^{n+1}) = \frac{1}{2\,m_K}\sum_{\sigma\in\mathcal{E}_K} \tau_\sigma (u_{K,\sigma}^{n+1} - u_K^{n+1})^2.$$

With this choice, if we hope to achieve a maximum principle without restrictions on the data, we need to reach a balance between the Joule term and the diffusive one. Thus, instead of (5), we propose to define the numerical flux for the diffusion term through the edge σ by:

$$F_{K,\sigma}^{n+1} = \tau_\sigma\, u_\sigma^{n+1}(u_K^{n+1} - u_{K,\sigma}^{n+1}),$$

where u_σ^{n+1} is another approximation of u^{n+1} at $\mathbf{x}_\sigma$ defined this time by: $u_\sigma^{n+1} = \max(u_K^{n+1}, u_{K,\sigma}^{n+1})$. Provided this balance between the diffusion term and the Joule one, Theorem 1 occurs.

3 Benchmarks

3.1 Maximum Principle: Case $f = 0$

This first benchmark consists in solving (1)–(3) with $\Omega = [0; 1]^2$ for $0 \leq t \leq T = 0.1$, with $f \equiv 0$, $\lambda = 2$, $\Gamma_D = \Gamma$, $\Gamma_N = \emptyset$ and $\forall t \in [0; T]$:

$$u_D(\mathbf{x}, t) = \begin{cases} 1 & \text{if } \mathbf{x} \in \Gamma \backslash \Gamma_H, \\ M & \text{if } \mathbf{x} \in \Gamma_H, \quad M > 1, \end{cases}$$

where $\Gamma_H = \{1\} \times (0,3; 0,7)$. Moreover, the given velocity field $\mathbf{v}(\mathbf{x}, t)$ is given by:

$$\mathbf{v}(\mathbf{x}, t) = 5\sqrt{t}\, e^{-25\|\mathbf{x}-\mathbf{c}\|^2} \begin{pmatrix} -y + c_2 \\ x - c_1 \end{pmatrix}, \quad \forall \mathbf{x} = (x, y)^T \in \Omega, \quad \forall t \in [0, T],$$

with $\mathbf{c} = (c_1, c_2)^T = (0.5, 0.5)^T$. Since $f \equiv 0$, Theorem 1 can be applied, so that the numerical solution should be bounded between 1 and M, what we aim to illustrate here. The simulations are performed on a triangulation $\mathcal{T}$ corresponding to $h = 3.62 \cdot 10^{-2}$. Since both schemes are implicit in time, a Newton solver is implemented associated with the adaptive time step Δt_n to compute u_h^{n+1} from u_h^n from (4). Iterations are performed until the accuracy on the residual in l^∞-norm is less than 10^{-10}. If it is not the case after 15 iterations, the time step is divided by 2 and the resolution is done again. Conversely, the solver tries to multiply by a factor 2 the time step periodically, which in any case remains bounded by h in order to preserve the accuracy in time. Several values of M are considered from $M = 2$ to $M = 50$. Results are displayed in Table 1. On the one hand, it is observed that the numerical solution is bounded between 1 and M whatever the value of M chosen, as theoretically expected. On the other hand, we investigate the time steps Δt_{min} and Δt_{max}, corresponding respectively to the smallest and the largest value of Δt_n used in $[0, T]$. First, it can be seen that the higher M is, the smaller Δt_{min} and Δt_{max} have to be in order to ensure the convergence process. Then, we remark that Scheme 1 leads to a value of Δt_{max} roughly ten times larger than the one used for Scheme 2, and consequently to a faster computation of the solution on the interval $[0, T]$.

Table 1 Verification of the maximum principle according to M. ✓ : it is satisfied

$M =$	Scheme 1			Scheme 2		
		Δt_{min}	Δt_{max}		Δt_{min}	Δt_{max}
2	✓	6.25×10^{-5}	5.00×10^{-4}	✓	7.81×10^{-6}	1.56×10^{-5}
10	✓	3.91×10^{-6}	3.13×10^{-5}	✓	1.95×10^{-6}	3.91×10^{-6}
20	✓	1.95×10^{-6}	1.56×10^{-5}	✓	6.25×10^{-7}	2.50×10^{-6}
50	✓	4.88×10^{-7}	7.81×10^{-6}	✓	2.50×10^{-7}	5.00×10^{-7}

3.2 Convergence Rate and Maximum Principle: Case $f \neq 0$

Now, we want to investigate numerically the convergence rate of the schemes on a regular solution and to illustrate Theorem 1 in one case corresponding to $f \neq 0$. To do that, we consider the exact solution u_{ex} given by:

$$u_{ex}(\mathbf{x}, t) = \sin\left(t + \frac{\pi}{6}\right)(5 - x^2(x+y)^2)$$

for $\mathbf{x} = (x, y)^T \in \Omega = [0; 1]^2$ and $0 \leq t \leq T = 0.1$, where the given velocity field is:

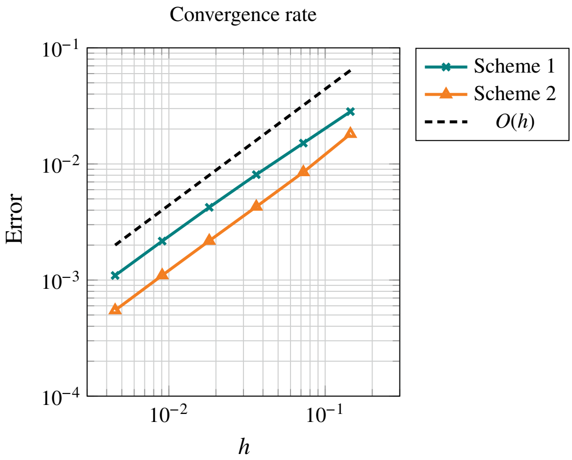
$$\mathbf{v}(\mathbf{x}, t) = \cos(t)\begin{pmatrix} -x^2(x-1)^2(y-1)(y-0.5)y \\ y^2(y-1)^2(x-1)(x-0.5)x \end{pmatrix}, \quad \forall \mathbf{x} = (x, y)^T \in \Omega, \quad \forall t \in [0, T].$$

In the computation, we define $\Gamma_N = \{0\} \times [0, 1]$ and $\Gamma_D = \Gamma \backslash \Gamma_N$. We set $\lambda = 1$ and the value of f in (1) is being computed accordingly. It can easily be checked that $f \geq 0$ in $\Omega \times [0, T]$, so that Theorem 1 ensures that the numerical solution u_h has to remain bounded from below during the whole simulation by:

$$m = \min_{\mathbf{x} \in \Omega} u_{ex}(\mathbf{x}, 0) = 0.5. \tag{8}$$

Simulations are performed on triangulations $\mathcal{T}_i$ $(i = 1, \ldots 6)$, so that $h = 0.145$ for $\mathcal{T}_1$, and the value of h is twice smaller for $\mathcal{T}_i$ than for $\mathcal{T}_{i-1}$ $(2 \leq i \leq 6)$. First, we observe that whatever the simulation considered, u_h remains bounded from below by m defined by (8), as theoretically expected. Then, the error in $L^\infty(0, T; \Omega)$ norm is plotted in Fig. 1 as a function of the mesh size h, in log-log scale. We observe that both schemes are first-order accurate in space. This behavior was clearly expected because of the upwind treatment of the convective term, but also because of the upwind choice in the Joule term (in Scheme 1) or in the diffusion term (in Scheme 2), required to obtain the estimation (7). We observe that the convergence rate is the same and that Scheme 1 leads to a value of Δt_n two orders of magnitude larger than the one used for Scheme 2 in order to make the Newton algorithm converging, even if Scheme 2 seems a little bit more accurate than Scheme 1. The advantage of

Fig. 1 Errors in $L^\infty(0, T; \Omega)$ norm



Scheme 1 is confirmed also considering smaller values of λ, or smaller magnitude of $\mathbf{v}$ or u_{ex}. Finally, other schemes have been proposed and analyzed in [2]. Even if they are of order two in the case $\mathbf{v} = 0$, they verify the maximum principle only under very restrictive conditions on the magnitude of $M - m$.

Acknowledgements This work was supported in part by the Labex CEMPI (ANR-11-LABX-0007-01).

References

1. Bradji, A., Herbin, R.: Discretization of coupled heat and electrical diffusion problems by finite-element and finite-volume methods. IMA J. Numer. Anal. **28**(3), 469–495 (2008). https://doi.org/10.1093/imanum/drm030
2. Calgaro, C., Colin, C., Creusé, E.: A combined finite volume - finite element scheme for a low-Mach system involving a Joule term. AIMS Math. **5**(1), 311–331 (2020)
3. Calgaro, C., Colin, C., Creusé, E., Zahrouni, E.: Approximation by an iterative method of a low-Mach model with temperature dependant viscosity. Math. Methods Appl. Sci. **42**, 250–271 (2019)
4. Chainais-Hillairet, C.: Discrete duality finite volume schemes for two-dimensional drift-diffusion and energy-transport models. Internat. J. Numer. Methods Fluids **59**(3), 239–257 (2009). https://doi.org/10.1002/fld.1393
5. Chainais-Hillairet, C., Peng, Y.J., Violet, I.: Numerical solutions of Euler-Poisson systems for potential flows. Appl. Numer. Math. **59**(2), 301–315 (2009). https://doi.org/10.1016/j.apnum.2008.02.006
6. Colin, C.: Analyse et simulation numérique par méthode combinée volumes finis—eléments finis de modèles de type faible mach. Ph.D. thesis, Université de Lille (2019)
7. Eymard, R., Gallouët, T., Herbin, R.: Finite volume methods. In: Handbook of numerical analysis, Vol. VII, Handb. Numer. Anal., VII, pp. 713–1020. North-Holland, Amsterdam (2000)

8. Eymard, R., Gallouët, T., Herbin, R.: A cell-centered finite-volume approximation for anisotropic diffusion operators on unstructured meshes in any space dimension. IMA J. Numer. Anal. **26**(2), 326–353 (2006). https://doi.org/10.1093/imanum/dri036
9. Huang, F., Tan, W.: On the strong solution of the ghost effect system. SIAM J. Math. Anal. **49**(5), 3496–3526 (2017). https://doi.org/10.1137/16M106964X
10. Levermore, C., Sun, W., Trivisa, K.: Local well-posedness of a ghost system effect. Indiana Univ. Math. J. **60**, 517–576 (2011)

On the L^2 Stability of Finite Volumes for Stationary First Order Systems

Michaël Ndjinga and Sédrick Kameni Ngwamou

Abstract The aim of this paper is two-folds. Firstly we study first order stationary systems of PDEs of the form $\sum_k A_k \partial_k U + KU = 0$ with $K \not> 0$ on $\mathbb{R}^d$. We prove that the classical assumption $K > 0$ is not necessary for the well-posedness of the system and is violated in the particular case of the first order Poisson problem. Secondly we prove the L^2 stability of the finite volume discretisations provided the term KU is appropriately discretised on faces. Our result relies on a discrete Gagliardo-Nirenberg-Sobolev inequality to be submitted [15].

Keywords First order systems · Elliptic systems · Finite volume method · Source upwinding

MSC (2010) 35F45 · 35J46 · 35J05 · 65N08 · 65N12

1 Introduction

Let $d \in \mathbb{N}^*$. We consider symmetric systems of first order PDEs

$$\sum_{k=1}^{d} A_k \partial_k U(x) + KU(x) = F(x), \tag{1}$$

M. Ndjinga (✉)
Université Paris-Saclay, CEA Saclay, DEN/DM2S/STMF, 91191
Gif-sur-Yvette, France
e-mail: michael.ndjinga@cea.fr

S. K. Ngwamou
Laboratoire d'analyse et applications, Université de Yaoundé I, 812
Yaoundé, Cameroun
e-mail: sedrick.ngwamou@aims-cameroon.org

R. Klöfkorn et al. (eds.), *Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples*, Springer Proceedings in Mathematics & Statistics 323,
https://doi.org/10.1007/978-3-030-43651-3_38

with unknown $U \in L^2(\mathbb{R}^d)$, where $A_k, k = 1, \dots d$, and K are $m \times m$ matrices and $F \in L^2(\mathbb{R}^d)$. (1) takes the conservative form $\nabla \times \mathscr{F}(U) + KU = F$ with linear flux $\mathscr{F}(U)\mathbf{n} = A(\mathbf{n})U$ and jacobian matrix $A(\mathbf{n}) = \sum_{k=1}^{d} n_k A_k$ where $\mathbf{n} = (n_1, \dots, n_d)$.

The classical theory of Friedrichs systems ([9, 11] Sect. 5) covers the case of symmetric systems with $K > 0$, using a variational formulation and Lax-Milgram theorem. The coercivity required for the Lax-Milgram theorem is a consequence of the assumption $K > 0$.

In several important cases however, one has to consider systems (1) with $K \not> 0$. The first class of examples are conservation laws with source term in the stationary regime. In the particular case of gas dynamics, taking into account a friction force [6], a Coriolis force [1] or a chemical reaction ([12] Chap. 2 Sect. 5) yields $K \not> 0$.

The second class consists in the mixed formulation of stationary diffusion problems $\nabla \times (D\boldsymbol{\nabla} u) = f$. The prototypical example of diffusion equation is the Poisson problem whose first order reduction is given in Example 1. Finite volume schemes for stationary diffusion on unstructured meshes can have a very complex design (see for instance [10]). The authors however believe that the discretisation of the mixed formulation of stationary diffusion on unstructured meshes will yield simpler schemes with smaller stencils and linear systems that are larger but with better condition number $\left(\mathscr{O}\left(\frac{1}{h}\right) \text{ instead of } \mathscr{O}\left(\frac{1}{h^2}\right)\right)$.

The main objective of this paper is to lay the ground for the numerical analysis of finite volume methods for stationary first order systems with $K \not> 0$. In this first account of our research we set the problem on $\mathbb{R}^d$ and emphasize the importance at both continuous and discrete level of a proper handling of the order 0 term KU, which plays an important role in the well-posedness and stability analysis.

We quickly review in Sect. 2 the well-posedness of the problem (1) on $\mathbb{R}^d$. We prove that $K > 0$ is not a necessary condition but only a particular case. There is even no need for K to be invertible as shown in Theorem 1 and Corollary 2. In this short paper we investigate only the case $\Omega = \mathbb{R}^d$ but the well-posedness results extend to the case of bounded Ω if the BNB theorem [9] is used instead of the Lax-Milgram theorem.

Then in Sect. 3 we study finite volume discretisations of the problem (1) and prove their stability. We consider first order upwind type schemes for the order 1 terms as often done for transient hyperbolic systems [16, 18] without source term. Source upwinding is usually performed in the hyperbolic community to guarantee a good capture of stationary states using transient schemes (well-balanced property) [2, 4, 13, 14] or an asymptotic preserving property [6, 13]. The assumption $K > 0$ yields a straightforward proof of L^2 stability (Theorem 3). However in the case $K \not> 0$, one needs a discrete Gagliardo-Nirenberg-Sobolev inequality and an appropriately discretisation of the term KU on the faces of the mesh (Theorem 4).

2 The Continuous Setting

In this section we give without proof two Theorems (1 and 2) that are straightforward applications of the Fourier transform on system (1) as done in [17] Sect. 3.1.

Existence of solutions to (1) is guaranteed provided $-iA(\boldsymbol{\xi}) + K$ is an invertible matrix for almost every $\xi \in \mathbb{R}^d$ and $(-iA(\boldsymbol{\xi}) + K)^{-1}\hat{F} \in L^2(\mathbb{R}^d)^m$.

Theorem 1 (Existence for first order systems) *Let $A_1, \ldots, A_d$, K be $m \times m$ matrices and $F \in L^2(\mathbb{R}^d)^m$ such that*

$$(-iA(\boldsymbol{\xi}) + K)^{-1}\,\hat{F}(\boldsymbol{\xi}) \in L^2(\mathbb{R}^d)^m. \tag{2}$$

Then (1) admits a unique solution $U \in L^2(\mathbb{R}^d)^m$.

Corollary 1 (Case of Friedrichs' systems) *Assume A_k, $k = 1, \ldots, d$ and K are symmetric matrices, and that $K > 0$. Then for any $F \in L^2(\mathbb{R}^d)^m$, the system (1) admits a unique solution U in $L^2(\mathbb{R}^d)^m$.*

Proof We prove that $\sigma_{\min}(iA(\boldsymbol{\xi}) + K)^{-1} \in L^\infty(\mathbb{R}^d)$ and then apply Theorem 1. Since A_k, $k = 1, \ldots, d$ and K are symmetric matrices, they are diagonalisable with real eigenvalues in an orthonormal basis of $\mathbb{R}^d$ and therefore

$$\begin{aligned} &\forall X \in \mathbb{C}^m, \quad {}^t\bar{X}A_kX \in \mathbb{R}, k = 1, \ldots, d,\\ &\forall X \in \mathbb{C}^m, \quad {}^t\bar{X}KX \in \mathbb{R}, \end{aligned}$$

from which we deduce

$$\forall X \in \mathbb{C}^m, \boldsymbol{\xi} \in \mathbb{R}^d \quad |{}^t\bar{X}(iA(\boldsymbol{\xi}) + K)X| = |i{}^t\bar{X}A(\boldsymbol{\xi})X + {}^t\bar{X}KX)| \geq \lambda_{\min}(K)||X||^2.$$

Since

$$\forall X \in \mathbb{C}^m, \boldsymbol{\xi} \in \mathbb{R}^d \quad |{}^t\bar{X}(iA(\boldsymbol{\xi}) + K)X| \leq ||X||\,||(iA(\boldsymbol{\xi}) + K)X||,$$

we deduce

$$\forall X \in \mathbb{C}^m, \boldsymbol{\xi} \in \mathbb{R}^d \quad \lambda_{\min}(K)||X||^2 \leq ||X||\,||(iA(\boldsymbol{\xi}) + K)X||$$

and finally since $K > 0$ implies $\lambda_{\min}(K) > 0$ we deduce

$$\sigma_{\min}(iA(\boldsymbol{\xi}) + K)^{-1} \leq \frac{1}{\lambda_{\min}(K)}.$$

Hence $\sigma_{\min}(iA(\boldsymbol{\xi}) + K)^{-1} \in L^\infty(\mathbb{R}^d)$. As a consequence for any $F \in L^2(\mathbb{R}^d)^m$, $||(iA(\boldsymbol{\xi}) + K)^{-1}\hat{F}||_2 \leq \sigma_{\min}(iA(\boldsymbol{\xi}) + K)^{-1}||\hat{F}||_2 \in L^2(\mathbb{R}^d)$ and Theorem 1 yields the existence of a unique solution U in $L^2(\mathbb{R}^d)^m$ to the system (1). □

Example 1 (*First order reduction of the Poisson problem*) The first order reduction of the Poisson problem $-\Delta u = f$ amounts to defining $\mathbf{v} = \nabla u$ and to solve the symmetric system

$$\begin{pmatrix} 0 & -\nabla\times \\ -\nabla & 0 \end{pmatrix}\begin{pmatrix} u \\ \mathbf{v} \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & \mathbb{I}_d \end{pmatrix}\begin{pmatrix} u \\ \mathbf{v} \end{pmatrix} = \begin{pmatrix} f \\ 0 \end{pmatrix}. \tag{3}$$

The system (3) takes the form (1) with the following symmetric jacobian and singular friction matrices as well as right hand side function

$$A_{Poisson}(\boldsymbol{\xi}) = \begin{pmatrix} 0 & {}^t\boldsymbol{\xi} \\ \boldsymbol{\xi} & 0 \end{pmatrix}, \quad K_{Poisson} = \begin{pmatrix} 0 & 0 \\ 0 & \mathbb{I}_d \end{pmatrix}, \quad F_{Poisson}(x) = \begin{pmatrix} -f(x) \\ 0 \end{pmatrix}. \tag{4}$$

We cannot use Corollary 1 for the well-posedness of system (3) since $K \not> 0$. Theorem 1 should be used instead.

Corollary 2 (Existence for first order Poisson system) *Let* $d \in \mathbb{N}^*, f \in L^2(\mathbb{R}^d)$ *such that*

$$\frac{\hat{f}}{||\boldsymbol{\xi}||^2} \in L^2(\boldsymbol{\xi}), \quad \frac{\hat{f}}{||\boldsymbol{\xi}||^2}\boldsymbol{\xi} \in L^2(\boldsymbol{\xi})^d. \tag{5}$$

Let the matrices $A_{poisson}(\boldsymbol{\xi})$, $K_{poisson}$ *and* $F_{poisson}(\boldsymbol{\xi})$ *defined in (4). Then the system (3) admits a unique solution* $(u, \mathbf{v}) \in L^2(\mathbb{R}^d)^{d+1}$.

Proof Since $(-iA_{Poisson}(\boldsymbol{\xi}) + K_{Poisson})^{-1}\,\hat{F}_{Poisson}(\boldsymbol{\xi}) = \frac{\hat{f}(\boldsymbol{\xi})}{||\boldsymbol{\xi}||^2}\begin{pmatrix} 1 \\ -i\boldsymbol{\xi} \end{pmatrix}$, according to Theorem (1), the assumptions Corollary 2 yield the existence of a unique solution in $L^2(\boldsymbol{\xi}) \times L^2(\boldsymbol{\xi})$ (actually $H^1(\boldsymbol{\xi}) \times L^2(\boldsymbol{\xi})$). □

The assumption (5) on f in Corollary (2) is to be compared with assumption (6) when the original second order problem is solved in L^2.

Theorem 2 (Existence for Poisson equation in L^2) *Let* $d \in \mathbb{N}^*, f \in L^2(\mathbb{R}^d)$ *such that*

$$\frac{\hat{f}}{||\boldsymbol{\xi}||^2} \in L^2(\boldsymbol{\xi}). \tag{6}$$

Then there exists a unique $u \in L^2(\mathbb{R}^d)$ *such that* $-\Delta u = f$ *in the weak (distributional) sense.*

Corollary 2 has more stringent assumptions than Theorem 2. Indeed the first order reduction require that $v = \nabla u \in L^2(\mathbb{R})^d$, hence the solution u given in Corollary (2) is actually in H^1. The first order reduction is not able to represent very weak solutions $u \notin H^1$. However this should not be a serious issue since H^1 is the usual solution space in classical variational formulations.

3 The Discrete Setting

In this section we consider the system (1) with symmetric matrices $A_1, \dots, A_d$ and K. We propose a numerical method that is stable under a similar assumption on $A_1, \dots, A_d, K$ as that of Theorem 1.

Using a finite volume discretisation, $\mathbb{R}^d = \cup_{i \in \mathbb{N}} C_i$ is decomposed into polygonal cells C_i with volume $v_i > 0$. Neighbouring cells C_i and C_j are separated by an interface f_{ij} having measure $s_{ij} > 0$ and unit normal $\mathbf{n}_{ij}$ oriented from C_i towards C_j ($\mathbf{n}_{ji} = -\mathbf{n}_{ij}$). The distance between the centers of mass of two neighbouring cells is noted d_{ij}. The set $\nu(i)$ contains all the indices j such that C_j and C_i have a common interface. $s_i = \sum_{j \in \nu(i)} s_{ij}$ is the perimeter of the cell C_i. We recall that Green theorem yield $\forall i \in \mathbb{N},\ \sum_{j \in \nu(i)} s_{ij} \mathbf{n}_{ij} = 0$.

We consider a finite volume discretisation where the unknown $U(x)$ is approximated by a piecewise constant function $\mathscr{U}$ taking the value U_i on the cell C_i. Similarly, the right hand side function $F(x)$ is approximated by a piecewise constant function $\mathscr{G}$ taking the value F_i on the cell C_i.

We propose to use an upwind scheme for the fluxes and an upwind scheme for the friction term KU:

$$\frac{1}{v_i} \sum_{j \in \nu(i)} s_{ij} \mathscr{F}_{ij} + \frac{1}{s_i} \sum_{j \in \nu(i)} s_{ij} K U_{ij} = F_i, \tag{7}$$

where the numerical fluxes $\mathscr{F}_{ij}$ and the source interfacial state U_{ij} take the upwind form:

$$\mathscr{F}_{ij} = \frac{\mathscr{F}(U_i) + \mathscr{F}(U_j)}{2} \mathbf{n}_{ij} + D^{\mathscr{F}}_{ij} \frac{U_i - U_j}{2} = \mathscr{F}(U_i) - \frac{A(\mathbf{n}_{ij}) - D^{\mathscr{F}}_{ij}}{2} (U_i - U_j) \tag{8}$$

$$U_{ij} = \frac{U_i + U_j}{2} + D^K_{ij} \frac{U_i - U_j}{2} = U_i - \frac{\mathbb{I}_m - D^K_{ij}}{2} (U_i - U_j), \tag{9}$$

where $D^{\mathscr{F}}_{ij}$ is the flux upwind matrix which is assumed to satisfy $D^{\mathscr{F}}_{ji} = D^{\mathscr{F}}_{ij}$. The choice $D^{\mathscr{F}}_{ji} = |A(\mathbf{n}_{ij})|$ gives the classical upwind scheme flux. $D^{\mathscr{F}}_{ij}$ is the source upwind matrix. The interfacial states U_{ij} and U_{ji} are identical provided $D^K_{ji} = -D^K_{ij}$. The choice $D^K_{ij} = \mathbb{I}_m$ gives the classical centered scheme $U_{ij} = U_i \neq U_{ji} = U_j$ for the source term KU.

Lemma 1 *Let $A_1, \dots, A_d$ and K be symmetric matrices. Consider a mesh $\mathscr{M}$, where each interface f_{ij} is associated with a flux upwind matrix D_{ij} and a source upwind matrix D^K_{ij} such that*

$$D^{\mathscr{F}}_{ij} = D^{\mathscr{F}}_{ji}, \quad D^K_{ji} = D^K_{ij} \tag{10}$$

Then any solution of the numerical scheme (7), (8) and (9) satisfies

$$\sum_{i\in\mathbb{N}} v_i {}^tU_iKU_i + \frac{1}{2}\sum_{f_{ij}} s_{ij} {}^t(U_i - U_j)\left(D^{\mathscr{F}}_{ij} - \frac{v_iK(\mathbb{I}_m - D^K_{ij})}{s_i}\right)(U_i - U_j) = \sum_{i\in\mathbb{N}} v_i {}^tU_iF_i. \tag{11}$$

Proof Using the expressions (8) and (9), (7) becomes

$$\frac{1}{v_i}\sum_{j\in\nu(i)} -s_{ij}\frac{A(\mathbf{n}_{ij}) - D^{\mathscr{F}}_{ij}}{2}(U_i - U_j) + \frac{1}{s_i}\sum_{j\in\nu(i)} s_{ij}K\left(U_i - \frac{\mathbb{I}_m - D^K_{ij}}{2}(U_i - U_j)\right) = F_i$$

$$\sum_{j\in\nu(i)} \frac{s_{ij}}{s_i}KU_i - \left(\frac{s_{ij}}{v_i}\frac{A(\mathbf{n}_{ij}) - D^{\mathscr{F}}_{ij}}{2} + \frac{s_{ij}}{s_i}K\frac{\mathbb{I}_m - D^K_{ij}}{2}\right)(U_i - U_j) = F_i$$

$$KU_i - \sum_{j\in\nu(i)}\left(\frac{s_{ij}}{v_i}\frac{A(\mathbf{n}_{ij}) - D^{\mathscr{F}}_{ij}}{2} + \frac{s_{ij}}{s_i}K\frac{\mathbb{I}_m - D^K_{ij}}{2}\right)(U_i - U_j) = F_i.$$

After taking the inner product with v_iU_i we obtain

$$v_i {}^tU_iKU_i - \sum_{j\in\nu(i)} s_{ij} {}^tU_i\left(\frac{A(\mathbf{n}_{ij}) - D^{\mathscr{F}}_{ij}}{2} + \frac{v_i}{s_i}K\frac{\mathbb{I}_m - D^K_{ij}}{2}\right)(U_i - U_j) = v_i {}^tU_iF_i.$$

Since $A(\mathbf{n}_{ji}) = -A(\mathbf{n}_{ij})$, $D^{\mathscr{F}}_{ji} = D^{\mathscr{F}}_{ij}$ and $D^K_{ji} = D^K_{ij}$, summing over i yields

$$\sum_{i\in\mathbb{N}} v_i {}^tU_iKU_i - \sum_{f_{ij}} s_{ij} {}^t(U_i + U_j)A(\mathbf{n}_{ij})(U_i - U_j) + s_{ij} {}^t(U_i - U_j) \tag{12}$$
$$\left(\frac{-s_iD^{\mathscr{F}}_{ij} + v_iK(\mathbb{I}_m - D^K_{ij})}{2s_i}\right)(U_i - U_j) = v_i {}^tU_iF_i.$$

We have ${}^t(U_i + U_j)A(\mathbf{n}_{ij})(U_i - U_j) = {}^tU_iA(\mathbf{n}_{ij})U_i - {}^tU_jA(\mathbf{n}_{ij})U_j$ since $A(\mathbf{n}_{ij})$ is a symmetric matrix and thus (13) becomes

$$\sum_{i\in\mathbb{N}} v_i {}^tU_iKU_i - \frac{1}{2}\sum_{f_{ij}} s_{ij}({}^tU_iA(\mathbf{n}_{ij})U_i - {}^tU_jA(\mathbf{n}_{ij})U_j) + s_{ij} {}^t(U_i - U_j)$$
$$\left(\frac{-s_iD^{\mathscr{F}}_{ij} + v_iK(\mathbb{I}_m - D^K_{ij})}{s_i}\right)(U_i - U_j) = v_i {}^tU_iF_i,$$

which in turn yields

$$\sum_{i\in\mathbb{N}} v_i {}^tU_iKU_i - \sum_{i} {}^tU_iA(\sum_{j\in\nu(i)} s_{ij}\mathbf{n}_{ij})U_i - \sum_{f_{ij}} s_{ij} {}^t(U_i - U_j)$$
$$\left(\frac{-s_iD^{\mathscr{F}}_{ij} + v_iK(\mathbb{I}_m - D^K_{ij})}{2s_i}\right)(U_i - U_j) = v_i {}^tU_iF_i,$$

and since $\sum_{j\in\nu(i)} s_{ij}\mathbf{n}_{ij} = 0$ (Green theorem) we finally obtain (11). □

We can now prove the stability of the scheme where the source are centered provided $K > 0$ and that the fluxes are upwind ($\forall U \in \mathbb{R}^m$, ${}^tUD_{ij}^{\mathscr{F}}U \geq 0$).

Theorem 3 (Stability of the source centered scheme for$K > 0$) *Let $A_1, \ldots, A_d$ and K be symmetric matrices. Assume $K > 0$. Consider a mesh $\mathscr{M}$, where each interface f_{ij} is associated with a flux upwind matrix D_{ij} and a source upwind matrix D_{ij}^K such that*

$$D_{ij}^{\mathscr{F}} = D_{ji}^{\mathscr{F}}, \quad D_{ij}^{\mathscr{F}} \geq 0, \quad D_{ij}^K = \mathbb{I}_m \tag{13}$$

Then any solution of the numerical scheme (7), (8) and (9) satisfies

$$||\mathscr{U}||_2 \leq \frac{1}{\lambda_{\min}(K)}||\bar{F}||_2. \tag{14}$$

Proof The result is a straightforward application of the Lemma 1.
From the assumptions $D_{ij}^K = \mathbb{I}_m$, (11) yields

$$\sum_{i\in\mathbb{N}} v_i {}^tU_iKU_i + \frac{1}{2}\sum_{f_{ij}} s_{ij}{}^t(U_i - U_j)D_{ij}^{\mathscr{F}}(U_i - U_j) = \sum_{i\in\mathbb{N}} v_i {}^tU_iF_i.$$

Since $D_{ij}^{\mathscr{F}} \geq 0$, we have

$$\sum_{i\in\mathbb{N}} v_i {}^tU_iKU_i \leq \sum_{i\in\mathbb{N}} v_i {}^tU_iF_i$$

The inequality (14) is a consequence of Cauchy-Schwarz inequality and of ${}^tUKU \geq \lambda_{\min}(K)||U||_2^2$. □

We can also prove the stability of a scheme where the term KU is upwinded appropriately. This result however requires a discrete Gagliardo-Nirenberg-Sobolev inequality on $\mathbb{R}^d$. We recall that the continuous Gagliardo-Nirenberg-Sobolev inequality $||u||_{p^*} \leq C||\nabla u||_p$ is valid on $\mathbb{R}^d$ for $1 \leq p < d$ and $p^* = \frac{dp}{d-p}$ (see Theorem 9.9 in [7]). The literature on finite volume methods present discrete inequalities Gagliardo-Nirenberg-Sobolev on bounded domains Ω and $p = 2$ [3, 8], except [5] which concerns a modified Galiardo-Nirenberg-Sobolev on so-called admissible meshes.

Theorem 4 (Stability of the source upwinded scheme) *Let $d > 2$, $A_1, \ldots, A_d$ and K be symmetric matrices. Consider a mesh $\mathscr{M}$, where each interface f_{ij} is associated with a flux upwind matrix D_{ij} and a source upwind matrix D_{ij}^K such that*

$$D_{ij}^{\mathscr{F}} = D_{ji}^{\mathscr{F}}, \quad D_{ji}^K = D_{ij}^K. \tag{15}$$

Assume a discrete Gagliardo-Nirenberg-Sobolev inequality ($d > 2$):

$$\exists C > 0, \quad ||\mathscr{U}||^2_{\frac{2d}{d-2}} \leq C \sum_{f_{ij}} \frac{s_{ij}}{d_{ij}}{}^t ||U_i - U_j||^2 \tag{16}$$

Assume that

$$\exists \alpha > 0, \quad \frac{1}{2}\lambda_{\min}\left(D^{\mathscr{F}}_{ij} - \frac{v_i K(\mathbb{I}_m - D^K_{ij})}{s_i}\right) \geq \frac{\alpha}{d_{ij}}$$

Then any solution of the numerical scheme (7), (8) and (9) satisfies

$$\lambda_{\min}(K)||\mathscr{U}||^2_2 + \frac{1}{C}||\mathscr{U}||^2_{\frac{2d}{d-2}} \leq ||\mathscr{G}||_2||\mathscr{U}||_2. \tag{17}$$

Proof The result is a straightforward application of the Lemma 1. □

4 Conclusion and Perspectives

First order systems with $K \not> 0$ are of particular interest in many applications. We studied some first order upwind type methods for stationary first order systems. The present study emphasised the importance of upwinding the term KU in order to obtain a stable discretisation.

We proposed simple finite volume methods applicable to second order elliptic equation via a first order reduction. These methods are based on upwinding both zero and first order terms. The stability result relies on a new discrete Sobolev-Gagiardo-Nirenberg inequality on $\mathbb{R}^d$ [15], which implies the various discrete Sobolev inequalities on bounded domains one can find in the literature [3, 8].

References

1. Audusse, E., Minh, H.D., Omnes, P., Penel, Y.: Analysis of a modified godunov scheme for the linear wave equation with coriolis source term on cartesian meshes. J. Comput. Phys. **373**, 91–129 (2018)
2. Bermudez, A., Vazquez, E.: Upwind methods for hyperbolic conservation laws with source terms. Comp. Fluids **23**(8), 1049–1071 (1994)
3. Bessemoulin-Chatard, M., Chainais-Hillairet, C., Filbet, F.: On discrete functional inequalities for some finite volume schemes. IMA J. Numer. Anal. **35**(3), 1125–1149 (2015)
4. Bouchut, F.: Nonlinear stability of finite volume methods for hyperbolic conservation laws, and well-balanced schemes for sources. In: Frontiers in Mathematics Series. Birkhäuser (2004)
5. Bouchut, F., Eymard, R., Prignet, A.: Finite volume schemes for the approximation via characteristics of linear convection equations with irregular data. J. Evol. Equ. **11**, 687–724 (2011)

6. Bouchut, F., Ounaissa, H., Perthame, B.: Upwinding of the source term at interfaces for euler equations with high friction. Comput. Math. Appl. **53**(3–4), 361–375 (2007)
7. Brezis, H.: Functional Analysis, Sobolev Spaces and PDEs, Universitext. Springer, New York (2010)
8. Coudière, Y., Gallouët, T., Herbin, R.: Discrete sobolev inequalities and Lp error estimates for approximate finite volume solutions of convection diffusion equations. M2AN Math. Model. Numer. Anal **35**, 767–778 (2001)
9. Ern, A., Guermond, J.L.: Theory and Practice of Finite Elements. Applied Mathematical Sciences, vol. 159. Springer, New York (2004)
10. Eymard, R., Gallouët, T., Herbin, R.: Discretisation of heterogeneous and anisotropic diffusion problems on general non-conforming meshes SUSHI: a scheme using stabilisation and hybrid interfaces. IMA J. Numer. Anal. **30**, 1009–1043 (2010)
11. Friedrichs, K.O.: Symmetric positive linear differential equations. Commun. Pure Appl. Math. **11**, 333–418 (1958)
12. Godlewski, E., Raviart, P.A.: Numerical Approximation of Hyperbolic Systems of Conservation Laws. Applied Mathematical Sciences, vol. 118. Springer, New York (1996)
13. Gosse, L.: Computing qualitatively correct approximations of balance laws. SIMAI Springer Series, vol. 2. Springer, Mailand (2013)
14. Greenberg, J.M., Leroux, A.Y.: A well-balanced scheme for the numerical processing of source terms in hyperbolic equations. SIAM J. Numer. Anal. **33** (1996)
15. Ndjinga, M.: A discrete gagliardo-nirenberg-sobolev inequality. To be submitted
16. Ndjinga, M.: L^2 stability of nonlinear finite volume schemes for linear hyperbolic systems. C. R. Acad. Sci. Paris **351**, 707–711 (2013)
17. Serre, D.: Systems of Conservation Laws I: Hyperbolicity, Entropies, Shock Waves. Cambridge University Press, Cambridge (1999)
18. Vila, J.P., Villedieu, P.: Convergence de la méthode des volumes finis pour les systèmes de friedrichs. C. R. Acad. Sci. Paris **325** (1997)

A New Class of L^2-Stable Schemes for the Isentropic Euler Equations on Staggered Grids

Michaël Ndjinga and Katia Ait-Ameur

Abstract Staggered schemes for compressible flows are highly non linear and the stability analysis has historically been performed with a heuristic approach and the tuning of numerical parameters [12]. We investigate the L^2-stability of staggered schemes by analysing their numerical diffusion operator. The analysis of the numerical diffusion operator gives new insight into the scheme and is a step towards a proof of linear stability or stability for almost constant initial data. For most classical staggered schemes [9–11, 14], we are able to prove the positivity of the numerical diffusion only in specific cases (constant sign velocities). We then propose a class of linearly L^2-stable staggered schemes for the isentropic Euler equations based on a carefully chosen numerical diffusion operator. We give an example of such a scheme and present some first numerical results on a Riemann problem.

Keywords Euler equations · Compressible flows · Finite volumes · Staggered grids · Stability analysis · Numerical diffusion

MSC (2010) 35L65 · 35Q35 · 65M08 · 65M12

M. Ndjinga
Université Paris-Saclay, CEA Saclay, DEN/DM2S/STMF,
91191 Gif-sur-Yvette, France
e-mail: michael.ndjinga@cea.fr

K. Ait-Ameur (✉)
Laboratoire Jacques Louis Lions (LJLL), Sorbonne Université,
75005 Paris, France
e-mail: aitameur.katia@gmail.com

Université Paris-Saclay, CEA Saclay, DEN/DM2S/STMF,
91191 Gif-Sur-Yvette Cedex, France

R. Klöfkorn et al. (eds.), *Finite Volumes for Complex Applications IX - Methods,
Theoretical Aspects, Examples*, Springer Proceedings in Mathematics & Statistics 323,
https://doi.org/10.1007/978-3-030-43651-3_39

1 Introduction

As an introduction to the issue, we consider a 1D conservative non linear hyperbolic system

$$\partial_t U(x,t) + \partial_x F(U)(x,t) = 0, \tag{1}$$

with unknown vector $U \in \mathbb{R}^m$ and Lipschitz flux $F : \mathbb{R}^m \to \mathbb{R}^m$ with real-diagonalisable Jacobian matrix $A(U) = \nabla_U F(U) \in \mathbb{R}^{m\times m}$.

When approximating smooth solutions U of (1) by a consistent numerical method on a regular mesh with space step $\triangle x$, the semi-discrete equations approximate to the first order in $\triangle x$ the following perturbed version of Eq. (1):

$$\partial_t U_{\triangle x} + \partial_x F(U_{\triangle x}) = \mathscr{D}(U_{\triangle x}, \triangle x) + o(\triangle x), \tag{2}$$

where $U_{\triangle x}$ is the numerical solution and $\mathscr{D}$ is a second order differential operator.

When the flux function F is linear, linear numerical methods yield a linear diffusion operator $\mathscr{D}(U, \triangle x) = \triangle x \partial_x (D \partial_x U)$. The matrix D comes from the upwind (off-centered) contributions of the discrete equations and gives a first insight into the scheme precision and stability. In the non linear case (F Lipschitz), the numerical diffusion operator can often be approximated to the first order by a non linear diffusion operator:

$$\mathscr{D}(U, \triangle x) = \triangle x \partial_x (D(U) \partial_x U) + o(\triangle x). \tag{3}$$

This is the case for instance for colocated schemes based on characteristic upwinding such as Godunov [6], Roe [15], VFRoe [13] or VFFC [4] schemes where the non linear numerical diffusion tensor is $D(U) = |A(U)|$.

We recall that in the case of symmetric hyperbolic systems (${}^t A(U) = A(U)$), any entropy solution to (1) preserves the L^2 norm (see [5] Example 3.2 in the Introduction chapter) and we would like the discrete L^2 norm of any scheme to be bounded as well. In the case of non-symmetric systems, one first symmetrises the system using entropic variables $V(U) = \nabla s(U)$ where s a strictly convex entropy of the system (1) is assumed to exist (see [5] Theorem 3.2 in the introduction chapter). The new symmetric system:

$$\partial_t V + \bar{A}(V) \partial_x V = 0, \text{ with } {}^t \bar{A} = \bar{A} \tag{4}$$

is linearly L^2-stable. Any numerical scheme yields a numerical diffusion $\bar{D}(V)$ in the symmetrised basis and we require that the diffusion operator $\bar{D}$ have positive symmetric part: ${}^t \bar{D} + \bar{D} \geq 0$.

In this first account of our research, we investigate the wave system which raises an issue that will remain with the more complex fluid model: the numerical treatment of the mass balance equation yields a non classical diffusion operator. In order to obtain a straightforwardly stable scheme we propose a new discretisation with positive numerical diffusion.

In Sect. 2, we determine the numerical diffusion of the staggered schemes and show it does not straightforwardly yield a linear stability. We then present a new class of staggered schemes and prove their linear stability in Sect. 3. Some first numerical results are given in Sect. 4.

2 The Numerical Diffusion of Staggered Schemes for the Linear Wave System

For compressible flows at low Mach numbers, it is customary to approximate the Euler system by the wave system (see for instance [2, 3]). The linear wave system is the first order reduction of the classical linear wave equation $\partial_{tt}\rho - c^2\partial_{xx}\rho = 0$ and can be written in the following conservative form:

$$\begin{cases} \partial_t \rho + \quad \partial_x q = 0 \\ \partial_t q + c^2 \partial_x \rho = 0 \end{cases}. \tag{5}$$

The analysis of the wave system can be extended to the Euler system but the calculations are lengthier and do not help the understanding of the main result.

The wave system (5) is a linear PDE system and the characteristic based upwind method is linear. Therefore the finite volume upwind scheme yields a classical linear diffusion operator (3) with constant diffusion tensor: $D = c\mathbb{I}_d$.

The upwind scheme can be proven to be stable. However the amount of numerical diffusion is proportional to the sound speed c and for low Mach number flows where $q \ll \rho c$ the schemes based on characteristics upwinding are not able to capture nearly incompressible solutions (see [2, 3] for more details).

On the contrary, staggered schemes are known to be more precise for low Mach number flows in practice and are very popular in the thermal hydraulics community [14]. However, their stability analysis has historically been based on heuristics [12]. Yet the conservative staggered schemes presented in [9, 10] is proven to be entropic and to satisfy a kinetic energy preservation [11]. Likewise in [1], the authors present a kinetic scheme on staggered grids for the barotropic Euler equations and derive stability conditions which preserve the positivity of the density and the decay of the discrete global entropy and satisfy a kinetic energy preservation. Unfortunately the boundedness of the entropy does not necessarily imply the boundedness of the solution. Indeed a strictly convex function is not necessarily bounded below. This is in particular the case for the full Euler system since the entropy involves the function—ln which is strictly convex but not bounded below (see [5] Example 3.3 in the introduction chapter). In the next subsection we show that the first order perturbed Eq. (2) associated to staggered schemes yields not the classical diffusion operator (3) but instead a strongly non linear numerical diffusion operator.

2.1 The Staggered Scheme of Herbin et al.

Using staggered schemes, the density and pressure are located on cells and the velocity on faces (nodes in 1D) [7, 8]. The momentum variable is usually split as a product between the density and the velocity: $\mathbf{q} = \rho\mathbf{u}$. The main difference between the various staggered schemes is the treatment of the convection term $\rho\mathbf{u} \otimes \mathbf{u}$ in the momentum equation. Since we work with the wave system, they differ only by the use of $\mathbf{u}$ (non conservative approach, see [14] Sect. 11.2) or $\rho\mathbf{u}$ as main variable. We consider the staggered scheme of [11] which is conservative and entropic. For simplicity, we present the semi-discrete equations of the fully implicit variant ([10] Sect. 3, [9] Sect. 2.1) for the 1D wave system:

$$\begin{cases} \partial_t \rho_i + \frac{1}{\Delta x}(\rho^{up}_{i+\frac{1}{2}} u_{i+\frac{1}{2}} - \rho^{up}_{i-\frac{1}{2}} u_{i-\frac{1}{2}}) = 0 \\ \partial_t \left(\bar{\rho}_{i+\frac{1}{2}} u_{i+\frac{1}{2}}\right) + c^2 \frac{1}{\Delta x}(\rho_{i+1} - \rho_i) = 0. \end{cases} \tag{6}$$

The mass flux ρu at the cell interfaces is defined using an upwind density $\rho^{up}_{i+\frac{1}{2}}$ defined as:

$$\begin{aligned} \rho^{up}_{i+\frac{1}{2}} &= \begin{cases} \rho_i & \text{if } u_{i+\frac{1}{2}} > 0 \\ \rho_{i+1} & \text{if } u_{i+\frac{1}{2}} \leq 0 \end{cases} \\ &= \frac{\rho_i + \rho_{i+1}}{2} + sign(u_{i+\frac{1}{2}}) \frac{\rho_i - \rho_{i+1}}{2}, \end{aligned} \tag{7}$$

which is the sum of a centered and an upwind term.

The expression of $\bar{\rho}_{i+\frac{1}{2}}$ in the discrete momentum equation amounts to an average of the neighbouring densities

$$\bar{\rho}_{i+\frac{1}{2}} = \frac{1}{2}(\rho_i + \rho_{i+1}). \tag{8}$$

2.2 The Numerical Diffusion

In this section, we assume that the exact solution is smooth and we determine the numerical diffusion of the scheme (6). The first order momentum perturbed equation is straightforward:

$$\partial_t(\rho u)(x_{i+\frac{1}{2}}, t) + c^2 \partial_x \rho(x_{i+\frac{1}{2}}, t) = \frac{c^2}{2}(\Delta x)\partial_{xx}\rho(x_{i+\frac{1}{2}}, t) + \mathscr{O}(\Delta x^2).$$

After some calculations, we find that the mass flux is given to the first order:

$$\frac{\rho^{up}_{i+\frac{1}{2}}u_{i+\frac{1}{2}} - \rho^{up}_{i-\frac{1}{2}}u_{i-\frac{1}{2}}}{\Delta x} = \partial_x(\rho u)(x_i,t) - \frac{\Delta x}{2} sign(u(x_i,t))\partial_x(u\partial_x\rho)(x_i,t) + \mathscr{O}(\Delta x^2).$$

Hence the following result on the numerical diffusion operator of the staggered scheme (6).

Theorem 1 (Numerical diffusion of staggered schemes) *The second order perturbation operator associated to the staggered scheme (6) on a 1D regular mesh with space step Δx is the strongly non linear operator:*

$$\mathscr{D}(U,\Delta x) = \Delta x \begin{pmatrix} sign(u) & 0 \\ 0 & 1 \end{pmatrix} \partial_x \left[\begin{pmatrix} u & 0 \\ c^2 & 0 \end{pmatrix} \partial_x \begin{pmatrix} \rho \\ q \end{pmatrix} \right] + o(\Delta x). \tag{9}$$

The numerical diffusion associated to the mass conservation law is the term $sign(u)\partial_x(u\partial_x\rho)$ and is decoupled from the momentum diffusion. The linear stability analysis of such a strongly non linear diffusion is not classical and we are not aware of any reference in the litterature.

If we assume that u does not change sign then the diffusion term simplifies to the weakly non linear diffusion term $\partial_x(|u|\partial_x\rho)$ which involves a positive diffusion coefficient $|u|$. The weakly non linear diffusion term $\partial_x(|u|\partial_x\rho)$ can be linearised around a constant state $(\rho_0, u_0 \neq 0$ as $\partial_x(|u_0|\partial_x\rho) + \partial_x(|u|\partial_x\rho_0) = |u_0|\partial_{xx}\rho)$. Hence if $u > 0$ or $u < 0$ the mass equation has a positive contribution on the diagonal of the numerical diffusion tensor D and thus has a stabilising effect.

If we allow u to change sign then the multiplication with $sign(u)$ makes things more complicated and we can not rule out potential instabilities. The linearisation is not trivial, even taking u smooth enough, since the function $sign(u)$ is not continuous. The consistency analysis is only a first step that requires smooth solutions but the final goal of capturing discontinuous weak solutions with velocity that change sign will raise even more issues.

3 A New Class of Schemes for the Isentropic Euler Equations

We now propose a class of staggered schemes for conservation laws (1) which admits a classical diffusion operator (3) and such that the diffusion tensor satisfies: $\bar{D} + {}^t\bar{D} \geq 0$. We will prove that these schemes are linearly L^2-stable.

We specify this new class in the particular case of the following 1D isothermal Euler equations in conservative form:

$$\begin{cases} \partial_t \rho + \partial_x q = 0 \\ \partial_t q + \partial_x \dfrac{q^2}{\rho} + \partial_x p = 0. \end{cases} \tag{10}$$

which takes the form (1) with

$$U = \begin{pmatrix} \rho \\ q \end{pmatrix}, \quad F(U) = \begin{pmatrix} q \\ \frac{q^2}{\rho} + p, \end{pmatrix}.$$

The results can be extended to the multidimensional Euler system but the calculations are lengthier and do not help the understanding of the main properties.

We consider the class $Stag$ of discrete staggered conservative schemes of the form:

$$U_i'(t) + \frac{F_{i,i+1} - F_{i-1,i}}{\Delta x} = 0, \quad \text{with: } U_i = \begin{pmatrix} \rho_i \\ q_{i+\frac{1}{2}} \end{pmatrix}, \text{ and:} \tag{11}$$

$$F_{i,i+1} = \frac{F(U_i) + F(U_{i+1})}{2} + D_{Stag}(U_i, U_{i+1})\frac{U_i - U_{i+1}}{2}, \tag{12}$$

where D_{Stag} is a 2×2 matrix-valued function. An example of a scheme in the class $Stag$ is the staggered centered scheme, which correspond to the case $D_{Stag} = 0$.

Schemes of the class $Stag$ admit a classical diffusion operator (3).

Theorem 2 (Classical diffusion of $Stag$ schemes) *Let $D_{Stag} : \mathbb{R}^2 \to \mathbb{R}^{2\times 2}$ be a matrix valued Lipschitz function. A staggered conservative scheme (11) with a numerical flux $F_{i,i+1}$ of the form (12) admits the following classical diffusion operator*

$$\mathscr{D}(U, \Delta x) = \Delta x \partial_x (D_{Stag}(U, U)\partial_x U) + o(\triangle x).$$

on a regular mesh with space step Δx.

Schemes of the class $Stag$ are L^2-stable.

Theorem 3 (L^2-stability of $Stag$ schemes) *A staggered conservative scheme (11) with a numerical flux $F_{i,i+1}$ of the form (12) such that the diffusion operator D_{Stag} satisfies: $\bar{D}_{Stag} + {}^t\bar{D}_{Stag} \geq 0$, is linearly L^2-stable.*

Proof After the symmetrisation (4) and linearisation, the isentropic Euler system takes the form:

$$\partial_t \bar{V} + \bar{A}(V_0)\partial_x \bar{V} = 0, \quad \bar{V} = \begin{pmatrix} c\rho \\ q - \rho u_0 \end{pmatrix}$$

We consider $\bar{V}$ a solution of the conservative scheme:

$$\bar{V}_i'(t) + \frac{\bar{F}_{i,i+1} - \bar{F}_{i-1,i}}{\Delta x} = 0, \text{ where: } \bar{F}_{i,i+1} = \bar{A}(V_0)\frac{\bar{V}_i + \bar{V}_{i+1}}{2} + \bar{D}_{Stag}(V_0)\frac{\bar{V}_i - \bar{V}_{i+1}}{2}$$

We compute the evolution in time of $||\bar{V}||_2^2$ using the symmetry of $\bar{A}(V_0)$ and the positiveness of $\bar{D}_{Stag}(V_0)$.

$$\begin{aligned}\frac{1}{2}\frac{d||\bar{V}||_2^2}{dt} &= \bar{V}\cdot\frac{d\bar{V}}{dt} = \sum_i \Delta x\bar{V}_i\cdot\frac{d\bar{V}_i}{dt} = -\sum_i \bar{V}_i\cdot(\bar{F}_{i,i+1}-\bar{F}_{i-1,i})\\ &= -\frac{1}{2}\sum_i \bar{V}_i\cdot\bar{A}(V_0)(\bar{V}_{i+1}-\bar{V}_{i-1}) - \frac{1}{2}\sum_i \bar{V}_i\cdot\bar{D}_{Stag}(V_0)(\bar{V}_i-\bar{V}_{i+1})\\ &+ \frac{1}{2}\sum_i \bar{V}_i\cdot\bar{D}_{Stag}(V_0)(\bar{V}_{i-1}-\bar{V}_i)\end{aligned}$$

Since: $\bar{A}(V_0) = {}^t\bar{A}(V_0)$, we have: $\sum_i \bar{V}_i\cdot\bar{A}(V_0)(\bar{V}_{i+1}-\bar{V}_{i-1}) = 0$, and:

$$\frac{1}{2}\frac{d||\bar{V}||_2^2}{dt} = -\frac{1}{2}\sum_i(\bar{V}_i-\bar{V}_{i+1})\cdot\bar{D}_{Stag}(V_0)(\bar{V}_i-\bar{V}_{i+1})$$

Since: $\bar{D}_{Stag}(V_0) + {}^t\bar{D}_{Stag}(V_0) \geq 0$, we obtain: $\frac{1}{2}\frac{d||\bar{V}||_2^2}{dt} \leq 0$. □

Corollary 1 *The numerical scheme (11)–(12) based on the following numerical flux:* $F_{i,i+1} = \begin{pmatrix}\bar{q}_{i+\frac{1}{2}}\\ \frac{\bar{q}_{i+1}^2}{\rho_{i+1}} + p_{i+1}\end{pmatrix}$, *with:*

$$\begin{aligned}\bar{q}_{i+\frac{1}{2}} &= q_{i+\frac{1}{2}} + (|u|-u)\frac{\rho_i-\rho_{i+1}}{2}\\ \frac{\bar{q}_{i+1}^2}{\rho_{i+1}} &= \frac{q_{i+\frac{1}{2}}^2}{\rho_i} + (|u|-u)\frac{q_{i+\frac{1}{2}}-q_{i+\frac{3}{2}}}{2}\end{aligned}$$

where u is the Roe average velocity [15]: $u = \frac{\frac{q_{i+\frac{1}{2}}}{\sqrt{\rho_i}}+\frac{q_{i+\frac{3}{2}}}{\sqrt{\rho_{i+1}}}}{\sqrt{\rho_i}+\sqrt{\rho_{i+1}}}$, *is linearly L^2-stable.*

Proof The diffusion operator of the proposed scheme is:

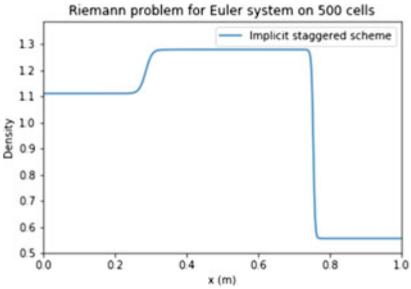
$$D_{Stag} = \begin{pmatrix}|u|-u & 1\\ -c^2-u^2 & |u|+u\end{pmatrix}, \quad \bar{D}_{Stag} = \begin{pmatrix}|u| & c\\ -c & |u|\end{pmatrix}. \tag{13}$$

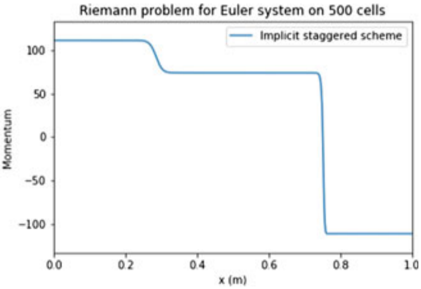
Since D_{Stag} verifies the property of Theorem 3, this numerical scheme is linearly L^2-stable.

In the next section, we implement an implicit version of this new numerical scheme.

4 Numerical Results and Conclusion

We assess the robustness of our new staggered scheme on a compressible fluid with isothermal equation of state $p = \rho c^2$ where the sound speed is $c = 300\,\text{m/s}$. We consider a Riemann problem for the isentropic Euler system (10) with left state ($\rho_L = \frac{10}{9}, q_L = 100\rho_L$) and right state ($\rho_R = \frac{\rho_L}{2}, q_R = -100\rho_L$). The solution displays a rarefaction (smooth) wave followed by a (discontinuous) shock wave. Our new method is able to capture both waves in a distinct and stable way. In a forthcoming longer paper we will present the details of the 2D version of the scheme and study its entropic character.





References

1. Berthelin, F., Goudon, T., Minjeaud, S.: Kinetic schemes on staggered grids for barotropic euler models: entropy-stability analysis. Math. Comput. **84**(295), 2221–2262 (2015)
2. Dellacherie, S.: Analysis of godunov type schemes applied to the compressible euler system at low mach number. J. Comp. Phys. **229**(4), 978–1016 (2010)
3. Dellacherie, S., Omnes, P., Rieper, F.: Analysis of godunov type schemes applied to the compressible euler system at low mach number. J. Comp. Phys. **229**(14), 5315–5338 (2010)
4. Ghidaglia, J., Kumbaro, A., Coq, G.L.: Une méthode volumes finis à flux caractéristiques pour la résolution numérique des systèmes hyperboliques de lois de conservation. Comptes Rendus de l'Acad. Sciences Paris, Série 1, vol. 322, pp. 981–988 (1996)
5. Godlewski, E., Raviart, P.A.: Numerical approximation of hyperbolic systems of conservation laws. In: Applied Mathematical Sciences, vol. 118. Springer, New York (1996)
6. Godunov, S.K.: A difference scheme for numerical solution of discontinuous solution of hydrodynamic equations. Mat. Sbornik. **47**, 271–306 (1959)
7. Harlow, F., Amsden, A.: Numerical calculation of almost incompressible flow. Journal of Computational Physics **3**, 80–93 (1968)
8. Harlow, F., Amsden, A.: A numerical fluid dynamics calculation method for all flow speeds. J. Comput. Phys. **8**, 197–213 (1971)
9. Herbin, R., Kheriji, W., Latché, J.C.: Staggered schemes for all speed flows. In: ESAIM: Proceedings, EDP Sciences, Congrès National de Mathématiques Appliquées et Industrielles, vol. 35, pp. 122–150 (2011)
10. Herbin, R., Kheriji, W., Latché, J.C.: On some implicit and semi-implicit staggered schemes for the shallow water and Euler equations. ESAIM: Math. Model. Numer. Anal. EDP Sciences, **48**(6), 1807–1857 (2014)

11. Herbin, R., Latché, J.C.: A kinetic energy preserving convection operator for the mac discretization of compressible Navier–Stokes equations. In: Mathematical Modelling and Numerical Analysis (2010). https://hal.archives-ouvertes.fr/hal-00477079/document
12. Hirt, C.W.: Heuristic stability theory for finite difference equations. J. Comp. Phys. **2**, 339–355 (1968)
13. Masella, J.M., Faille, I., Gallouët, T.: On an approximate godunov scheme. Int. J. Comput. Fluid Dyn. **12**, 133–149 (1999)
14. Prosperetti, A., Tryggvason, G.: Computational methods for multiphase flow. Cambridge University Press, Cambridge (2009)
15. Roe, P.L.: Approximate riemann solvers, parameter vectors and difference schemes. J. Comput. Phys. **43**, 357 (1981)

Convergence of a TPFA Finite Volume Scheme for Mixed-Dimensional Flow Problems

Wietse M. Boon and Jan M. Nordbotten

Abstract A two-point flux approximation (TPFA) finite volume method is considered for mixed-dimensional fracture flow problems. Its construction is based on applying a face-based quadrature rule to a conforming, mixed finite element scheme of lowest order. A concise argument shows linear convergence in theory, which we confirm in practice by a numerical experiment.

Keywords Mixed-dimensional · Fracture flow · Mixed finite element · Two-point flux approximation

MSC (2010) 65N08 · 65N12 · 65N30

1 Introduction

Problems involving geometric features with high aspect ratios, such as (thin) aquifers [5], fractures [10], or faults [7], are often favorably modeled with the feature being of a topologically lower dimension than that of the ambient domain. The case of fracture networks is particularly appealing in this context, since the intersection of fractures (and the intersection of intersections) can be treated consistently in a recursive manner [3]. In recent work, the authors have exploited this perspective to derive and analyze mixed-finite element (MFE) methods for flow in fractured porous media.

As with many low-order mixed finite element methods, the method discussed above allows for the derivation of a finite volume variant. Having this possibility is a noteworthy advantage since finite volume methods may be preferred in certain

W. M. Boon (✉)
KTH Royal Institute of Technology, Lindstedtsvägen 25, 11428 Stockholm, Sweden
e-mail: Wietse@kth.se

J. M. Nordbotten
University of Bergen, Postboks 7803, 5020 Bergen, Norway
e-mail: Jan.Nordbotten@uib.no

R. Klöfkorn et al. (eds.), *Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples*, Springer Proceedings in Mathematics & Statistics 323,
https://doi.org/10.1007/978-3-030-43651-3_40

applications for several reasons. First, due to the fewer degrees of freedom and resulting definite system, finite volume methods are still considered computationally favorable in reservoir simulation practice [6]. Furthermore, the finite volume structure allows for the incorporation of upstream weighting, which is important in order to capture accurately flow and transport phenomena for coupled problems (see e.g. [9]).

There are two main ways of deriving finite volume methods from mixed-finite element methods. The most straight-forward approach, which we will follow in this paper, is to consider the Raviart–Thomas ($\mathbb{RT}_0$) spaces for flux, and use a face-based quadrature to obtain a two-point flux approximation (TPFA) scheme. This construction is originally due to Russell and Wheeler [11], and was further refined by Baranger et al. [1]. In the context of fractured porous media, it leads to a method structurally similar to the method introduced by Karimi-Fard et al. [8]. A more elaborate approach, outside the scope of this paper, is to base the construction on Brezzi–Douglas–Marini ($\mathbb{BDM}_1$) elements for flow and apply corner-based quadrature to reduce the scheme to a multi-point flux approximation (MPFA) type finite volume method (see e.g. [13]). This construction is advantageous for problems with anisotropy in the permeability coefficients, as well as for more complex grids (e.g. quadrilateral grids). In the context of fractured porous media, this approach leads to a method similar to the MPFA method presented by Sandve et al. [12].

In this paper, we thus present the TPFA hybridization of $\mathbb{RT}_0$-based mixed-finite element methods for mixed-dimensional flow problems. This has several implications. Firstly, it directly establishes a finite volume scheme in this setting. Secondly, it shows the close connection between mixed-dimensional flow models and the equi-dimensional model discretized by Karimi-Fard et al. [8], and thus also giving a notion of convergence for that discretization. Thirdly, we note that the TPFA-type discretizations can be advantageous as preconditioners for the MFE discretzations [4].

2 Mixed-Dimensional Flow Model

In this section, we present the fracture flow model and introduce the notational conventions this work adheres to. We start by defining the mixed-dimensional geometry obtained after dimensional reduction of all fractures and intersections. Afterwards, the model equations are described, both in their strong form and the corresponding variational formulation.

2.1 Geometry and Notation

We follow the notation introduced in [3]. Consider a bounded Lipschitz domain $Y \subset \mathbb{R}^n$ with $n \in \{2, 3\}$. Each fracture included in Y is represented by a $(n-1)$-dimensional manifold Ω_i with i its index from the set I^{n-1}. For simplicity, we assume that all fractures have zero curvature. We identify the intersection of multiple fractures as a $(n-2)$-manifold Ω_i with index $i \in I^{n-2}$. In the case that $n = 3$, the same reasoning is applied once more at the intersection point of intersection lines, and we introduce a 0-manifold Ω_i there with $i \in I^0$. The remaining regions of the domain Y corresponding to the bulk (or matrix) of the medium, are considered n-manifolds in their own right and we refer to these as Ω_i with $i \in I^n$.

The set of indexes encompassing all manifolds is then defined as $I := \bigcup_{d=0}^{n} I^d$ and we let d_i denote the dimension of manifold Ω_i. At times, we employ binary relations in the superscript of index sets to denote certain subsets. For example, let $I^{d>0} := \bigcup_{d=1}^{n} I^d$ and $I^{d<n} := \bigcup_{d=0}^{n-1} I^d$.

We have an additional interest in the boundaries of manifolds that coincide with lower-dimensional neighbors because these form the location at which we will impose coupling conditions. For that purpose, we use the notation Γ_j for an interface of dimension d_j that satisfies the following two properties. First, an index $\check{j} \in I^{d_j}$ exists such that Γ_j coincides physically with $\Omega_{\check{j}}$. Secondly, an index $\hat{j} \in I^{d_j+1}$ exists such that $\Gamma_j \subseteq \partial\Omega_{\hat{j}}$. Note the use of the hat and check notation to refer to higher- and respectively lower-dimensional neighbors. The set of all interface indexes is denoted by J.

Fracture tips are then characterized for each Ω_i, with $d_i < n$, as the part of its boundary that does not coincide with a lower-dimensional neighbor or the boundary of the domain Y. Since we impose no-flux conditions at these tips, we refer to this part of the boundary as $\partial_u\Omega_i$.

Next, we impose the key restriction that the collection of manifolds forms a disjoint decomposition of the original domain Y in the sense that

$$Y = \bigcup_{i \in I} (\Omega_i \cup \partial_u\Omega_i). \tag{1}$$

In order to keep track of the interfaces Γ_j that are relevant for a manifold Ω_i, we introduce the following index sets:

$$\hat{J}_i := \{j \in J \,|\, \check{j} = i\}, \qquad \check{J}_i := \{j \in J \,|\, \hat{j} = i\}. \tag{2}$$

In short, $\hat{J}_i$ and $\check{J}_i$ contain the indexes of interfaces that connect Ω_i to its higher- and respectively lower-dimensional neighbors with codimension one. It follows naturally that $\check{J}_i = \emptyset$ for $i \in I^0$ and $\hat{J}_i = \emptyset$ for $i \in I^n$.

We impose a final restriction on the geometry, namely that $\hat{J}_i$ is non-empty for all $i \in I^{d<n}$. In other words, all lower-dimensional manifolds are located at the boundary of a higher-dimensional neighbor of codimension one.

2.2 Model Equations

We introduce the fracture flow problem in three steps. Starting with the strong form of the equations, we then introduce the relevant function spaces, and conclude the section with the corresponding variational formulation.

For each manifold Ω_i with $i \in I$, let p_i denote the scalar pressure. Moreover, for $i \in I^{d>0}$, let $\mathbf{u}_i$ denote the tangential velocity as a d_i-vector, obtained after integration over the cross-section of the thin inclusion. For a precise presentation of the dimensional reduction that leads to the definition of this variable, we refer the reader to [3].

The model problem is then given by the following three equations, describing Darcy's law tangential to each manifold, normal to each interface, and mass conservation, respectively.

$$\mathbf{u}_i + K_\| \nabla_i p_i = 0, \qquad \text{in } \Omega_i,\ i \in I^{d>0}, \tag{3a}$$

$$\mathbf{n}_j \cdot \mathbf{u}_{\hat{j}} + K_\perp (p_{\check{j}} - p_{\hat{j}}) = 0, \qquad \text{on } \Gamma_j,\ j \in J, \tag{3b}$$

$$\nabla_i \cdot \mathbf{u}_i - \sum_{j \in \hat{J}_i} \mathbf{n}_j \cdot \mathbf{u}_{\hat{j}} = f_i, \qquad \text{on } \Omega_i,\ i \in I. \tag{3c}$$

Here, $\mathbf{n}_j$ is the unit vector normal to Γ_j and oriented outward with respect to $\Omega_{\hat{j}}$. Moreover, ∇_i is the del operator on Ω_i given by its tangential bundle. It follows by definition that the first term in the final equation is zero for $i \in I^0$. Moreover, we note that $\hat{J}_i = \emptyset$ for $i \in I^n$, and hence the second term in this equation is zero for the top-dimensional subdomains.

Additionally, $K_\|$ is the effective tangential permeability given by a $d_i \times d_i$ symmetric, positive definite tensor and $K_\perp$ is the effective, normal permeability given by a positive scalar on each interface Γ_j. We remark that all scalings with apertures and other small cross-sectional measures are thus incorporated in these effective permeabilities.

We set homogeneous pressure boundary conditions on the boundary ∂Y. Together with the no-flux condition at fracture tips, we thus close the system with

$$p_i = 0, \qquad \text{on } \partial Y, \tag{3d}$$

$$\mathbf{n} \cdot \mathbf{u}_i = 0, \qquad \text{on } \partial_u \Omega_i, \qquad \forall i \in I^{d<n}. \tag{3e}$$

Next, we present the function spaces that we use to pose the variational formulation of (3). It is convenient to collect the variables into the mixed-dimensional functions $\mathfrak{u} \in \mathfrak{U}$ and $\mathfrak{p} \in \mathfrak{P}$ defined as

$$\mathfrak{u}|_{\Omega_i} := \mathbf{u}_i, \qquad \mathfrak{p}|_{\Omega_i} := p_i,$$

with the corresponding function spaces given by

$$\mathfrak{U} := \prod_{i \in I^{d>0}} \{\mathbf{u}_i \in H(div, \Omega_i) |\, \mathbf{n}_j \cdot \mathbf{u}_i \in L^2(\Gamma_j),\ \forall j \in \check{J}_i\}, \quad \mathfrak{P} := \prod_{i \in I} L^2(\Omega_i).$$

Here, $H(div, \Omega_i)$ denotes the space of square integrable vector fields with square integrable divergence and we thus consider its subspace with well-defined normal trace on Γ. The two spaces are related by the mixed-dimensional divergence operator $\mathfrak{D}\cdot : \mathfrak{U} \to \mathfrak{P}$, which is defined such that

$$(\mathfrak{D} \cdot \mathfrak{u})|_{\Omega_i} := \nabla_i \cdot \mathbf{u}_i - \sum_{j \in \hat{J}_i} \mathbf{n}_j \cdot \mathbf{u}_{\hat{j}}, \qquad \forall i \in I.$$

Finally, we have all the ingredients to state the variational formulation of (3): Find $(\mathfrak{u}, \mathfrak{p}) \in \mathfrak{U} \times \mathfrak{P}$ such that

$$(K_{\parallel}^{-1} \mathfrak{u}, \tilde{\mathfrak{u}})_{\Omega} + (K_{\perp}^{-1} \mathbf{n} \cdot \mathfrak{u}, \mathbf{n} \cdot \tilde{\mathfrak{u}})_{\Gamma} - (\mathfrak{D} \cdot \tilde{\mathfrak{u}}, \mathfrak{p})_{\Omega} = 0, \qquad \forall \tilde{\mathfrak{u}} \in \mathfrak{U}, \tag{4a}$$

$$(\mathfrak{D} \cdot \mathfrak{u}, \tilde{\mathfrak{p}})_{\Omega} = (\mathfrak{f}, \tilde{\mathfrak{p}})_{\Omega}, \qquad \forall \tilde{\mathfrak{p}} \in \mathfrak{P}. \tag{4b}$$

Here, the inner products are naturally defined as the sum over all manifolds Ω_i or interfaces Γ_j. Moreover, $\mathbf{n} \cdot \mathfrak{u}$ represents the normal trace of the Darcy velocity $\mathbf{u}_{\hat{j}}$ on each Γ_j with $j \in J$. Finally, $\mathfrak{f}$ is the source function defined such that $\mathfrak{f}|_{\Omega_i} = f_i$.

As observed in [3], this system has a typical saddle point structure, and we identify the bilinear forms $a : \mathfrak{U} \times \mathfrak{U} \to \mathbb{R}$ and $b : \mathfrak{U} \times \mathfrak{P} \to \mathbb{R}$ as

$$a(\mathfrak{u}, \tilde{\mathfrak{u}}) := (K_{\parallel}^{-1} \mathfrak{u}, \tilde{\mathfrak{u}})_{\Omega} + (K_{\perp}^{-1} \mathbf{n} \cdot \mathfrak{u}, \mathbf{n} \cdot \tilde{\mathfrak{u}})_{\Gamma}, \qquad b(\mathfrak{u}, \mathfrak{p}) := (\mathfrak{D} \cdot \mathfrak{u}, \mathfrak{p})_{\Omega}. \tag{5}$$

3 Discretization

In this section, we introduce the discretization of the model problem. Starting with the generation of the mesh, we continue with a description of a low-order mixed finite element method. Through hybridization, we then obtain the desired finite volume method.

For each $i \in I$ let $\Omega_{h,i}$ be a regular tesselation of Ω_i consisting of d_i-simplices. We assume that all grids are matching in the sense that each interface mesh $\Gamma_{h,j}$ is defined as the trace mesh of $\Omega_{h,\hat{j}}$. Moreover, each element e in $\Gamma_{h,j}$ is physically colocated with a unique element $\check{e}$ of the lower-dimensional neighbor $\Omega_{h,\check{j}}$. Finally, to ensure consistency of the TPFA scheme, we only consider K-orthogonal grids.

3.1 Mixed Finite Element Method

With the mesh defined, we now introduce the stable finite element pair of lowest order that forms a discrete subspace of $\mathfrak{U} \times \mathfrak{P}$:

$$\mathfrak{U}_h := \prod_{i \in I^{d>0}} \mathbb{RT}_0(\Omega_{h,i}), \qquad \mathfrak{P}_h := \prod_{i \in I} \mathbb{P}_0(\Omega_{h,i}), \tag{6}$$

In other words, the flux space is given by the Raviart–Thomas(–Nedelec) elements for $d_i \in \{2, 3\}$ and the linear Lagrange element for $d_i = 1$. Hence, this space has one degree of freedom per $(d_i - 1)$-dimensional face in the meshes with $d_i > 0$. The pressure space is defined as the piecewise constants and therefore has one degree of freedom per element in all dimensions.

Theorem 1 *A constant $C > 0$ exists such that the mixed finite element solution $(\mathfrak{u}_h, \mathfrak{p}_h)$ satisfies*

$$\|\mathfrak{u} - \mathfrak{u}_h\|_\Omega + \|\mathfrak{p} - \mathfrak{p}_h\|_\Omega \leq Ch \left(\|\mathfrak{u}\|_{H^1(\Omega)} + \|\mathbf{n} \cdot \mathfrak{u}\|_{H^1(\Gamma)} + \|\mathfrak{p}\|_{H^1(\Omega)} \right) \tag{7}$$

Proof See [3], Theorem 3.4. □

3.2 Finite Volume Method

In this section, we consider the technique described by [1, 2] to construct a TPFA finite volume scheme. During this procedure, the bilinear form a from (5) is replaced by a simpler form a_L. Its diagonal structure then allows us to eliminate the flux variables and obtain a TPFA finite volume scheme.

We first introduce a few necessary definitions. For each element $e \in \Omega_{h,i}$, let $\mathbf{x}_e$ be its circumcenter. In other words, $\mathbf{x}_e$ is the center of the unique d_i-sphere that passes through its vertices. We limit our exposition to the case in which $\mathbf{x}_e$ lies in the interior of e. This is the case if, for example, the mesh $\Omega_{h,i}$ consists of simplices with acute angles. Moreover, to ensure consistency of TPFA, we herein consider the simpler case in which $K_\|$ is isotropic, i.e. a positive scalar field on each Ω_i with $i \in I^{d>0}$.

Given an element element e_1 in the mesh $\Omega_{h,i}$ with $i \in I^{d>0}$, let σ be one of its faces. We remark that σ is a $(d_i - 1)$-simplex, by construction. Using this face as a base, we construct a d_i-simplex by connecting it with the circumcenter $\mathbf{x}_{e_1}$. If applicable, the procedure is repeated for the second element e_2 in the mesh that has σ as a face. The patch L_σ is then defined as the union of these constructed sub-simplices bordering on σ.

Geometrically, we observe that the measure of the patch L_σ is given by

$$|L_\sigma| = \frac{1}{d_i}|\sigma||l_\sigma|, \tag{8}$$

with l_σ defined as the distance between the two adjacent circumcenters for internal faces. For boundary faces, l_σ is the distance between σ and the circumcenter $\mathbf{x}_{e_1}$.

Next, we use the disjoint patches L_σ to construct the following, piecewise constant function space $\mathfrak{U}_L$:

$$\mathfrak{U}_L := \prod_{i \in I^{d>0}} \{\mathbf{v}_L \in L^2(\Omega_i)^{d_i} : \mathbf{v}_L|_{L_\sigma} = c_\sigma \mathbf{n}_\sigma \text{ with } c_\sigma \in \mathbb{R},\ \forall \sigma \in \mathscr{F}_i\}. \tag{9}$$

Here, $\mathbf{n}_\sigma$ denotes the unique normal vector associated with face σ and $\mathscr{F}_i$ is the set of all faces of $\Omega_{h,i}$. The associated projection operator $\Pi_L : \mathfrak{U}_h \to \mathfrak{U}_L$ is defined as

$$(\Pi_L \mathbf{v}_h)|_{L_\sigma} = (\mathbf{v}_h \cdot \mathbf{n}_\sigma)|_\sigma\ \mathbf{n}_\sigma, \qquad\qquad \forall \sigma \in \bigcup_{i \in I^{d>0}} \mathscr{F}_i. \tag{10}$$

We are now ready to define the bilinear form $a_L : \mathfrak{U}_h \times \mathfrak{U}_h \to \mathbb{R}$ as

$$a_L(\mathfrak{u}_h, \tilde{\mathfrak{u}}_h) := \sum_{i \in I^{d>0}} \left((d_i K_\parallel^{-1} \Pi_L \mathbf{u}_{h,i}, \Pi_L \tilde{\mathbf{u}}_{h,i})_{\Omega_i} + \sum_{j \in \check{J}_i} \langle K_\perp^{-1} \mathbf{n} \cdot \mathbf{u}_{h,i}, \mathbf{n} \cdot \tilde{\mathbf{u}}_{h,i} \rangle_{\Gamma_j} \right). \tag{11}$$

Note the scaling with the dimension d_i in the first term. We elaborate the reason for this in Remark 1 at the end of the section.

Due to the structure of the bilinear form a_L, we can now eliminate the flux variable to obtain a linear system in pressure only. In particular, we invert the corresponding, diagonal matrix A_L to obtain the system

$$\mathsf{B}^\mathsf{T} \mathsf{A}_L^{-1} \mathsf{B} \mathsf{p}_h = \mathsf{f}_h. \tag{12}$$

Here, B is the matrix associated with the bilinear form b and p_h and f_h are the vector representations of $\mathfrak{p}_h$ and $\mathfrak{f}$. After solving for the pressure $\mathfrak{p}_h$, we may then reconstruct $\mathfrak{u}_h$ by setting $\mathsf{u}_h := \mathsf{A}_L^{-1} \mathsf{B} \mathsf{p}_h$.

Theorem 2 *A constant $C > 0$ exists such that the finite volume solution $(\mathfrak{u}_h, \mathfrak{p}_h)$ satisfies*

$$\|\mathfrak{u} - \mathfrak{u}_h\|_\Omega + \|\mathfrak{p} - \mathfrak{p}_h\|_\Omega \le Ch \left(\|\mathfrak{u}\|_{H^1(\Omega)} + \|\mathbf{n} \cdot \mathfrak{u}\|_{H^1(\Gamma)} + \|\mathfrak{p}\|_{H^1(\Omega)} \right). \tag{13}$$

Moreover, the system in (12) is symmetric and positive definite.

Proof In [1], it is shown that the replacement of a with a_L introduces a consistency error of the same order as the approximation error in $\mathbb{RT}_0$. Hence, a triangle inequality on each d-dimensional mesh with (7) from Theorem 1 provides (13).

Symmetry of $\mathsf{B}^\mathsf{T} \mathsf{A}_L^{-1} \mathsf{B}$ is apparent. Due to the inf-sup condition on b (see [3]), it follows that $\mathsf{B} \mathsf{p}_h = 0$ implies $\mathsf{p}_h = 0$. In turn, the system is positive definite. □

Remark 1 We conclude this section by showing how the use of a_L leads to the two-point flux approximation (TPFA) of the tangential Darcy's law. Let the mesh $\Omega_{h,i}$ be given with $i \in I^{d>0}$ and let $(\tilde{\mathbf{u}}_h, p_h) \in \mathbb{RT}_0(\Omega_{h,i}) \times \mathbb{P}_0(\Omega_{h,i})$. Starting with the weak formulation, we use the divergence theorem and (8) to derive:

$$\begin{aligned}
(d_i K_{\|}^{-1} \Pi_L \mathbf{u}_h, \Pi_L \tilde{\mathbf{u}}_h)_{\Omega_i} &= (\nabla \cdot \tilde{\mathbf{u}}_h, p_h)_{\Omega_i} \\
&= \sum_{e \in \Omega_{h,i}} \langle p_h^e, \tilde{\mathbf{u}}_h \cdot \mathbf{n}_{e,\sigma} \rangle_{\partial e} \\
&= \sum_{e \in \Omega_{h,i}} \langle p_h^e, \Pi_L \tilde{\mathbf{u}}_h \cdot \mathbf{n}_{e,\sigma} \rangle_{\partial e} \\
&= \sum_{\sigma \in \mathscr{F}_i} \langle p_h^{e_1} \mathbf{n}_{e_1,\sigma} + p_h^{e_2} \mathbf{n}_{e_2,\sigma}, \Pi_L \tilde{\mathbf{u}}_h \rangle_{\sigma} \\
&= \sum_{\sigma \in \mathscr{F}_i} \frac{d_i}{|l_\sigma|} (p_h^{e_1} \mathbf{n}_{e_1,\sigma} + p_h^{e_2} \mathbf{n}_{e_2,\sigma}, \Pi_L \tilde{\mathbf{u}}_h)_{L_\sigma}.
\end{aligned}$$

Here, p_h^e is the evaluation of p_h in element e and $p_h^{e_2}$ is zero if the face σ is on the boundary. Moreover, $\mathbf{n}_{e_1,\sigma}$ denotes the unit vector that is normal to σ, oriented outward with respect to element e_1.

It follows that using a_L from (11) is equivalent to imposing the TPFA stencil:

$$\Pi_L \mathbf{u}_h = K_{\|} \frac{p_h^{e_1} \mathbf{n}_{e_1,\sigma} + p_h^{e_2} \mathbf{n}_{e_2,\sigma}}{|l_\sigma|}, \qquad \text{on } L_\sigma. \tag{14}$$

4 Numerical Experiment

To confirm the convergence of both numerical schemes, we show the numerical results using test cases designed to highlight some of the typical challenges associated with fracture flow simulation. First, we will introduce the set-up and describe the chosen parameters therein are discussed afterwards, followed by an evaluation of the results.

Let the domain Y be the unit square and the fractures given as depicted in Fig. 1 (left). A pressure drop is simulated by imposing unit pressure at the top and zero pressure at the bottom boundary. On the remaining sides, a no-flow boundary condition is imposed. For simplicity, the source function f is set to zero.

Let us continue by defining the parameters for the test cases. We set $K_{\|}$ to the 2×2 identity tensor in the bulk and we set $K_{\|} = 10^2$ in the fractures. We investigate two cases by varying the effective normal permeability. Flow into the fractures is stimulated in Case 1 by setting $K_{\perp}$ to 10^2. On the other hand, we set $K_{\perp}$ to 1 in Case 2 to capture the influence of blocking features and resulting pressure discontinuities. Finally, the normal permeabilities at the intersection are set to one in both cases, for simplicity.

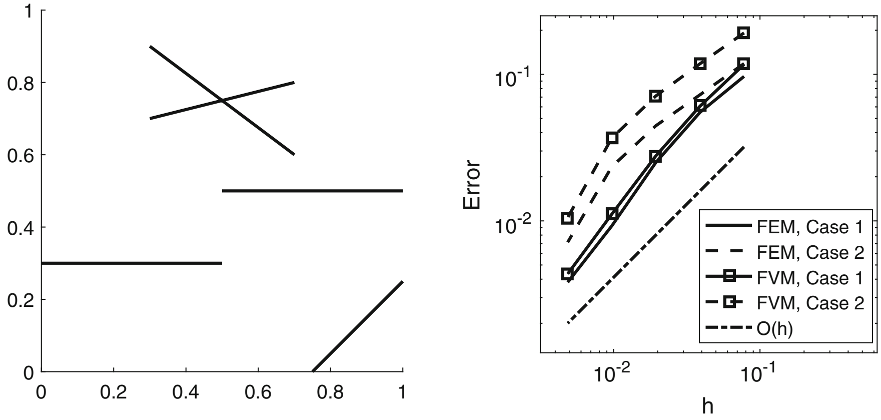


Fig. 1 (Left) The square domain contains an intersection and multiple fracture endings. (Right) The error in the energy norm decreases linearly with the mesh size (h) for the mixed finite element method (FEM) and the TPFA finite volume scheme (FVM) in both test cases

The numerical experiments were performed on six consecutively refined grids. All solutions were then compared to the solution on the finest grid using the norms from Theorems 1 and 2. The results for each case are shown in Fig. 1 (right). As expected, the results show that both methods converge linearly for both cases, with the TPFA variant suffering a modest loss of accuracy.

References

1. Baranger, J., Maitre, J.F., Oudin, F.: Connection between finite volume and mixed finite element methods. ESAIM: Math. Model. Numer. Anal. **30**(4), 445–465 (1996)
2. Boffi, D., Fortin, M., Brezzi, F.: Mixed Finite Element Methods and Applications. Springer Series in Computational Mathematics. Springer, Berlin, Heidelberg (2013)
3. Boon, W.M., Nordbotten, J.M., Yotov, I.: Robust discretization of flow in fractured porous media. SIAM J. Numer. Anal. **56**(4), 2203–2233 (2018)
4. Budiša, A., Hu, X.: Block preconditioners for mixed-dimensional discretization of flow in fractured porous media. arXiv:1905.13513 (2019)
5. Freeze, R., Cherry, J.: Groundwater. 0-13-365312-9. Prentice-Hall, Upper Saddle River (1979)
6. GeoQuest Schlumberger: Eclipse Reference Manual. Schlumberger, Houston, TX (2014)
7. Heimisson, E.R., Dunham, E.M., Almquist, M.: Poroelastic effects destabilize mildly rate-strengthening friction to generate stable slow slip pulses. J. Mech. Phys. Solids **130**, 262–279 (2019)
8. Karimi-Fard, M., Durlofsky, L.J., Aziz, K., et al.: An efficient discrete fracture model applicable for general purpose reservoir simulators. In: SPE Reservoir Simulation Symposium. Society of Petroleum Engineers, Houston (2003)
9. LeVeque, R.J.: Numerical Methods for Conservation Laws, vol. 132. Springer, Berlin (1992)
10. Martin, V., Jaffré, J., Roberts, J.E.: Modeling fractures and barriers as interfaces for flow in porous media. SIAM J. Sci. Comput. **26**(5), 1667–1691 (2005)

11. Russell, T.F., Wheeler, M.F.: Finite element and finite difference methods for continuous flows in porous media. In: The Mathematics of Reservoir Simulation, pp. 35–106. SIAM, Philadelphia (1983)
12. Sandve, T.H., Berre, I., Nordbotten, J.M.: An efficient multi-point flux approximation method for discrete fracture-matrix simulations. J. Comput. Phys. **231**(9), 3784–3800 (2012)
13. Wheeler, M.F., Xue, G., Yotov, I.: A family of multipoint flux mixed finite element methods for elliptic problems on general grids. Procedia Comput. Sci. **4**, 918–927 (2011)

A Relaxation Method for the Simulation of Possibly Non-hyperbolic Polymer Flooding Models with Inaccessible Pore Volume Effect

Guissel Lagnol Dongmo Nguepi, Benjamin Braconnier, Christophe Preux, Quang-Huy Tran, and Christophe Berthon

Abstract Polymer flooding models used in the simulation of enhanced oil recovery of reservoirs commonly involve a system of conservation laws that may be ill-posed, especially when an inaccessible pore volume (IPV) empirical law is considered. Depending on the IPV law, the flow model is either weakly hyperbolic with resonance or non-hyperbolic with complex eigenvalues. In this paper, we propose a Suliciu-type relaxation, which unconditionally ensures hyperbolicity for any IPV law. This approximation gives rise to a new numerical scheme, which is compared with the classical upwind scheme and the exact solution whenever possible.

Keywords Polymer flooding · Exclusion volume · Relaxation method · Weakly hyperbolic and non-hyperbolic systems

MSC (2010) 65M08 · 65N08 · 37D50

1 Polymer Flooding Models with IPV Effect

We are interested in the simulation of enhanced oil recovery (EOR) using polymers. Polymers are injected in the oil field to increase the water viscosity, thus reducing water-oil interfaces instabilities and improving oil recovery. To highlight the difficulties associated with this problem, let us consider a simplified model describing the incompressible flow of a water-polymer mixture and oil in a 1-D porous medium. This simplified model reads [9]

G. L. Dongmo Nguepi (✉) · B. Braconnier · C. Preux · Q.-H. Tran
IFP Energies nouvelles, 1 et 4 avenue de Bois-Préau, 92852
Rueil-Malmaison Cedex, France
e-mail: guissel-lagnol.dongmo-nguepi@ifpen.fr

C. Berthon
University of Nantes, Laboratoire de Mathématiques Jean Leray,
UMR 6629, Département de Mathématiques, 2 rue de la Houssinière, BP 92208,
44322 Nantes Cedex 3, France
e-mail: christophe.berthon@univ-nantes.fr

R. Klöfkorn et al. (eds.), *Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples*, Springer Proceedings in Mathematics & Statistics 323,
https://doi.org/10.1007/978-3-030-43651-3_41

$$\partial_t(s) \quad + \partial_x(f) \quad = 0, \tag{1a}$$

$$\partial_t(sc) + \partial_x(fc\gamma) = 0, \tag{1b}$$

where $s(x, t) \in [0, 1]$ denotes the water saturation and $c(x, t) \in [0, 1]$ the polymer mass fraction. In (1), polymer molecules are assumed to be in the water phase only and to not adsorb on the rock. The water fractional flux f is a function of s and c, given by

$$f(s, c) = \frac{\Lambda_w(s)/R_m(c)}{\Lambda_w(s)/R_m(c) + \Lambda_o(1 - s)}, \tag{2}$$

in which the smooth functions $\Lambda_w(s)$ and $\Lambda_o(1 - s)$ represent the (known) mobilities of the water and oil phases. $R_m(c)$ is the mobility reduction factor of the water phase. In real applications, it depends on the polymer adsorption and on the water shear thinning. Here, we consider the analytical formula

$$R_m(c) = 1 + a_1 c + a_2 c^2 + a_4 c^{15/4}, \qquad a_{1,2,4} \geq 0, \tag{3}$$

due to de Gennes [5]. When $R_m \equiv 1$, the fractional flux f does not depend on c and boils down to the Buckley-Leverett one. In the general case, f is an S-shaped function of s at fixed c and a non-increasing function of c at fixed s.

The acceleration factor γ expresses the inaccessible pore volume (IPV) effect and has to be supplied as a function of (s, c). Once an IPV law $\gamma(s, c)$ is selected, system (1) becomes algebraically closed and can be put under the abstract form

$$\partial_t \mathbf{w} + \partial_x \mathbf{f}(\mathbf{w}) = 0, \tag{4}$$

with $\mathbf{w} = (s, sc)^T$ and $\mathbf{f} = (f, fc\gamma)^T$. The question then arises as to the hyperbolicity of (4). In other words, it is essential to know if the Jacobian matrix $\nabla_{\mathbf{w}}\mathbf{f}(\mathbf{w})$ has real eigenvalues corresponding to wave propagation speeds and has a basis of real eigenvectors. The answer depends on the IPV law under consideration. Below are the three most well-known cases.

- No IPV, i.e.,

$$\gamma(s, c) \equiv 1. \tag{5}$$

  Then, model (1) coincides with the Keyfitz–Kranzer system [8], which is weakly hyperbolic in the following sense: the eigenvalues are always real, equal to $\lambda_c = f/s =: u$ (c-wave) and $\lambda_s = \partial_s f = u + s\partial_s u$ (s-wave); but on the resonance curve $\Gamma = \{(s, c) \mid \partial_s u = 0\}$ where the eigenvalues $\lambda_c = \lambda_s$ collapse to each other, the basis of eigenvectors is lost. For some initial data, the Riemann problem may have several solutions. Isaacson and Temple [7] advocated an additional entropy condition to recover uniqueness.
- Percolation IPV, i.e.,

$$\gamma(s, c) = \frac{s}{s - s_\bullet}, \qquad s_\bullet \in (0, 1), \tag{6}$$

was proposed in [1] to keep the system hyperbolic for $s \in [s_\flat, 1]$ where $s_\flat \in (s_\bullet, 1]$ is the irreducible water saturation. Then, system (1) exhibits a similar weakly hyperbolic behavior: the eigenvalues are always real, equal to $\lambda_c = \gamma u$ (c-wave) and $\lambda_s = u + s\partial_s u + (\gamma - 1)c\partial_c u$ (s-wave), where $u := f/s$; but on $\Gamma = \{(s, c) \mid s\partial_s(\gamma u) + (\gamma - 1)c\partial_c(\gamma u) = 0\}$ where they collapse to each other, the basis of eigenvectors is lost. Again, resorting to an additional entropy condition similar to [7], it is still possible to recover uniqueness for the solution of the Riemann problem.

- Constant IPV, i.e.,

$$\gamma(s, c) \equiv \gamma_* > 1. \tag{7}$$

Then, system (1) is not hyperbolic, in the sense that the eigenvalues may become complex. More accurately, this is shown [1] to occur on the manifold $\Gamma = \{(s, c) \mid \partial_s u = 0\}$ (the resonance curve for the no-IPV law) for $\gamma_* = 1 + \eta$ when $\eta > 0$ is small enough. Unfortunately, this IPV law is still widely used by reservoir engineers and implemented with the upwind scheme. For some initial data, they observe polymer accumulation at the saturation front, which gives rise to very high mass fractions and numerical instabilities.

Under relevant physical justifications, other IPV laws are derived in [6] to avoid possible δ-shock solutions when $s_\bullet$ is greater than the irreducible saturation. The latter IPV laws are not studied in this work.

The no IPV law (5) and the percolation IPV law (6) are equivalent, up to a change of variables. This result seems to have never been formally stated before.

Theorem 1 *Setting* $(\mathtt{S}, \mathtt{C}) = (s/\gamma, \gamma c) = (s - s_\bullet, \gamma c)$ *and* $\mathtt{F}(\mathtt{S}, \mathtt{C}) = f(s, c)$, *system (1) using the percolation IPV law (6) can be transformed into*

$$\partial_t(\mathtt{S}) + \partial_x(\mathtt{F}) = 0, \tag{8a}$$

$$\partial_t(\mathtt{SC}) + \partial_x(\mathtt{FC}) = 0, \tag{8b}$$

which is system (1) using the no IPV law (5).

Proof By (1a), we have $\partial_t \mathtt{S} = \partial_t(s - s_\bullet) = \partial_t s = -\partial_x f = -\partial_x \mathtt{F}$, hence (8a). From $\mathtt{SC} = (s/\gamma)(\gamma c) = sc$ and $\mathtt{FC} = fc\gamma$, we see that (8b) is none other than (1b). □

Because of the possibly non-hyperbolic nature of the original model (1), it is difficult to design a reliable and robust numerical scheme [4, 6]. Here, we propose a new relaxation scheme based on the Suliciu relaxation method [10]. Our motivation is to supply practitioners with a numerical method whose stability is always guaranteed by the very fact that it can be interpreted as the Godunov scheme applied to a hyperbolic approximation of the original model. This is important for engineers. In this respect, our approach bears some similarities with Baudin et al.'s work [2] where the hyperbolicity of the original model is not taken for granted.

2 A Relaxation Method Ensuring Hyperbolicity

The first step toward building a new numerical scheme is to rewrite (1) under a form that better separates transportation from acceleration, namely,

$$\begin{aligned} \partial_t(s) &+ \partial_x(su) &= 0, \qquad (9a)\\ \partial_t(sc) &+ \partial_x(scu+q) &= 0, \qquad (9b)\end{aligned}$$

where $q(s,c) = (\gamma-1)scu$ will be called the IPV deviation. In (9), which looks more like a hydrodynamics model, the genuine nonlinearities are encapsulated in the functions $u(s,c)$ and $q(s,c)$. In the second step, we attempt to circumvent these nonlinearities by approximating (9) by the relaxation model

$$\begin{aligned} \partial_t(s) &+ \partial_x(sU) &&= 0, &&(10a)\\ \partial_t(sc) &+ \partial_x(scU+Q) &&= 0, &&(10b)\\ \partial_t(sU) &+ \partial_x(sU^2 - a^2/s) &&= \varepsilon^{-1}s(u(s,c)-U), &&(10c)\\ \partial_t(sQ) &+ \partial_x(sQU + b^2c) &&= \varepsilon^{-1}s(q(s,c)-Q), &&(10d)\end{aligned}$$

where $\varepsilon > 0$ is the relaxation time, a and b are relaxation coefficients to be chosen later. The variables (U, Q) are relaxation counterparts of (u, q) and should be seen as independent of (s, c). System (10) can be abstractly reformulated as

$$\partial_t \mathbf{W} + \partial_x \mathbf{F}(\mathbf{W}) = \varepsilon^{-1} s \mathbf{R}, \qquad (11)$$

with $\mathbf{W} = (s, sc, sU, sQ)^T$, $\mathbf{F} = (sU, scU + Q, sU^2 - a^2/s, SQU + b^2c)^T$ and $\mathbf{R} = (0, 0, u - U, q - Q)^T$. The benefits of working with (11) are numerous.

Lemma 1 *For $s > 0$ and $a, b > 0$, the Jacobian matrix $\nabla_{\mathbf{W}}\mathbf{F}(\mathbf{W})$ has real eigenvalues with a basis of eigenvectors. The eigenfields, given by $\lambda_{U\pm} = U \pm a/s$ and $\lambda_{Q\pm} = U \pm b/s$, are all linearly degenerated and have Riemann invariants $\mathscr{I}_{U\pm} = \{c, Q, U \pm a/s\}$ and $\mathscr{I}_{Q\pm} = \{s, U, Q \pm bc\}$.*

Proof Calculations are easier to perform in the set of variables $\mathbf{V} = (s, c, U, Q)^T$ and with the quasi-linear form $\partial_t \mathbf{V} + \mathbf{J}\partial_x \mathbf{V} = \varepsilon^{-1}\mathbf{R}$, where the matrix $\mathbf{J}$ can be shown to have the four eigenvalues $\lambda_{U\pm}$ and $\lambda_{Q\pm}$. In this set of variables, the right eigenvectors are given by $\mathbf{r}_{U\pm} = \left(1, 0, \pm a/s^2, 0\right)^T$ and $\mathbf{r}_{Q\pm} = (0, 1, 0, \pm b)^T$, from which linear degeneracy and various sets of Riemann invariants can be deduced. □

Linear degeneracy and explicit knowledge of Riemann invariants make the Riemann problem easy to solve. Before giving the details of the numerical scheme, we have to make sure that (11) is a dissipative approximation to (4).

Theorem 2 *At the first order in $\varepsilon \downarrow 0$, the first two components $\mathbf{w} = (s, sc)$ of the relaxation solution $\mathbf{W}$ of (11) solve the equivalent equation*

$$\partial_t \mathbf{w} + \partial_x \mathbf{f}(\mathbf{w}) = \varepsilon \partial_x (s\mathbf{P}^{-1}\mathbf{D}\mathbf{P}\partial_x \mathbf{w}), \tag{12}$$

where

$$\mathbf{P} = \begin{bmatrix} 1 & 0 \\ -cs & s \end{bmatrix}, \quad \mathbf{D} = \frac{1}{s^3}\begin{bmatrix} a^2 - s^4(\partial_s u)^2 - s^2\partial_c u \partial_s q & -\partial_c u(s^2\partial_s u + \partial_c q) \\ -s^2\partial_s q(s^2\partial_s u + \partial_c q) & b^2 - s^2\partial_s q \partial_c u - (\partial_c q)^2 \end{bmatrix}.$$

Furthermore, it is possible to choose $a, b > 0$ *large enough such that* $\mathbf{D}$ *has positive eigenvalues, in which case the relaxation approximation is dissipative.*

Proof Inserting the Chapman-Enskog expansions $U = u + \varepsilon U_1$ and $Q = q + \varepsilon Q_1$ into the first two equations of (10) and moving the terms in ε to the right-hand side, we obtain the equivalent equation

$$\partial_t \mathbf{w} + \partial_x \mathbf{f}(\mathbf{w}) = \varepsilon \partial_x \left\{ \begin{bmatrix} s & 0 \\ sc & 1 \end{bmatrix} \begin{bmatrix} -U_1 \\ -Q_1 \end{bmatrix} \right\} = \varepsilon \partial_x \left\{ s\mathbf{P}^{-1} \begin{bmatrix} -U_1 \\ -Q_1 \end{bmatrix} \right\}. \tag{13}$$

Inserting the Chapman-Enskog expansions $U = u + \varepsilon U_1$ and $Q = q + \varepsilon Q_1$ into the last two equations of (10) yields the zeroth-order approximations $-sU_1 = \partial_t(su) + \partial_x(su^2 - a^2/s)$ and $-sQ_1 = \partial_t(sq) + \partial_x(squ + b^2 c)$. Using (9), we can express $\partial_t(su) + \partial_x(su^2)$ and $\partial_t(sq) + \partial_x(squ)$ as combinations of $\partial_x u$ and $\partial_x q$. Expanding these derivatives, we end up with

$$\begin{bmatrix} -U_1 \\ -Q_1 \end{bmatrix} = \widetilde{\mathbf{D}}\, \partial_x \begin{bmatrix} s \\ c \end{bmatrix} = \widetilde{\mathbf{D}} \begin{bmatrix} 1 & 0 \\ -c/s & 1/s \end{bmatrix} (\mathbf{P}^{-1}\mathbf{P})\, \partial_x \begin{bmatrix} s \\ sc \end{bmatrix} = \mathbf{D}\mathbf{P}\, \partial_x \mathbf{w}, \tag{14}$$

with

$$\widetilde{\mathbf{D}} = \frac{1}{s^3}\begin{bmatrix} a^2 - s^4(\partial_s u)^2 - s^2\partial_c u \partial_s q & -s^2\partial_c u(s^2\partial_s u + \partial_c q) \\ -s^2\partial_s q(s^2\partial_s u + \partial_c q) & s^2(b^2 - s^2\partial_s q \partial_c u - (\partial_c q)^2) \end{bmatrix}.$$

Combining (13) and (14) leads to the desired result. To prove that a, b can be chosen large enough so that $\mathbf{D}$ has positive eigenvalues, see Baudin et al. [2]. □

In view of Lemma 1 and Theorem 2, it seems that we run into trouble when $s = 0$. The issue of void has been addressed in other contexts by Bouchut [3]. Here, the same methodology will be applied: we consider that the parameters a and b are transported at the water velocity. For the sake of brevity, we have written our relaxation system (11) without this trick and do not provide more details here.

For convenience, the parameters a and b are also selected such that $a > b$ in order to enforce the ordering $\lambda_{U-} < \lambda_{Q-} < \lambda_{Q+} < \lambda_{U+}$ of eigenvalues. To go from time t^n to time $t^{n+1} = t^n + \Delta t$, we follow the two-stage procedure:

1. *Free evolution*: $\varepsilon = +\infty$. The relaxation model (10) is solved without the source terms, starting from the initial data $\mathbf{W}^n$ and for the time lapse equal to Δt. The solution obtained is designated by $\mathbf{W}^{n+1,-}$. At the discrete level in space, this is achieved by the Godunov scheme subject to an appropriate CFL condition.

2. *Projection to equilibrium*: $\varepsilon = 0$. We keep the first two components of $\mathbf{W}^{n+1,-}$, that is, $(s,\ sc)^{n+1} = (s,\ sc)^{n+1,-}$ but replace the last two components by their equilibrium values $(sU,\ sQ)^{n+1} = (s^{n+1}u(s^{n+1}, c^{n+1}),\ s^{n+1}q(s^{n+1}, c^{n+1}))$.

3 Numerical Validation

We consider a 1-D test configuration that represents water and polymer injection in an oil saturated plug. The domain has length $L = 1$ and is meshed with 1000 uniform cells, each of size $\Delta x = 0.001$. At the initial time, the plug is initialized at

$$(s,\ c)(x, t = 0) = \begin{cases} (1,\ 10^{-4}) & \text{if } x < 0.1, \\ (s_\flat = 0.1,\ 0) & \text{if } x > 0.1. \end{cases}$$

On the left boundary, we prescribe the Dirichlet condition $(s, c)(x = 0, t) = (1, 10^{-4})$. On the right boundary, we impose a homogeneous Neumann condition. This scenario corresponds to a water and polymer injection through a well located at $x = 0.1$ with an injection speed $u = 1$ in an oil-saturated plug. The plug corresponds to the slice $[0.1, 1.0]$ of the computational domain. The simulations are stopped at time $T = 0.4$ and we plot the saturation and mass fraction profiles along the computational domain.

Using $s_\flat = 0.1$ as the irreducible water saturation and assuming that the residual oil saturation is equal to 0, we consider a very simple Brooks-Corey model with the exponent equal to 2 for water and oil and unity maximum relative permeabilities. In other words, the phase mobilities are $\Lambda_w(s) = (\frac{s-s_\flat}{1-s_\flat})^2$ and $\Lambda_o(1-s) = (\frac{1-s}{1-s_\flat})^2$. For the mobility reduction R_m, we consider the formula (3) with coefficients $a_1 = 2.0 \cdot 10^3$, $a_2 = 2.8 \cdot 10^6$, $a_4 = 9.0 \cdot 10^{19/2}$. These values are adapted from the fitted experimental data given in [4] to fit to our dimensionless system. This configuration is simulated with the percolation IPV model (6) and the constant IPV model.

For the first case, we use the percolation IPV given by formula (6) with $s_\bullet = 0.05$. The simulation is performed with the upwind scheme and with our relaxation scheme. The results are compared with the exact solution and plotted in Fig. 1. The solution structure is the following: for $0.1 < x < 0.59$ we have a rarefaction wave (s-wave), at $x = 0.59$ we have the contact discontinuity (c-wave) with the resonance embedded, then for $x = 0.63$ we have a shock wave (s-wave). The results obtained with the relaxation scheme and the upwind scheme are quite similar. The contact discontinuity is poorly resolved because it is smeared. In fact, at this point we have the resonance, the s-wave and the c-wave have the same speed, the s-wave is no longer genuinely nonlinear and the rarefaction and the contact discontinuity are not clearly separated by the numerical schemes. Concerning the mass fraction profile, with this IPV model, it decreases slightly in the rarefaction wave and there is no polymer accumulation (or higher mass fraction) at the contact discontinuity, which is not the physically

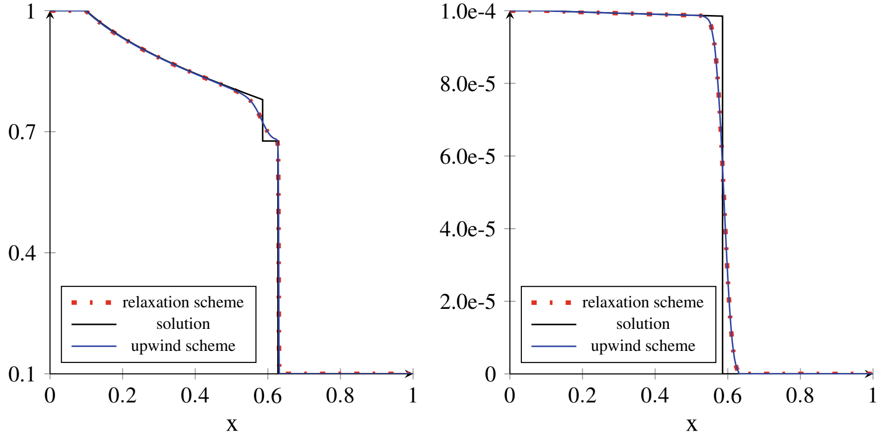


Fig. 1 Water saturation profiles (*left*) and polymer mass fraction (*right*) obtained with the percolation IPV law at time $T = 0.4$

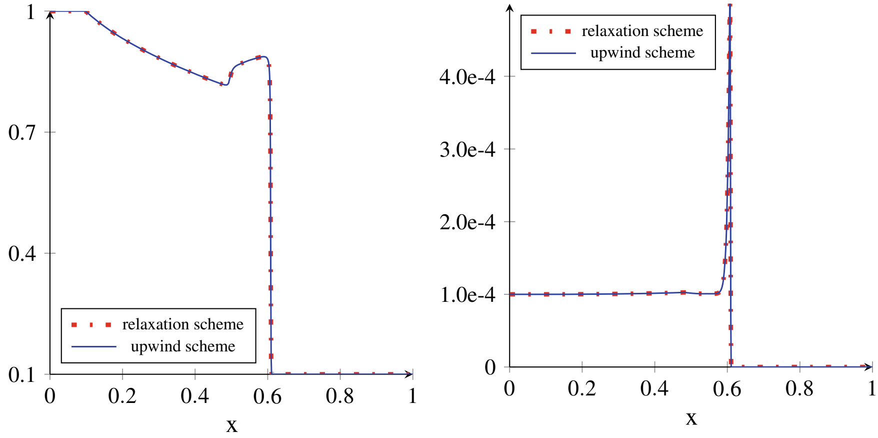


Fig. 2 Water saturation profiles (*left*) and polymer mass fraction (*right*) obtained with the constant IPV law at time $T = 0.4$

expected behavior. Nevertheless, this accumulation can be observed for other initial data.

For the second case, we use the constant IPV $\gamma = 1.2$. The results are plotted in Fig. 2. The simulation is performed with the upwind scheme and with our relaxation scheme. For this case with a non-hyperbolic region, there is no known exact solution. The solution structure is the following: for $0.1 < x < 0.49$ we have a rarefaction wave (s-wave), for $0.49 < x < 0.6$ we have an accumulation of water, followed by a peak in polymer concentration at $x = 0.6$ due to hyperbolicity loss. In fact, compared with the previous case, the contact discontinuity catches the shock waves and this yields to the hyperbolicity loss and a $\delta-$shock. At this point, the polymer mass fraction is supposed to be infinite. The relaxation scheme and the upwind scheme give equivalent results. The polymer mass fraction peak is about 5 times higher than the initial left mass fraction. This shows that for some given initial data, we can obtain very different results with different IPV model. Thus, validating the numerical results is a difficult task in our context. Until now, there was no mathematical hint to validate or reject the solution provided by the upwind scheme. With our relaxation model, we prove formally that the solution obtained here by both scheme is a dissipative limit of the polymer flooding solution and that it can be considered by physicists.

4 Conclusion

In this paper, we have introduced a simplified polymer flooding model and two IPV laws for polymers. The polymer flooding model is well- or ill-posed, depending on the IPV law. It is either weakly hyperbolic with a resonance region, or non-hyperbolic. To approximate numerically its solutions, we proposed a relaxation model that is unconditionally hyperbolic for any IPV model. We proved that this relaxation model is linearly degenerated and that the Riemann problem solutions are easy to compute. We also proved that the relaxation model is a dissipative approximation of the original model for a small enough relaxation time. The associated numerical scheme provides relevant results in resonant and non-hyperbolic cases. Future works will strive: (i) to reduce the numerical diffusion in the resonance regions, (ii) to try other IPV laws and compute reference solutions, and (iii) to take into account the polymer adsorption and shear thinning effects in the mobility reduction function.

References

1. Bartelds, G.A., Bruining, J., Molenaar, J.: The modeling of velocity enhancement in polymer flooding. Transp. Porous Media **26**(1), 75–88 (1997). https://doi.org/10.1023/A:1006563532277
2. Baudin, M., Berthon, C., Coquel, F., Masson, R., Tran, Q.H.: A relaxation method for two-phase flow models with hydrodynamic closure law. Numer. Math. **99**(3), 411–440 (2005). https://doi.org/10.1007/s00211-004-0558-1
3. Bouchut, F.: Nonlinear Stability of Finite Volume Methods for Hyperbolic Conservation Laws and Well-Balanced Schemes for Sources. In: Frontiers in Mathematics. Birkhäuser, Basel (2004). https://doi.org/10.1007/b93802
4. Braconnier, B., Preux, C., Flauraud, É., Tran, Q.H., Berthon, C.: An analysis of physical models and numerical schemes for polymer flooding simulations. Comput. Geosci. **21**(5), 1267–1279 (2017). https://doi.org/10.1007/s10596-017-9637-0
5. de Gennes, P.G.: Scaling Concepts in Polymer Physics. Cornell University Press, Ithaca (1979)
6. Hilden, S.T., Nilsen, H.M., Raynaud, X.: Study of the well-posedness of models for the inaccessible pore volume in polymer flooding. Transp. Porous Media **114**(1), 65–86 (2016). https://doi.org/10.1007/s11242-016-0725-8

7. Isaacson, E.L., Temple, J.B.: Analysis of a singular hyperbolic system of conservation laws. J. Diff. Equ. **65**(2), 250–268 (1986). https://doi.org/10.1016/0022-0396(86)90037-9
8. Keyfitz, B.L., Kranzer, H.C.: A system of non-strictly hyperbolic conservation laws arising in elasticity theory. Arch. Rat. Mech. Anal. **72**(3), 219–241 (1980). https://doi.org/10.1007/BF00281590
9. Pope, G.A.: The application of fractional flow theory to enhanced oil recovery. SPE J. **20**(3), 191–205 (1980). https://doi.org/10.2118/7660-PA.SPE-7660-PA
10. Suliciu, I.: On the thermodynamics of fluids with relaxation and phase transitions. Fluids with relaxation. Int. J. Engin. Sci. **36**, 921–947 (1988). https://doi.org/10.1016/S0020-7225(98)00005-6

The FVC Scheme on Unstructured Meshes for the Two-Dimensional Shallow Water Equations

Moussa Ziggaf, Mohamed Boubekeur, Imad kissami, Fayssal Benkhaldoun, and Imad El Mahi

Abstract The fluid flow transport and hydrodynamic problems often take the form of hyperbolic systems of conservation laws. In this work we will present a new scheme of finite volume methods for solving these evolution equations. It is a family of finite volume Eulerian–Lagrangian methods for the solution of non-linear problems in two space dimensions on unstructured triangular meshes. The proposed approach belongs to the class of predictor-corrector procedures where the numerical fluxes are reconstructed using the method of characteristics, while an Eulerian method is used to discretize the conservation equation in a finite volume framework. The scheme is accurate, conservative and it combines advantages of the modified method of characteristics to accurately solve the non-linear conservation laws with a finite volume method to discretize the equations. The proposed Finite Volume Characteristics (FVC) scheme is also non-oscillatory and avoids the need to solve

M. Ziggaf (✉) · I. kissami · F. Benkhaldoun · I. E. Mahi
CSEHS, Mohammed VI Polytechnic University, Lot 660,
43150 Bengeurir, Morocco
e-mail: Moussa.ziggaf@um6p.ma

I. kissami
e-mail: Imad.KISSAMI@um6p.ma

F. Benkhaldoun
e-mail: fayssal@math.univ-paris13.fr

I. E. Mahi
e-mail: Imad.ELMAHI@um6p.ma

M. Ziggaf · F. Benkhaldoun
LAGA, CNRS, UMR 7539, Université Sorbonne Paris Nord,
F-93430 Villetaneuse, France

M. Ziggaf · I. E. Mahi
ENSAO, LM2N, Complexe Universitaire, B.P. 669, 60000 Oujda, Morocco

M. Boubekeur
Laboratoire Analyse, Géométrie et Applications, LAGA, CNRS, UMR 7539,
Université Sorbonne Paris Nord, 93430 Villetaneuse, France
e-mail: boubekeur@math.univ-paris13.fr

R. Klöfkorn et al. (eds.), *Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples*, Springer Proceedings in Mathematics & Statistics 323,
https://doi.org/10.1007/978-3-030-43651-3_42

a Riemann problem. Several test examples will be presented for the shallow water equations. The results will be compared to those obtained with the Roe.

Keywords Shallow water equations · Finite volume method · Method of characteristics · FVC scheme · Unstructured meshes

1 Introduction

Incompressible Navier–Stokes equations have been widely used in the literature to simulate water flows including eddy diffusion and Coriolis forces, see for example [6, 16]. However, for free-surface flows these models often become complicated due to the presence of moving boundaries within the flow domain and also due to the inclusion of hydrostatic pressure. Under certain assumptions these models can be replaced by the well-established shallow water equations. Indeed, the shallow water equations can be derived by depth-averaging the three-dimensional Navier–Stokes equations assuming that the pressure is hydrostatic and the vertical scale is far smaller than the horizontal scale, see [1]. In their depth-averaged form, shallow water equations have been used to model many engineering problems in hydraulics and free-surface flows including tides in coastal regions, rivers, open channel flows, etc. see for instance [5, 11]. Developing highly accurate numerical solvers for shallow water equations presents a challenge due to the non-linear aspect of these equations and their coupling through the source terms. More precisely, the difficulty in these models lies in the coupling terms involving some derivatives of the physical variables that make the system non-conservative and sometimes non-hyperbolic. A class of Eulerian–Lagrangian methods have also been used in [4] to solve the two-dimensional shallow water equations. This method avoids the solution of Riemann problem and belongs to the finite volume predictor-corrector type methods. The predictor stage uses the method of characteristics to reconstruct the numerical fluxes whereas the corrector stage recovers the conservation equations in the finite volume framework. Numerical results reported in [4] for two-dimensional shallow water equations demonstrate that this method is robust and more accurate than the Roe and SRNH schemes, but this previous work was limited to the **Cartesian mesh**. In this paper, the method is extended to the **unstructured mesh**. The results presented here show highly accurate solution by using our proposed finite volume characteristics method and confirm its capability to provide accurate and efficient simulations using unstructured meshes for shallow water flows, including Coriolis forces. This paper is organized as follows. The rotating shallow water equations and their projected speed model are presented in Sect. 2. In Sect. 3, the numerical method is formulated for the reconstruction of the FVC scheme. The Sect. 4 is devoted to numerical results for several test examples for partial dam-break problem and rotating shallow water equations. Finally, the Sect. 5 contains concluding remarks and perspectives.

2 Mathematical Model

2.1 The Rotating Shallow Water Model

The shallow water equations for the free-surface flow in two dimensions with the Coriolis forces are formulated as

$$\begin{cases} \partial_t h + \partial_x(hu) + \partial_y(hv) = 0 \\ \partial_t(hu) + \partial_x(hu^2 + \frac{1}{2}gh^2) + \partial_y(huv) = f_c hv \\ \partial_t(hv) + \partial_x(huv) + \partial_y(hv^2 + \frac{1}{2}gh^2) = -f_c hu \end{cases} \tag{1}$$

where g is the gravitational acceleration, f_c is the Coriolis force, h is the water depth, u and v are the depth-averaged velocities. It is well known that the system (1) is strictly hyperbolic with real and distinct eigenvalues. The conservative form of (1) is

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

$$W = \begin{pmatrix} h \\ hu \\ hv \end{pmatrix}, \quad \mathbb{F}(W) = \left(\begin{pmatrix} hu \\ hu^2 + \frac{1}{2}gh^2 \\ huv \end{pmatrix}, \begin{pmatrix} hv \\ huv \\ hv^2 + \frac{1}{2}gh^2 \end{pmatrix} \right)^T, \quad Q(W) = \begin{pmatrix} 0 \\ f_c hv \\ -f_c hu \end{pmatrix}$$

The system of Eq. (2) has to be solved in a bounded spatial domain Ω, with given boundary and initial conditions.

2.2 Construction of the Projected Speed Model

In this section we adopt the same calculation techniques used in the Sect. 2 of [4] in order to get the projected speed model. The differential form of the projected speed model is

$$\begin{cases} \dfrac{\partial h}{\partial t} + \dfrac{\partial hu_\eta}{\partial \eta} = 0, \\ \dfrac{\partial hu_\eta}{\partial t} + \dfrac{\partial}{\partial \eta}\left(hu_\eta{}^2 + \dfrac{1}{2}gh^2 \right) = f_c hu_\tau, \\ \dfrac{\partial hu_\tau}{\partial t} + \dfrac{\partial}{\partial \eta}\left(hu_\eta u_\tau \right) = -f_c hu_\eta, \end{cases} \tag{3}$$

The system (3) can be rewritten as a transport equation form

$$\frac{\partial \mathbf{U}}{\partial t}(t, X) + u_\eta(t, X)\frac{\partial \mathbf{U}}{\partial \eta}(t, X) = \mathbf{F}(\mathbf{U}, f_c), \quad \forall\ X = (x, y) \in \Omega \subset \mathbb{R}^2, \ \ t > t_0 \tag{4}$$

with, $\mathbf{U}=\begin{pmatrix} h \\ u_\eta \\ u_\tau \end{pmatrix}$, $\begin{pmatrix} u_\tau \\ u_\eta \end{pmatrix}=\begin{pmatrix} vn_x - un_y \\ un_x + vn_y \end{pmatrix}$, *and* $\mathbf{F}(\mathbf{U},f_c)=\begin{pmatrix} -h\partial_\eta(u_\eta) \\ -g\partial_\eta(h)+f_c u_\tau \\ -f_c u_\eta \end{pmatrix}$

The system of Eq. (4) is used only to reconstruct the numerical fluxes while the finite volume method is applied directly to the conservative system (2), see [3, 14].

3 Finite Volume Characteristics Scheme

In this section we present the finite volume characteristics method for the numerical solution of the shallow water Eq. (1). The method consists of two steps and can be interpreted as a predictor-corrector approach. The first step deals with the finite volume discretization of the equations whereas in the second step, the reconstruction of the numerical fluxes is discussed.

3.1 Finite Volume Discretization

The classical finite volume discretization of the system (2) without the bathymetry terms is the volume integral over the total volume of the cell T_i, which gives

$$\frac{dW_i}{dt}+\frac{1}{|T_i|}\sum_{j\in N(i)}|\gamma_{ij}|\Phi(W_{ij},\mathbf{n}_{ij})=Q_i \tag{5}$$

where $W_i=\dfrac{1}{|T_i|}\displaystyle\int_{T_i}W\,dV$, $\Phi(W_{ij},\mathbf{n}_{ij})\simeq\dfrac{1}{|\gamma_{ij}|}\displaystyle\int_{\gamma_{ij}}\mathbb{F}(W)\cdot\mathbf{n}_{ij}d\sigma$,
$|T_i|$ denotes the area of the cell T_i and γ_{ij} is the edge surrounding the cell T_i and $N(i)$ is the neighbouring triangles of the cell T_i. $\Phi(W_{ij},\mathbf{n}_{ij})$ is the numerical flux computed at the interface between the cells i and j. The intermediate solution W_{ij} is reconstructed using the characteristic method in the predictor stage. The time discretization of (5) is performed by a first order explicit Euler scheme. The time domain is divided into N subintervals $[t_n, t_{n+1}]$ with time step $\Delta t=t_{n+1}-t_n$ for $n=0,1,\ldots,N$. W^n is the value of a generic function W at time t_n. The fully-discrete formulation of the system (2) is given by

$$W_i^{n+1}=W_i^n-\frac{\Delta t}{|T_i|}\sum_{j\in N(i)}|\gamma_{ij}|\Phi(W_{ij}^n,\mathbf{n}_{ij})+\Delta t Q_i^n \tag{6}$$

3.2 Flux Construction

In the present study, we reconstruct the numerical flux $\Phi(W_{ij}^n, \mathbf{n}_{ij})$ using the method of characteristics. The fundamental idea of this method is to impose a regular grid at the new time level and to backtrack the flow trajectories to the previous time level, for more details see [13, 15]. At the previous time level, the quantities that are needed are evaluated by interpolation from their known values on a regular grid.

3.2.1 Method of Characteristics

The characteristic curves associated with the Eq. (4) are solutions of the initial-value problem

$$\begin{cases} \dfrac{dX^c(t)}{dt} = u_\eta(t, X^c(t)) \cdot \mathbf{n} \quad t \in [t_n, t_n + \alpha \Delta t], \quad \alpha > 0 \\ X^c(t_n + \alpha \Delta t) = X^* \end{cases} \tag{7}$$

The solution of (7) can be expressed in an integral form as

$$X^c(t_n) = X^* - \int_{t_n}^{t_n+\alpha\Delta t} u_\eta(s, X^c(s)) \cdot \mathbf{n}\, ds \tag{8}$$

This integral can be calculated using the integral approximation methods. In our simulations we used a first-order Euler method to approximate the integral in (8). The numerical fluxes in (6) are reconstructed using the solution of the transport Eq. (4) which is given by

$$\mathbf{U}(t_n + \alpha\Delta t, X^*) = \mathbf{U}(X^c(t_n)) + \int_{t_n}^{t_n+\alpha\Delta t} \mathbf{F}(\mathbf{U}(X^c(s), s), f_c)\, ds \tag{9}$$

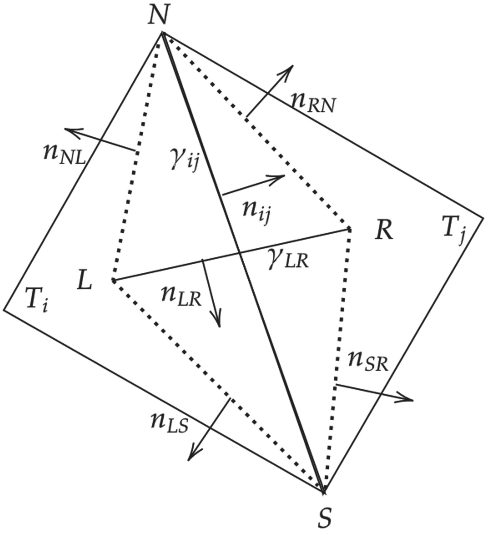
where $\mathbf{U}(t_n + \alpha\Delta t, X^*)$ is the solution at the characteristic feet. It is computed by interpolation of the departure point $X^c(t_n)$ on the mesh. we used the scattered interpolation methods proposed in [2]. The integral in (9) is calculated using the mind-point rule. This approximation is formulated as

$$\mathbf{U}_{ij}^n = \hat{\mathbf{U}}_{ij}^n + \alpha\Delta t \mathbf{F}(\hat{\mathbf{U}}_{ij}^n, f_c) \tag{10}$$

where $\hat{\mathbf{U}}_{ij}^n$ is the interpolated solution. To approximate $\mathbf{F}(\mathbf{U}, f_c)$, (i.e. $\partial_\eta(u_\eta), \partial_\eta(h), \ldots$) we need to approximate these derivatives at the interfaces, for that we use the diamond cell as expressed in Fig. 1. For more details see the Sect. 3.1.1.2 of [9]. The gradient value at the interface is

$$\nabla u_{ij} = \frac{1}{2\mu_{_{SRNL}}} \Big\{ (u_{_S} - u_{_N})\mathbf{n_{LR}}|\gamma_{_{LR}}| + (u_{_R} - u_{_L})\mathbf{n_{ij}}|\gamma_{ij}| \Big\} \tag{11}$$

Fig. 1 Diamond cell in 2D



where μ_{SRNL} is the area of the co-volume *SRNL*. After the discretization of the source term $\mathbf{F}(\mathbf{U}, f_c)$, the district equations system (10) can be written as

$$h^n_{ij} = \hat{h}^n_{ij} - \alpha\Delta t\hat{h}^n_{ij}\nabla(\hat{u}_\eta)^n_{ij}$$
$$(u_\eta)^n_{ij} = (\hat{u}_\eta)^n_{ij} - \alpha g\Delta t\nabla\hat{h}^n_{ij} + \alpha\Delta t f_c(\hat{u}_\tau)^n_{ij}$$
$$(u_\tau)^n_{ij} = (\hat{u}_\tau)^n_{ij} - \alpha\Delta t f_c(\hat{u}_\eta)^n_{ij}$$

Once these projected states are calculated, the states W_{ij} are recovered by using the transformations, $u^n_{ij} = (u_\eta)^n_{ij}n_x - (u_\tau)^n_{ij}n_y, \quad v^n_{ij} = (u_\tau)^n_{ij}n_x + (u_\eta)^n_{ij}n_y$

► *The FVC scheme on unstructured meshes for the present model*

$$\left|\begin{array}{lcl} W^n_{ij} & = & (h^n_{ij} \quad h^n_{ij}u^n_{ij} \quad h^n_{ij}v^n_{ij})^T, \qquad \Phi(W^n_{ij}, \mathbf{n}_{ij}) \;=\; \mathbb{F}(W^n_{ij})\cdot\mathbf{n}_{ij} \\ W^{n+1}_i & = & W^n_i - \dfrac{\Delta t}{|T_i|}\displaystyle\sum_{j\in N(i)} |\gamma_{ij}|\Phi(W^n_{ij}, \mathbf{n}_{ij}) + \Delta t Q^n_i \end{array}\right.$$

4 Results

In this section we perform numerical tests with our Finite Volume Characteristics scheme on unstructured meshes for the two-dimensional shallow water equations. In all our computations a fixed Courant number $CFL = 0.8$ and $\alpha = 1.2$, are used while the time step Δt is varied according to the stability condition $\Delta t = CFL\dfrac{\min_i |\gamma_{ij}|}{\sqrt{2\alpha}\lambda^n_{ij}}, \quad \lambda^n_{ij} = \max_p\{|u^n_{pi} + \sqrt(gh^n_{pi})|, \ |v^n_{pj} + \sqrt(gh^n_{pj})|\}$. The used computer is an Intel Core i7-8565U CPU @ 1.80GHz × 8, with 15 GB RAM.

4.1 Accuracy Test Example

The accuracy of the proposed unstructured FVC scheme for a shallow water system is checked, it is compared to the analytical solution. We solve the shallow water Eq. (1) without source terms in the squared domain $\Omega = [0, 100] \times [0, 100]$ with initial solution for the water depth as the dam-break problem $h(0, x, y) = 4\,\text{m}, \quad (x, y) < (0, 0); \;\; h(0, x, y) = 2\,\text{m}, \quad (x, y) > (0, 0);$ and $u(0, x, y) = v(0, x, y) = 0\,\text{m/s}$. We also compare the results obtained using our FVC scheme on an unstructured mesh to those obtained using the well established Roe scheme in [12]. The results presented in the table below are obtained with the relative L^1-error norm corresponding to the water depth defined as $\dfrac{\sum_{i=1}^{N_{ele}} |T_i||h_i^n - h(t_n, x_i, y_i)|}{\sum_{i=1}^{N_{ele}} |T_i||h(t_n, x_i, y_i)|}$, where h_i^n and $h(t_n, x_i, y_i)$ are respectively, the computed and exact water depth at the cell T_i, and N_{ele} denotes the total number of cells. The Relative L^1-error is obtained for the accuracy test example at time $t = 5.5\,\text{s}$ using the Roe and FVC schemes for different unstructured mesh. We remark that the relative L^1-error for the FVC scheme is smaller than for Roe scheme, but the convergence order is still the same and it is close to 1 (Fig. 2).

# Cells	Roe	FVC
	L^1-error	L^1-error
2592	2.0867×10^{-2}	1.5695×10^{-2}
5000	1.7154×10^{-2}	1.1913×10^{-2}
10082	1.3389×10^{-2}	8.3061×10^{-3}
20073	1.0112×10^{-2}	5.1472×10^{-3}

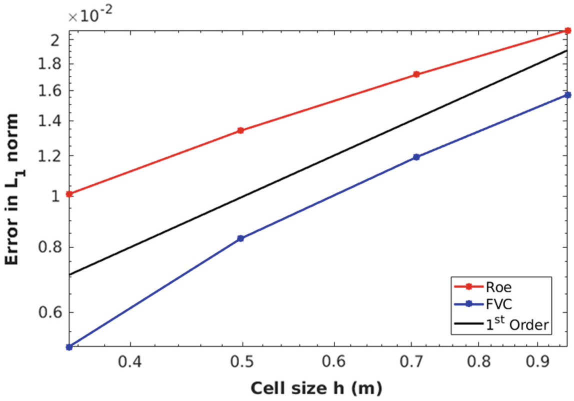


Fig. 2 Convergence rates and L^1-error. Comparison between FVC and Roe schemes on an unstructured mesh using the same code structure

4.2 Circular Dam-Break Problem

This benchmark was used in [4] to represent the FVC scheme on structured Cartesian mesh. We solve the shallow water Eq. (1) on a flat bottom in the spatial domain $\Omega = [-10, 10] \times [-10, 10]$ equipped with the following initial conditions

$$h(0,x,y) = 1 + \frac{1}{4}\left(1 - \tanh\left(\frac{\sqrt{ax^2 + by^2} - 1}{c}\right)\right), \quad u(0,x,y) = v(0,x,y) = 0\,\text{m/s}\,,$$

where $a = \frac{5}{2}$, $b = \frac{2}{5}$, and $c = 0.1$, $g = 1\,\text{m/s}^2$ and $f_c = 1\,\text{Kg m/s}^2$ as in [4] (Fig. 3). The domain Ω is discretized with unstructured triangular mesh of 10052 cells. In this simulation we applied the Neumann conditions on all boundaries (see the Sect. 7.5.2 in [10]).

As it can be clearly seen, the results obtained using FVC scheme on unstructured mesh are very similar to those performed with FVC on structured Cartesian mesh (see Sect. 4.2 in [4]). The rotational movement due to the effect of Coriolis forces provides an ellipsoid profile, which implies non radial symmetry.

4.3 Partial Dam-Break Problem

This benchmark consists of studying the torrential flow (i.e. Froude number $F_r > 1$) due to a partial and asymmetrical dam-break. This benchmark was proposed in [7]. Let's study a basin 200 m wide, 200 m long and flat bottom, without friction. Water is retained in the left part of the basin.

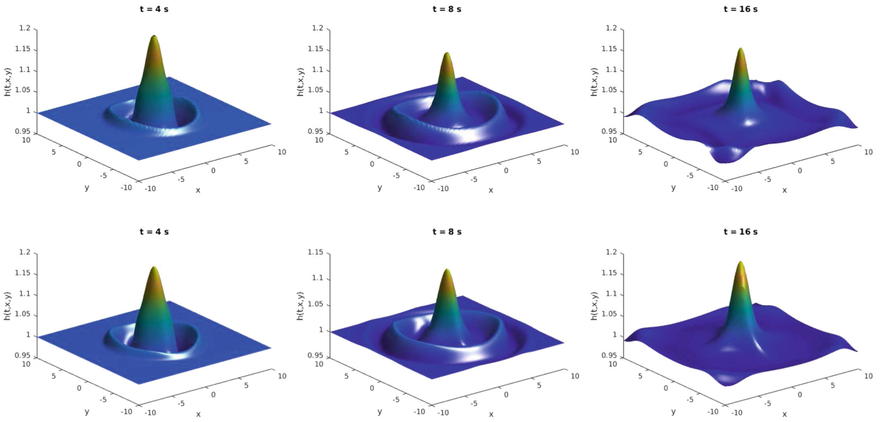
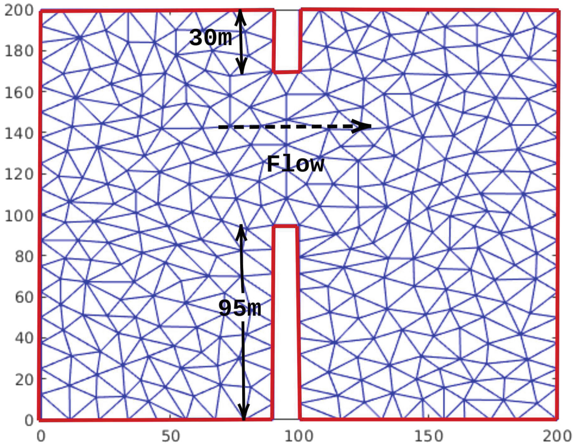


Fig. 3 Water depth obtained at different times, using FVC on unstructured mesh (first line) and FVC on a Cartesian mesh (second line)

Fig. 4 Partial dam-break domain



The thickness of the dam is 10 m on the flow direction. see Fig. 4. Initially $hr/hl = 0.5$ is fixed with $hl = 4$ m as water depth in the reservoir and $hr = 2$ m as the water level downstream of the dam. The water in the basin is at rest at $t = 0$. When the region occupied by the fluid is bordered by a solid surface, the fluid can not pass through it. Its speed is necessarily zero in the direction perpendicular to the surface. On the other hand, it is not necessarily null in the tangential directions. In this simulation, a no-slip boundary condition is imposed on all walls see Sect. 3.2 in [8]. The domain studied was discretized in 20002 non-uniform cells. The duration of simulation is 8.2 s counted from the partial dam break.

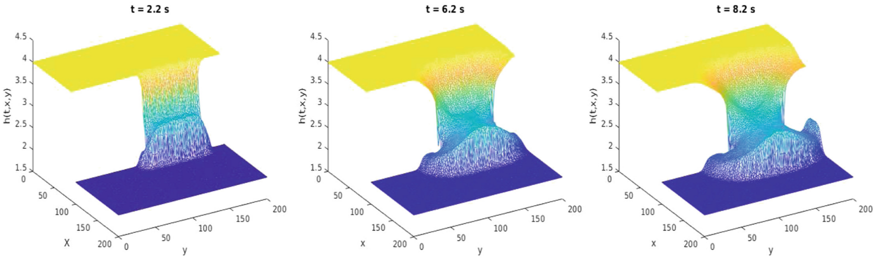


Fig. 5 Water depth for the partial dam-break problem on flat bottom obtained at different times ($t = 2.2$, 6.2 and 8.2 s) using FVC scheme on an unstructured mesh

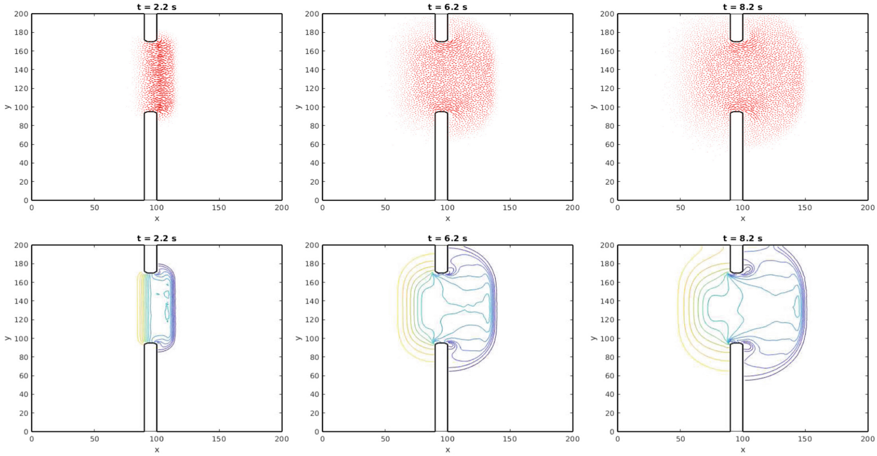


Fig. 6 Velocity fields and contours for the partial dam-break problem corresponding to the water depth represented in the Fig. 5

5 Conclusion

A finite volume-characteristics method to solve two-dimensional shallow water equations on unstructured meshes has been presented (Fig. 6). This method combines the advantages of the finite volume discretization and the method of characteristics, it solves also steady flows without large numerical errors and compute the numerical flux corresponding to the real state of water flow without relying on Riemann problem solvers. The reasonable accuracy can be obtained easily and no special treatment is needed to maintain a numerical balance, because it is performed automatically in the integrated numerical flux function. Finally, the proposed approach does not require either non-linear solution of algebraic equations or special front tracking techniques. Furthermore, it has strong applicability to various problems in rotating shallow water flows as shown in the numerical results. The outlook of this work is to extend this approach to a multi-layers model of shallow water equations with a bathymetry where we can guarantee a balance between the gradient flux and the source term. In a further step, we will work on coupling this model with the transport convection equation.

References

1. Abbott, M.: Elements Of The Theory Of Free Surface Flows, vol. 001. Pitman, London (1979)
2. Amidror, I.: Scattered data interpolation methods for electronic imaging systems: a survey. J. Electron. Imaging **11**(ARTICLE), 157–76 (2002)
3. Benkhaldoun, F., Elmahi, I., Seaïd, M.: A new finite volume method for flux-gradient and source-term balancing in shallow water equations. Comput. Methods Appl. Mech. Eng. **199**(49–52), 3324–3335 (2010)

4. Benkhaldoun, F., Sari, S., Seaid, M.: Projection finite volume method for shallow water flows. Math. Comput. Simul. **118**, 87–101 (2015)
5. Churuksaeva, V., Starchenko, A.: Mathematical modeling of a river stream based on a shallow water approach. Procedia Comput. Sci. **66**, 200–209 (2015)
6. Codina, R.: Numerical solution of the incompressible navier-stokes equations with coriolis forces based on the discretization of the total time derivative. J. Comput. Phys. **148**(2), 467–496 (1999)
7. Fennema, R.J., Hanif Chaudhry, M.: Implicit methods for two-dimensional unsteady free-surface flows. J. Hydraul. Res. **27**(3), 321–332 (1989)
8. Godlewski, E., Raviart, P.A.: Numerical Approximation of Hyperbolic Systems of Conservation Laws, vol. 118. Springer Science & Business Media, Berlin (2013)
9. Karel, J.: Numerical simulation of streamer propagation on unstructured dynamically adapted grids. Ph.D. thesis (2014)
10. Mazumder, S.: Numerical Methods for Partial Differential Equations: Finite Difference and Finite, Volume Methods. Academic Press, Cambridge (2015)
11. Özgen, I., Zhao, J., Liang, D., Hinkelmann, R.: Urban flood modeling using shallow water equations with depth-dependent anisotropic porosity. J. Hydrol. **541**, 1165–1184 (2016)
12. Roe, P.L.: Approximate riemann solvers, parameter vectors, and difference schemes. J. Comput. Phys. **43**(2), 357–372 (1981)
13. Roe, P.L.: Characteristic-based schemes for the euler equations. Annu. Rev. Fluid Mech. **18**(1), 337–365 (1986)
14. Sahmim, S., Benkhaldoun, F., Alcrudo, F.: A sign matrix based scheme for non-homogeneous pdes with an analysis of the convergence stagnation phenomenon. J. Comput. Phys. **226**(2), 1753–1783 (2007)
15. Seaïd, M.: On the quasi-monotone modified method of characteristics for transport-diffusion problems with reactive sources. Comput. Methods Appl. Math. **2**(2), 186–210 (2001)
16. Yan, J., Deng, X., Korobenko, A., Bazilevs, Y.: Free-surface flow modeling and simulation of horizontal-axis tidal-stream turbines. Comput. Fluids **158**, 157–166 (2017)

Numerical Analysis of a Finite Volume Scheme for the Optimal Control of Groundwater Pollution

Catherine Choquet, Moussa Mory Diédhiou, and Houssein Nasser El Dine

Abstract This paper is devoted to an optimal control problem of the underground water contaminated by agricultural pollution, the spatiotemporal objective taking into account the trade-off between the fertilizer used by the farmer to increase profits and the cleaning costs which are necessary to treat the water before it is distributed to users. The constraint is a hydrogeological model for the spread of the pollution in the aquifer which consists in a system of a parabolic partial differential equation and an elliptic equation. Hydrogeological and economic modelling are thus combined in the problem. We propose a finite volume scheme based on a two-point flux approximation with upwind mobilities of an optimal control. Numerical simulations are provided to illustrate the 2D and 3D optimal solutions.

Keywords Optimal control problem · Hydrogeological state equations · Nonlinearly coupled problem · Parabolic and elliptic PDEs · Finite volume scheme

MSC (2010) 37N40 · 76R99 · 37N35 · 65M12 · 65M08 · 76S05

1 Introduction

The preservation of water resources is a major issue in view of the growing world population. This water resource is today threatened by different kinds of pollution. In Europe, for instance, agriculture is the main pollutant source, with 50–80% of the total nitrogen and phosphorus loaded with fresh water (see [6]). According to the report of the parliamentary office for the evaluation of scientific choices (see [10]),

C. Choquet · M. Mory Diédhiou (✉) · H. Nasser El Dine
MIA Lab, La Rochelle University, Avenue A. Einstein, 17031 La Rochelle, France
e-mail: moussa_mory.diedhiou@univ-lr.fr

C. Choquet
e-mail: catherine.choquet@univ-lr.fr

H. Nasser El Dine
e-mail: houssein.nasser_el_dine@univ-lr.fr

R. Klöfkorn et al. (eds.), *Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples*, Springer Proceedings in Mathematics & Statistics 323,
https://doi.org/10.1007/978-3-030-43651-3_43

the cost of denitrification of water is around half euro per cubic meter. For finding a compromise between the benefits and the cleaning costs, a recent model is derived in [8] through an optimal control formulation.

Hydrogeologically, modelling the spread of a miscible pollutant in groundwater amounts to model the flow of incompressible miscible fluids in porous media. Here convection-diffusion and reaction phenomena are taken into account in order to be more realistic. The economic point of view consisting in finding an optimal policy by taking into account the costs of decontamination is classical. Nevertheless, most of the existing models do not depend on space and are thus unrealistic especially due to the delay between the application of any policy and its effects induced by the small flow speed in aquifers (see [4, 7]). Here the space dependence is fully included in the model. Many of the economic models are also restricted to linear state equations, when in reality the hydrogeological modelling leads to strongly nonlinear equations governing the transport of pollutant in the groundwater. Recents works of Comte and *al.* in [1, 8] have taken into account both the dependence on space and a more realistic (nonlinear) equation of state for modelling the transport and the diffusion of pollutant in the groundwater. The works [1, 8] are mainly devoted to the mathematical analysis of the problem, providing existence, uniqueness and asymptotic results. Our work may be viewed as the numerical implementation of the theoretical results in [1]. In [8], a mixed finite element scheme was proposed and used to get numerical illustrations, the implementation being done with FreeFEM. The few test cases are very academic and limited to the two-dimensional framework and to very short time scales, far from the scale at which the problem needs to be studied, from at least a few months to several years.

Here rather, we propose a finite volume architecture embedded in an iterative fixed point approximation to handle the full 3D optimal control problem. Notice also that the scheme has been built to withstand the strong parameter and scale contrasts induced by concrete applications.

The paper is organized as follows. In Sect. 2 we present the optimal control problem and introduce the adjoint problem which will be used for the next. Our numerical scheme is presented. In Sect. 3, some numerical tests are performed to illustrate the solution of the optimal control problem.

2 Presentation of the Problem

We consider $\Omega \subset \mathbb{R}^d$, $d \in \mathbb{N}$, $d \leq 3$, a bounded domain, representing an area containing both the area affected by the pollution and the groundwater collection wells. We assume that the boundary $\partial\Omega$ of Ω is such that $\partial\Omega \in \mathscr{C}^1$. In the case of the 3D example treated in the last section, the boundary of Ω is divided into six subsets for taking into account different physical boundary conditions, namely $\partial\Omega = \cup_{i=1}^{6}\Gamma_i$ where Γ_1, Γ_2, Γ_3, Γ_4 are the laterals faces and respectively Γ_5, Γ_6 are the top, bottom face. The time horizon is denoted by T, with $0 < T < \infty$. We set $\Omega_T = \Omega\times]0, T[$.

2.1 The Optimal Control Problem

Let $\bar{p} > 0$ be the maximal pollution load. The quantity $\bar{p}$ is for instance the maximal fertilizer load that can be applied on the field. Such a quantity exists due to obvious practical constraints imposed to the farmer. The natural admissible set of control is defined as

$$E = \left\{p \in L^2(\Omega_T);\, 0 \leq p(t,x) \leq \bar{p}\, \text{a.e. in}\, \Omega_T\right\}.$$

We consider a standard central planner objective, which is written as (see [11])

$$J(p) =$$
$$\int_0^T \left(\int_{\mathscr{L}} f(x, p(t,x))dx - \int_{\Omega} D(x, c(t,x))dx \right) e^{-\rho t} dt - \varsigma \int_{\Omega} D(x, c(T,x))dx\, e^{-\rho T},$$

where functions f and D respectively model the benefit of the farmer and the cleaning costs depending on the position of the production wells and on the pollutant concentration in the groundwater c, $\mathscr{L} \subset \Omega$ is a specific localization of the fertilizer load, ρ is the social discount rate, $0 < \rho < 1$, and ς is the weight of the terminal costs.

The problem is to find $\max_{p \in E} J(p)$ subject to state constraints

$$\text{div}(\text{v}) = \theta, \quad v = -\kappa \nabla \phi \quad \text{in} \quad \Omega_T, \qquad (1)$$

$$R\psi \partial_t c + \text{div}(v_c) = -r(c) + \text{X}_{\mathscr{L}}\, p, \quad v_c = -S(v)\psi \nabla c + cv \quad \text{in} \quad \Omega_T, \qquad (2)$$

completed by the following initial condition $c_{|t=0} = c_0$ and boundary conditions

$$\begin{cases} v \cdot n = 0, \qquad v_c \cdot n = 0 \quad \text{on} \quad \Gamma_0 = \Gamma_2 \cup \Gamma_4 \cup \Gamma_5 \cup \Gamma_6, \\ \phi = 0, \qquad v_c \cdot n = f_1 \quad \text{on} \quad \Gamma = \Gamma_1 \cup \Gamma_3, \end{cases} \qquad (3)$$

where n is the outward normal to the boundary, v is the Darcy velocity, $r(c)$ is the chemical reaction term, $S(\text{v}) = S_m \times \mathbb{I}_{\mathbb{R}^d} + S_p(\text{v})$ is a nonlinear dispersion tensor depending on the longitudinal and transverse components of the dispersion and of the Darcy velocity, κ is the mobility of the fluid in the soil, $\text{X}_{\mathscr{L}}$ is the characteristic function of the set $\mathscr{L}$, $\theta(x)$ is a source term and f_1 is the Neumann condition on Γ. Other parameters are $R > 0$, a so-called retardation factor due to the possible instantaneous reactions, $S_m > 0$ the diffusion coefficient $\psi > 0$.

2.2 The Adjoint Problem

The existence and the uniqueness of the solution of the optimal control problem introduced in Sect. 2.1 is proved in [1]. Its characterization using the necessary optimality conditions of the first order thus makes sense. Cancelling the variations of the Lagrangian associated to the problem with respect to the control p, to the state variable c and to the final state $c(T, x)$, respectively provides the terminal condition, the optimality condition (8), the adjoint equation (6) and the boundary conditions in (7) below. With the state equations, we get the adjoining problem equivalent to the optimal control problem, which is to find (ϕ, c, μ) such that

$$\operatorname{div}(v) = \theta, \quad v = -\kappa \nabla \phi, \tag{4}$$

$$R\psi \partial_t c + \operatorname{div}(v_c) = -r(c) + \mathrm{X}_{\mathscr{L}} p, \quad v_c = -S(v)\psi \nabla c + cv, \tag{5}$$

$$R\psi \partial_t \mu + \operatorname{div}(v_\mu) = -r'(c) - (R\psi\rho + \theta)\mu + \partial_c D(x, c), v_\mu = -S(v)\psi \nabla \mu - \mu v, \tag{6}$$

in Ω_T, completed by the following initial condition and boundary condition

$$\begin{cases} v \cdot n = 0, \quad v_c \cdot n = 0, \quad v_\mu \cdot n = 0 \quad \text{on} \quad \Gamma_0, \\ \phi = 0, \quad v_c \cdot n = f_1, \quad v_\mu \cdot n = f_2 \quad \text{on} \quad \Gamma, \\ c_{|t=0} = c_0, \quad R\psi \mu_{|t=0} = \varsigma \partial_c D(\cdot, c_{|t=T}), \end{cases} \tag{7}$$

$$\partial_p f(x, p) = \mu. \tag{8}$$

Notice that we have used a time variable change in the adjoint unknown so that its terminal condition appears as an initial condition.

We are going to construct a finite volume scheme on an orthogonal admissible mesh coupled with an iterative fixed point algorithm for the approximation of (4)–(8). Here, for the sake of the simplicity in the notations, we treat the case where

$$\kappa = k \times \mathbb{I}_{\mathbb{R}^d}, \qquad S(\mathrm{v}) = S_m \times \mathbb{I}_{\mathbb{R}^d} + S_p(\mathrm{v}), \qquad S_p(v) = 0,$$

where k and S_m are given positive real numbers. We finally introduce some physically relevant assumptions for our system.

(H1) The function θ is a nonnegative such that $\theta^* \leq \theta(x) \leq \theta_*$.
(H2) The functions f_1, and f_2 belong to $L^2(\Gamma_1 \cup \Gamma_3)$.
(H3) The reaction function r belongs to $\mathscr{C}^1$.
(H4) The benefit function f is strictly concave and belongs to $\mathscr{C}^1$.
(H5) The function D is bounded and continuous such that $\frac{\partial D}{\partial c} \geq 0$ and $\frac{\partial^2 D}{\partial c^2} \geq 0$.

2.3 The Numerical Scheme

In this section, we explicit the discretization of the adjoint problem (4)–(7) using a finite volume scheme [9]. The nonlinear coupling between (5) and (6) through (8) is handled with an iterative fixed point algorithm. The complete approximation scheme is described in the present subsection. Notice that the mathematical analysis of the scheme is postponed to a forthcoming paper.

Let $\mathscr{T}$ be a regular and admissible mesh of the domaine Ω, constituting of open convex polygons called control volumes with maximum size (diameter) h. For all $K \in\mathscr{T}$, let x_K denote the center of K, $N(K)$ the set of the neighbours of K i.e the set of cells of $\mathscr{T}$ which have a common interface with K, by $N_{int}(K)$ the set of the neighbours of K located in the interior of $\mathscr{T}$, by $N_{ext}(K)$ the set of edges of K on the boundary $\partial\Omega$. Furthermore, for all $L \in N_{int}(K)$ denote by $d_{K,L}$ the distance between x_K and x_L, by $\sigma_{K,L}$ the interface between K and L, by $\eta_{K,L}$ the unit normal vector to $\sigma_{K,L}$ outward to K. And for all $\sigma \in N_{ext}(K)$, denote by $d_{K,\sigma}$ the distance from x_K to σ. For all $K \in \mathscr{T}$, we denote by $|K|$ the measure of K.
The admissibility of $\mathscr{T}$ implies that $\overline{\Omega} = \cup_{K\in\mathscr{T}} K \cap L = \emptyset$ if $K, L \in \mathscr{T}$, there exists a finite sequence of points $(x_K)_{K\in\mathscr{T}}$ and the straight line $\overline{x_K x_L}$ is orthogonal to the edge $\sigma_{K,L}$. We also need some regularity property for the mesh:

$$\min_{K\in\mathscr{T}, L\in N(K)} d_{K,L}/\mathrm{diam}(K) \geq \vartheta, \quad \text{for some} \quad \vartheta \in \mathbb{R}^+.$$

Let N_T be the number of time steps. We set $\Delta t = \frac{T}{N_T}$, $t^n = n\Delta t$, $0 \leq n \leq N_T$.

A finite volume scheme for the discretization of the problem (4)–(8) is given by the following set of equations with unknowns $P = (p_K^{m,n})_{K\in\mathscr{T}}$, $\Phi = (\phi_K)_{K\in\mathscr{T}}$, $C = (c_K^{m,n})_{K\in\mathscr{T}}$, $n \in [0, N_T]$ and $\Upsilon = (\mu_K^{m,n})_{K\in\mathscr{T}}$, $n \in [0, N_T]$, for all $K \in \mathscr{T}$ and for all $n \in [0, N_T]$, for all $m \in \mathbb{N}$.

$$c_K^{m,0} = \frac{1}{|K|}\int_K c_0(x)\,\mathrm{d}x, \quad p_K^{0,n} = \frac{1}{|K|}\int_K p_0(x)\,\mathrm{d}x, \tag{9}$$

$$-\sum_{L\in N(K)} |\sigma_{K,L}|\frac{\phi_L - \phi_K}{d_{K,L}} = |K|\theta_K, \tag{10}$$

$$R\psi|K|\frac{c_K^{m,n+1} - c_K^{m,n}}{\Delta t} - \psi S(v)\sum_{L\in N(K)} |\sigma_{K,L}|\frac{c_L^{m,n+1} - c_K^{m,n+1}}{d_{K,L}} \tag{11}$$

$$- \sum_{L\in N(K)} |\sigma_{K,L}|\frac{\phi_L - \phi_K}{d_{K,L}} c_K^{m,n+1} = -|K|\Big(r(c_K^{m,n}) - p_K^{m,n}\Big) + |\partial K \cap \partial\Omega| f_1,$$

$$R\psi|K|\frac{\mu_K^{m,n+1} - \mu_K^{m,n}}{\Delta t} - \psi S(v)\sum_{L\in N(K)} |\sigma_{K,L}|\frac{\mu_L^{m,n+1} - \mu_K^{m,n+1}}{d_{K,L}} \tag{12}$$

$$+ \sum_{L\in N(K)} |\sigma_{K,L}|\frac{\phi_L - \phi_K}{d_{K,L}}\mu_K^{m,n+1} = -|K| r'(c_K^{m,n+1}) - (R\psi\rho + \theta_K)\mu_K^{m,n}$$

$$+ \partial_c D(c_K^{m,N-n}) + |\partial K \cap \partial\Omega| f_2,$$

$$p_K^{m+1,n} = \left(f'\right)^{-1}(\mu_K^{m+1,n}), \quad \text{for} \quad m \in \mathbb{N}. \tag{13}$$

A proper convergence of the latter scheme ensures the following existence result for the optimal control problem (the proof is postponed to a forthcoming paper):

Theorem 1 *Assume the weak convergence of* (ϕ_K), $(c_K^{m,n})$, $(\mu_K^{m,n})$ *in* $L^2(0,T;H^1(\Omega))$, *the convergence of* $(c_K^{m,n})$ *and* $(\mu_K^{m,n})$ *in* $L^2(\Omega_T)$ *as* $m \to \infty$, $N_T \to \infty$, $\sup_{k\in\mathscr{T}} |K| \to 0$. *Then there exist* $\phi \in L^\infty(0,T;H^1_\Gamma(\Omega))$, $c \in L^2(0,T;H^1(\Omega))$ *and* $\mu \in L^2(0,T;H^1(\Omega))$ *that constitute a weak solution of the system* (4)–(8) *in the following sense: for all* $(\varphi,\Phi) \in H^1_\Gamma(\Omega) \times L^2(0,T;H^1(\Omega))$

$$\int_\Omega \kappa \nabla\phi\nabla\varphi\,dx = \int_\Omega \theta\varphi\,dx,$$

$$R\psi \int_0^T \int_\Omega \partial_t c\Phi + \int_0^T \int_\Omega S(v)\psi\nabla c\cdot\nabla\Phi\,dxdt - \int_0^T \int_\Omega cv\cdot\nabla\Phi\,dxdt$$
$$= -\int_0^T \int_\Omega \big(r(c) + X_{\mathscr{L}}\,p\big)\Phi + \int_0^T \int_\Gamma f_1\Phi\,d\sigma\,dt,$$

$$R\psi \int_0^T \int_\Omega \partial_t \mu\Phi + \int_0^T \int_\Omega S(v)\psi\nabla\mu\cdot\nabla\Phi\,dxdt + \int_0^T \int_\Omega \mu v\cdot\nabla\Phi\,dxdt$$
$$= -\int_0^T \int_\Omega \Big(r'(c) + (R\psi\rho+\theta)\mu - \partial_c D(x,c)\Big)\Phi + \int_0^T \int_\Gamma f_2\Phi\,d\sigma\,dt.$$

3 Numerical Tests

In this part, we present some results obtained with two Python codes of the algorithm described in Sect. 2.3, respectively for the 2D and for the 3D setting. The number of loops (that is the number of increments m in (13)) necessary to arrive at the stopping criterion for the fixed point is denoted by NN. Namely $NN = \min_{m\in\mathbb{N}}\{m+1\}$, such as $||p_h^{m+1} - p_h^m|| \le \varepsilon$ where ε is a given threshold.

We chose the same data characterizing the porous medium structure of the underground as in [8]. But, for showing in a simple way the qualitative realism of the results, we include two parameters, respectively α in the function f and β in D, for weighting respectively the benefit linked with the pollution and the cleaning costs. In the example of the agricultural pollution, α represents the price of cultivated species per tonne and β will take into account the production flow rate of the well and cost of treatment per cubic meter:

$$S_p = 0,\ S_m = 0.01,\ R = 1,\ k = 39.04,\ \psi = 0.3,\ \rho = 0.05,\ \Omega = [0,1]\times[0,1]\times[0,1],$$
$$r(c) = c^2,\ f(p) = \alpha\frac{\ln(p)}{10^{-3}}\mathrm{X}_{\mathscr{L}},\ D(c) = \beta c^2\mathrm{X}_{\mathscr{W}},\ p^0 = 0.005$$
$$\Delta t = 0.01,\ dx = dy = dz = 1/N$$

the set $\mathscr{W}$ corresponding to the water production wells area, with $N = 30$ characterizing the space discretization. The initial spreading p^0 and $\mathscr{W}$ are represented in Fig. 1.

We present two tests: **Test 1** ($\alpha = 1, \beta = 1$) produced the Figs. 2 and 3, **Test 2** ($\alpha = 0.5, \beta = 1$) produced the Figs. 4 and 5.

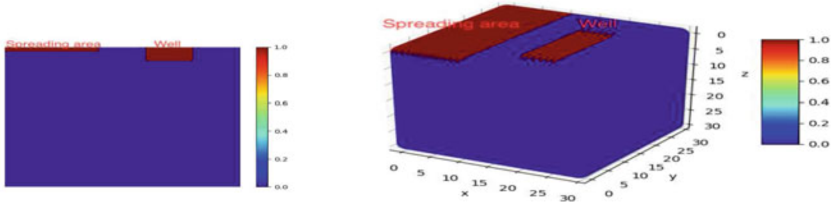


Fig. 1 2D and 3D computational domains with the initial position of the spreading area and the production well $\mathscr{W}$. We note that the vertical section in the middle of the well of the 3D domain relative to the y axis corresponds to the 2D domain

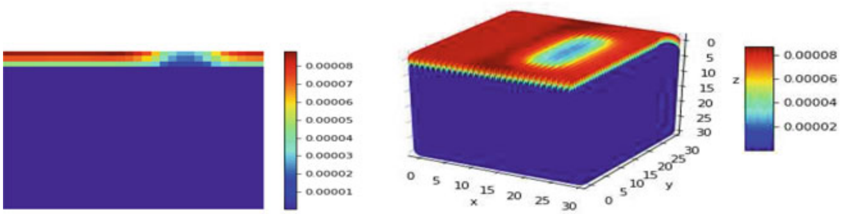


Fig. 2 **Test 1**. The optimal amount of fertilizer to be load at time $T = 1$ day with a stop criterion $\varepsilon = 5.10^{-6}$ in the $2D$ and $3D$ schemes at $NN = 9$

For example for the **Test 1**, we find that with a stop threshold of $\varepsilon = 5.10^{-6}$, our $2D$ and $3D$ programs turned 9 times ($NN = 9$). At the end, we observe that the amount of optimal fertilizer to be spread over the entire field evolves with time. We choose to represent the optimal solution for the fertilizer application with respect to the trade-off modelled by the functional J at time $T = 1$ day. We also note the absence of fertilizer prescribed around the production well both in the 2D and 3D settings see Fig. 2. This fact which is of course explained by the compromise between benefit and cleaning costs. This same trade-off explains that, when in **Test 1** we make the choice to greatly favor the farmer (see $\alpha = 1$ relative to ($\alpha = 0.5$) in **Test 2**), the amount of fertilizer load is larger, with a maximum quantity close to $p_{\max} \simeq 0.00008$ at $NN = 9$, while in **Test 2**, the maximum quantity of fertilizer is reduced to half $p_{\max} \simeq 0.00004$ at $NN = 7$.

Conclusion: From a qualitative point of view, the scheme gives satisfactory results. The Figs. 3 and 5 show the same convergence errors from $NN = 6$ to $NN = 9$ as much as in 2D and 3D. Although the scheme is convergent, the number of iterations needed for the fixed point part is very dependent on the functions appearing in functional J (for instance $NN = 6$ to $NN = 9$ for **Test 1** and **Test 2** respectively). This is where an improvement of the schema must be implemented.

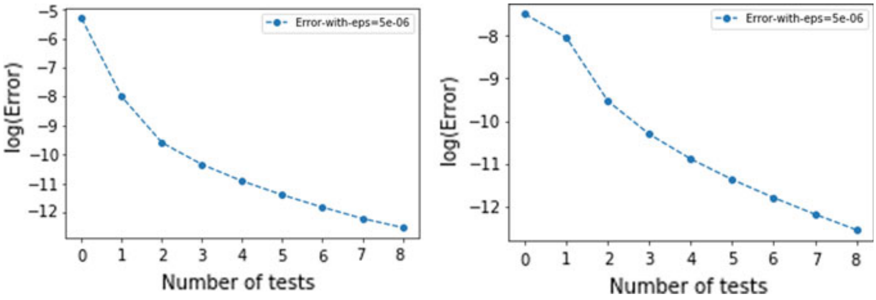


Fig. 3 **Test 1**. Error $||p_h^{m+1} - p_h^m||$ in the $2D$ scheme (left) and $3D$ scheme (right) at $NN = 9$

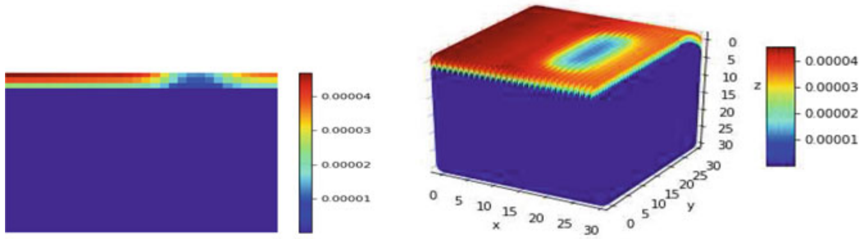


Fig. 4 **Test 2**. The optimal amount of fertilizer to be load at time $T = 1$ day with a stop criterion $\varepsilon = 5.10^{-6}$ in the $2D$ and $3D$ schemes at $NN = 7$

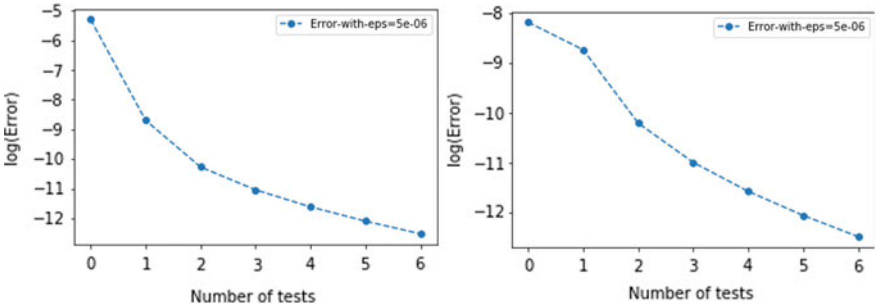


Fig. 5 **Test 2**. Error $||p_h^{m+1} - p_h^m||$ in the $2D$ scheme (left) and $3D$ scheme (right) at $NN = 7$

References

1. Augeraud-Véron, E., Choquet, C., Comte, É.: Optimal control for a groundwater pollution ruled by a convection-diffusion-reaction problem. J. Optim. Theory Appl. **173**(3), 941–966 (2017)
2. Augeraud-Véron, E., Leandri, M.: Optimal pollution control with distributed delays. J. Math. Econ. **55**, 24–32 (2014)
3. Bear, J., Verruijt, A.: Theory and applications of transport in porous media. Reidel, Modeling of groundwater flow and pollution, Dordrecht (1987)
4. Bordenave, P., Bouraoui, F., Gascuel-Odoux, C., Molenat, J., Merot, P.: Décalages temporels entre modifications des pratiques agricoles et diminution de nitrate dans les eaux superficielles.

Actes de colloques-IFREMER 311–333 (2001)

5. Bourgeois, C., Jayet, P.A.: Regulation of relationships between heterogeneous farmers and an aquifer accounting for lag effects. Aust. J. Agric. Resour. Econ. **60**(1), 39–59 (2016)
6. Bradley, A.M.: PDE-constrained optimization and the adjoint method (2010)
7. Carpenter, S.R., Ludwig, D., Brock, W.A.: Management of eutrophication for lakes subject to potentially irreversible change. Ecol. Appl. **9**(3), 751–771 (1999)
8. Comte, E.: Pollution agricole des ressources en eau: approches couplées hydrogéologique et économique (Doctoral dissertation) (2017)
9. Eymard, R., Gallouët, T., Herbin, R.: Finite volume methods. Handb. Numer. Anal. **7**, 713–1018 (2000)
10. Miquel, G.: La qualité de l'eau et l'assainissement en France (rapport). Office parlementaire d'évaluation des choix scientifiques et (2003). technologiques
11. Winkler, R.: A note on the optimal control of Stocks accumulating with a delay. Macroecon. Dyn. **15**(4), 565–578 (2011)

Space-Time Discontinuous Galerkin Methods for Linear Hyperbolic Systems and the Application to the Forward Problem in Seismic Imaging

Willy Dörfler, Christian Wieners, and Daniel Ziegler

Abstract We consider a p-adaptive discontinuous Galerkin method in space and time for linear hyperbolic systems. This is applied to the visco-acoustic wave equation in the formulation as first-order system. The method is applied to the forward problem in seismic imaging, and we study the convergence of the fully adaptive parallel method by the numerical evaluation of measurements in form of seismograms. The method is based on a formulation of Generalized Standard Linear Solids as symmetric Friedrichs system and an inf-sup stable variational Petrov–Galerkin setting. With respect to suitable DG norms the discretization is p-robust inf-sup stable, and the approximation of material parameters can be estimated by a Strang type argument. In order to restrict the computation to the domain of interest, an absorbing boundary layer in included. Numerical results for a benchmark configuration in geophysics are obtained with a p-adaptive method based on a dual-primal error estimator with respect to a goal functional corresponding to the seismic measurements. The linear system is solved in parallel with a space-time multigrid method.

Keywords Linear Hyperbolic systems · Space-time methods · Adaptivity · Wave equation · Parallel solution methods

MSC (2010) 35L50 · 65M60 · 65N30

W. Dörfler · C. Wieners (✉) · D. Ziegler
Karlsruhe Institute of Technology, Englerstraße 2, 76131 Karlsruhe, Germany
e-mail: christian.wieners@kit.edu

W. Dörfler
e-mail: willy.doerfler@kit.edu

D. Ziegler
e-mail: daniel.ziegler@kit.edu

R. Klöfkorn et al. (eds.), *Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples*, Springer Proceedings in Mathematics & Statistics 323,
https://doi.org/10.1007/978-3-030-43651-3_44

1 Linear Hyperbolic Systems in Space and Time

We consider the linear evolution equation in a bounded Lipschitz domain $\Omega \subset \mathbb{R}^d$

$$M\partial_t \mathbf{u}(t) + A\mathbf{u}(t) + D\mathbf{u}(t) = \mathbf{f} \qquad t \in (0, T)\,, \qquad \mathbf{u}(0) = \mathbf{u}_0 \tag{1}$$

subject to initial and boundary conditions corresponding to a hyperbolic system with $\mathbf{u}(t, x) \in \mathbb{R}^m$, $M, D \in \mathrm{L}_\infty(\Omega, \mathbb{R}^{m\times m}_{\mathrm{sym}})$, and the first-order differential operator

$$A\mathbf{y} = \sum_{j=1}^{d} B_j \partial_j \mathbf{y}\,, \qquad B_j = (B_{jkl})_{k,l=1,\dots,m} \in \mathbb{R}^{m\times m}_{\mathrm{sym}}\,, \quad j = 1, \dots, d\,.$$

Defining $B_{\mathbf{n}} = \sum_{j=0}^{d} n_j B_j$ for the normal vector $\mathbf{n} = (n_j)_{j=1,\dots,d}$ a.e. on $\partial\Omega$ we observe

$$\begin{aligned}\big(A\mathbf{y}, \mathbf{z}\big)_\Omega + \big(\mathbf{y}, A\mathbf{z}\big)_\Omega &= \sum_{jkl} B_{jkl}\Big(\big(\partial_j y_k, z_l\big)_\Omega + \big(y_k, \partial_j z_l\big)_\Omega\Big) \\ &= \sum_{jkl} B_{jkl} \int_\Omega \partial_j (y_k z_l)\,\mathrm{d}x = \sum_{jkl} B_{jkl} \int_{\partial\Omega} n_j (y_k z_l)\,\mathrm{d}x \\ &= \big(B_{\mathbf{n}}\mathbf{y}, \mathbf{z}\big)_{\partial\Omega} = \big(\mathbf{y}, B_{\mathbf{n}}\mathbf{z}\big)_{\partial\Omega}\,, \qquad \mathbf{y}, \mathbf{z} \in \mathrm{C}^1(\overline{\Omega}; \mathbb{R}^m)\,.\end{aligned}$$

In order to obtain a well-defined hyperbolic system, boundary conditions have to be imposed. Therefore, $\partial\Omega$ is decomposed into boundary parts Γ_k such that

$$\big(A\mathbf{y}, \mathbf{y}\big)_\Omega = 0 \qquad \text{if} \quad (B_{\mathbf{n}}\mathbf{y})_k = \mathbf{0} \ \text{ on } \ \Gamma_k\,, \quad k = 1, \dots, m\,.$$

For a solution in the space-time cylinder $Q = (0, T) \times \Omega$ we define discrete approximations $\mathbf{u}_h$ and L_h of the solution $\mathbf{u}$ and the operator $L = M\partial_t + A + D$ with suitable ansatz and test spaces $V_h, W_h \subset \mathrm{L}_2(\Omega, \mathbb{R}^m)$ solving

$$\mathbf{u}_h \in \mathbf{u}_0 + V_h\colon \qquad \big(L_h\mathbf{u}_h, \mathbf{w}_h\big)_Q = \big(\mathbf{f}, \mathbf{w}_h\big)_Q\,, \qquad \mathbf{w}_h \in W_h \tag{2}$$

such that the limit $h \to 0$ yields a weak solution of (1) characterized by

$$\big(\mathbf{u}, L^*\mathbf{w}\big)_Q = (\mathbf{f}, \mathbf{w})_Q + \big(\mathbf{u}_0, \mathbf{w}(0)\big)_\Omega - (\mathbf{g}, \mathbf{w})_{(0,T)\times\partial\Omega}\,, \qquad \mathbf{w} \in \mathcal{V}^* \tag{3}$$

with boundary data $\mathbf{g}$, the adjoint operator $L^* = -M\partial_t - A + D$, and the test space

$$\begin{aligned}\mathcal{V}^* = \big\{\mathbf{w} \in \mathrm{C}^1(Q; \mathbb{R}^m) \cap \mathrm{C}^0(\overline{Q}; \mathbb{R}^m)\colon\ & \mathbf{w}(T) = \mathbf{0}\,, \\ & (B_{\mathbf{n}}\mathbf{w})_k = 0 \ \text{ on } \ (0, T) \times \Gamma_k\,,\ k = 1, \dots, m\big\}\,.\end{aligned}$$

This setting applies to wave equations with damping. It can be extended to impedance boundary conditions of Robin type as it is now explained for a special case.

Our basic example is the acoustic wave equation for velocity $\mathbf{v}$ and pressure p

$$\begin{aligned} \rho\,\partial_t\mathbf{v} - \nabla p &= \mathbf{f}_0 && \text{in } (0,T)\times\Omega\,,\\ \partial_t p - \kappa\nabla\cdot\mathbf{v} &= 0 && \text{in } (0,T)\times\Omega\,,\\ \mathbf{v}(0) &= \mathbf{v}^0 && \text{in } \Omega \text{ at } t=0\,,\\ p(0) &= p^0 && \text{in } \Omega \text{ at } t=0\,,\\ p(t) &= p_{\mathrm{D}}(t) && \text{on } \Gamma_{\mathrm{D}} \text{ for } t\in(0,T)\,,\\ \mathbf{n}\cdot\mathbf{v}(t) &= g_{\mathrm{N}}(t) && \text{on } \Gamma_{\mathrm{N}} \text{ for } t\in(0,T)\,,\\ \mathbf{n}\cdot\mathbf{v}(t) + \zeta p(t) &= g_{\mathrm{R}}(t) && \text{on } \Gamma_{\mathrm{R}} \text{ for } t\in(0,T) \end{aligned} \tag{4}$$

with the density ρ, permeability κ, impedance $\zeta = \sqrt{\kappa\rho}$, initial values $(\mathbf{v}^0, p^0)$, and boundary conditions $(p_{\mathrm{D}}, g_{\mathrm{N}}, g_{\mathrm{R}})$ on $\Gamma_{\mathrm{N}}\cup\Gamma_{\mathrm{D}}\cup\Gamma_{\mathrm{R}} = \partial\Omega$.

This extends to visco-acoustics using the retarded material law

$$\partial_t p(t) = \kappa\nabla\cdot\mathbf{v}(t) + \int_0^t \partial_t\kappa(t-s)\nabla\cdot\mathbf{v}(s)\,\mathrm{d}s\,, \qquad \kappa(s) = \sum_{j=1}^r \kappa_j \exp\Big(-\frac{s}{\tau_j}\Big)$$

with permeability $\kappa = \kappa_0 + \kappa_1 + \cdots + \kappa_r$ and relaxation times $\tau_1,\ldots,\tau_r > 0$. This model approximates dispersive wave propagation within a given frequency range, see [5, Chap. 5] for the corresponding visco-elastic Maxwell model for Generalized Standard Linear Solids. Defining

$$p_j(t) = \int_0^t \exp\Big(\frac{s-t}{\tau_j}\Big)\kappa_j\nabla\cdot\mathbf{v}(s)\,\mathrm{d}s\,, \quad j=1,\ldots,r\,, \quad p = p_0+\cdots+p_r$$

results in the first-order system for linear visco-acoustic waves

$$\begin{aligned} \rho\,\partial_t\mathbf{v} - \nabla(p_0+\cdots+p_r) &= \mathbf{f}_0\,,\\ \partial_t p_0 - \kappa_0\nabla\cdot\mathbf{v} \qquad &= 0\,,\\ \partial_t p_j - \kappa_j\nabla\cdot\mathbf{v} + \tau_j^{-1}p_j &= 0\,, \qquad j=1,\ldots,r, \end{aligned}$$

corresponding to $m = d+1+r$ components with

$$\mathbf{u} = \big(\mathbf{v}, p_0,\ldots,p_r\big)^\top, \quad \mathbf{f} = \big(\mathbf{f}_0, 0,\ldots,0\big)^\top, \quad \mathbf{g} = \big(p_{\mathrm{D}}\mathbf{n}, g_{\mathrm{N}}+g_{\mathrm{R}},\ldots,g_{\mathrm{N}}+g_{\mathrm{R}}\big)^\top,$$

$$M\mathbf{u} = \begin{pmatrix} \rho\mathbf{v}\\ \kappa_0^{-1}p_0\\ \vdots\\ \kappa_r^{-1}p_r \end{pmatrix}, \; A\mathbf{u} = -\begin{pmatrix} \nabla p\\ \nabla\cdot\mathbf{v}\\ \vdots\\ \nabla\cdot\mathbf{v} \end{pmatrix}, \; D_{\mathrm{V}}\mathbf{u} = \begin{pmatrix} \mathbf{0}\\ 0\\ (\kappa_1\tau_1)^{-1}p_1\\ \vdots\\ (\kappa_r\tau_r)^{-1}p_r \end{pmatrix}, \; D_{\mathrm{R}}\mathbf{u} = \begin{pmatrix} \mathbf{0}\\ \zeta p_0\\ \zeta p_1\\ \vdots\\ \zeta p_r \end{pmatrix},$$

$$B_{\mathbf{n}}\mathbf{u} = \big(p\mathbf{n}, \mathbf{v}\cdot\mathbf{n},\ldots,\mathbf{v}\cdot\mathbf{n}\big)^\top, \quad \langle D\mathbf{u},\mathbf{w}\rangle = (D_{\mathrm{V}}\mathbf{u},\mathbf{w})_Q + (D_{\mathrm{R}}\mathbf{u},\mathbf{w})_{(0,T)\times\partial\Gamma_{\mathrm{R}}}$$

on the boundaries $\Gamma_j = \Gamma_{\mathrm{D}}$ for $j = 1, \ldots, d$ and $\Gamma_{d+1+j} = \Gamma_{\mathrm{N}} \cup \Gamma_{\mathrm{R}}$ for $j = 0, \ldots, r$, where (2) and (3) are complemented by a boundary integral on $(0, T) \times \partial\Gamma_{\mathrm{R}}$. The boundary data $(p_{\mathrm{D}}, g_{\mathrm{N}}, g_{\mathrm{R}})$ are extended to $\partial\Omega$ by zero.

2 Space-Time Discontinuous Galerkin Methods

We use tensor product space-time cells based on a decomposition in time

$$0 = t_0 < t_1 < \cdots < t_N = T\,, \qquad I_h = (t_0, t_1) \cup \cdots \cup (t_{N-1}, t_N) \subset I = (0, T)\,,$$

combined with decompositions in space $\Omega_{n,h} = \bigcup_{K \in \mathscr{K}_n} K$ into open cells $K \subset \Omega \subset \mathbb{R}^d$ with skeleton $\partial\Omega_{n,h} = \overline{\Omega} \setminus \Omega_{n,h}$ for $n = 1, \ldots, N$. This defines the set of space-time cells

$$\mathscr{R} = \Big\{ R = (t_{n-1}, t_n) \times K \colon K \in \mathscr{K}_n\,,\ n = 1, \ldots, N \Big\}\,,$$

and we obtain a decomposition $Q_h = \bigcup_{R \in \mathscr{R}} R$ of the space-time cylinder $Q = I \times \Omega$. For every $R = (t_{n-1}, t_n) \times K$ we select polynomial degrees $p_R = p_{n,K} \geq 1$ in time and $q_R = q_{n,K} \geq 0$ in space. In (t_{n-1}, t_n) we define the discontinuous space

$$Y_{n,h} = \prod_{K \in \mathscr{K}_n} \mathbb{P}_{q_{n,K}}(K; \mathbb{R}^m) \subset \mathbb{P}(\Omega_{n,h}; \mathbb{R}^m) \subset \mathrm{L}_2(\Omega; \mathbb{R}^m)\,,$$

a positive definite approximation $M_{n,h} \in \mathrm{L}_\infty(\Omega; \mathbb{R}^{m \times m}_{\mathrm{sym}})$ of M, and the projection $\Pi_{n,h} \colon \mathrm{L}_2(\Omega; \mathbb{R}^m) \to Y_{n,h}$ with

$$\big(M_{n,h}\Pi_{n,h}\mathbf{y}, \mathbf{z}_h\big)_\Omega = \big(M_{n,h}\mathbf{y}, \mathbf{z}_h\big)_\Omega\,, \qquad \mathbf{y} \in \mathrm{L}_2(\Omega; \mathbb{R}^m)\,,\ \mathbf{z}_h \in Y_{n,h}\,.$$

For the variational problem (2), we define the discontinuous ansatz and test spaces

$$\begin{aligned} V_h = \Big\{ \mathbf{v}_h \in & \prod_{R=(t_{n-1},t_n)\times K \in \mathscr{R}} \mathbb{P}_{p_R} \otimes \mathbb{P}_{q_R}(K; \mathbb{R}^m) \subset \mathbb{P}(Q_h; \mathbb{R}^m) \colon \\ & \mathbf{v}_h(0) = \mathbf{0} \text{ for } t = 0\,,\ \mathbf{v}_{n,h}(t_{n-1}) = \Pi_{n,h}\mathbf{v}_{n-1,h}(t_{n-1}) \text{ for } n = 2, \ldots, N \Big\}\,, \\ W_h = & \prod_{R=(t_{n-1},t_n)\times K \in \mathscr{R}} \mathbb{P}_{p_R - 1} \otimes \mathbb{P}_{q_R}(K; \mathbb{R}^m) \subset \mathbb{P}(Q_h; \mathbb{R}^m) \subset \mathrm{L}_2(Q; \mathbb{R}^m)\,. \end{aligned}$$

By construction, we have $\partial_t V_h = W_h$ in I_h and $\dim V_h = \dim W_h$.

We define the discrete operator $L_h = M_h\partial_t + A_h + D_h$ by the uniformly positive definite operator $M_h \in \mathrm{L}_\infty(\Omega; \mathbb{R}^{m \times m}_{\mathrm{sym}})$ with $M_h|_{(t_{n-1},t_n)} = M_{n,h}$, a positive semi-definite operator $D_h \in \mathrm{L}_\infty(\Omega; \mathbb{R}^{m\times m}_{\mathrm{sym}})$, and the discontinuous Galerkin approximation [2] with full upwind flux $A_h|_{(t_{n-1},t_n)} = A_{n,h} \in \mathscr{L}\big(Y_{n,h}, Y_{n,h}\big)$ given by

$$\begin{aligned}\big(A_{n,h}\mathbf{y}_h, \mathbf{z}_h\big)_\Omega &= -\big(\nabla\cdot\mathbf{v}_{h,K}, q_{h,K}\big)_{\Omega_{h,h}} - \big(\nabla p_{h,K}, \mathbf{w}_{h,K}\big)_{\Omega_{h,h}}\\ &\quad - \sum_{K\in\mathscr{K}_{n,h}}\sum_{F\in\mathscr{F}_K}\frac{1}{\zeta_K+\zeta_{K_F}}\big([p_h]_{K,F} + \zeta_{K_F}\mathbf{n}_K\cdot[\mathbf{v}_h]_{K,F}\,,\, q_{K,h} + \zeta_K\mathbf{n}_K\cdot\mathbf{w}_{h,K}\big)_F\end{aligned}$$

for $\mathbf{y}_h = (\mathbf{v}_h, p_{0,h}, \dots, p_{r,h})$, $\mathbf{z}_h = (\mathbf{w}_h, q_{0,h}, \dots, q_{r,h}) \in Y_{n,h}$ with $p_h = p_{0,h} + \cdots + p_{r,h}$, $q_h = q_{0,h} + \cdots + q_{r,h}$, where $\zeta_K = \sqrt{(\kappa\rho)|_K}$ is the impedance and $\mathscr{F}_K \subset \partial K$ are the set of faces. For inner faces $F \subset \Omega$ let K_F be the neighboring cell such that $\overline{F} = \partial K \cap \partial K_F$, and we define $[p_h]_{K,F} = p_{K_F,h} - p_{K,h}$ and $[\mathbf{v}_h]_{K,F} = \mathbf{v}_{K_F,h} - \mathbf{v}_{K,h}$. On boundary faces $F \subset \partial\Omega$, we set $\zeta_{K_F} = \zeta_K$, for $F \subset \Gamma_{\mathrm{D}}$ we set $\mathbf{n}_K\cdot[\mathbf{v}_h]_{K,F} = 0$ and $[p_h]_{K,F} = -2p_h$, for $F \subset \Gamma_{\mathrm{N}}$ we set $\mathbf{n}_K\cdot[\mathbf{v}_h]_{K,F} = -2\mathbf{n}_K\cdot\mathbf{v}_h$ and $[p_h]_{K,F} = 0$, and for $F \subset \Gamma_{\mathrm{R}}$ we set $\mathbf{n}_K\cdot[\mathbf{v}_h]_{K,F} = -2\mathbf{n}_K\cdot\mathbf{v}_h$ and $[p_h]_{K,F} = -2p_h$.

Inf-sup stability of the Petrov–Galerkin approximation (2) can be provided for fixed polynomial degrees $p_R \equiv p$, $q_R \equiv p$ [1, Lemma 3], and a Strang-type argument as in [4, Theorem 10] yields convergence for consistent data. If the solution is sufficiently regular, convergence of order $(\triangle t)^p + (\triangle x)^q$ in the graph norm of the operator is achieved [1, Theorem 1]. For a discontinuous Galerkin formulation also in time, inf-sup stability is also established for the fully adaptive case in [8, Theorem 3.1].

3 Application to a Benchmark Configuration in Geophysics

The discretization method is evaluated for an application to the Marmousi benchmark, see Fig. 1 for the configuration and [8, Chap. 5.2] for the parameters.

In order to avoid artificial reflections from the boundary of the computational domain, an absorbing boundary layer of width $\ell > 0$ is included. There we select a reduced velocity by scaling the material parameters ρ and κ_j depending on the distance $s < \ell$ to the boundary with an increasing function γ with $\gamma(s) \in (0, 1)$ for $s < \ell$ and $\gamma(s) = 1$ for $s \geq \ell$, so that the impedance remains constant, i.e.,

$$\tilde{c} = \gamma(s)c\,, \qquad \tilde{\rho} = \rho/\gamma(s)\,, \qquad \tilde{\kappa}_j = \gamma(s)\kappa_j\,. \tag{5}$$

A wave signal is initiated by a local source at $(t_{\mathrm{S}}, x_{\mathrm{S}}) \in Q$, and the reflections of the wave at material interfaces are measured at receivers $x_{\mathrm{R}_0}, \dots, x_{\mathrm{R}_m} \in \Omega$. The overall adaptive space-time solution method is realized as follows:

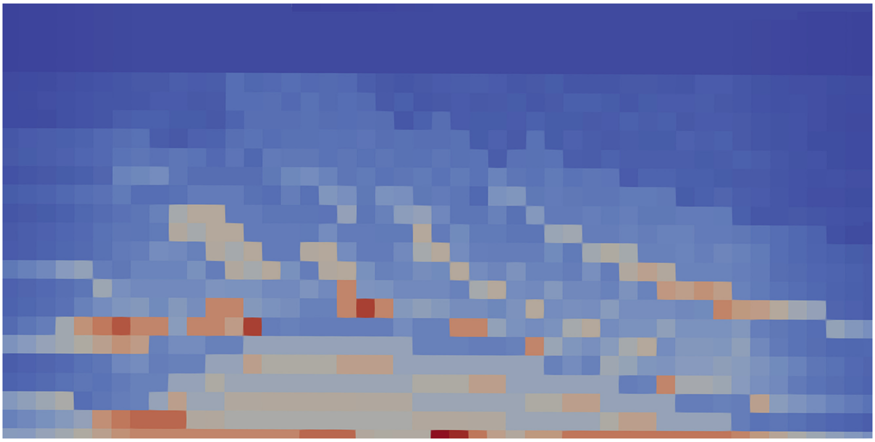


Fig. 1 Distribution of the wave velocity $c_\mathrm{p} = \sqrt{\kappa/\rho}$ for the Marmousi benchmark [7] with absorbing boundary layers (5). We use the domain $\Omega = (0, 6) \times (-3, 0)$ [km^2], homogeneous Neumann boundary conditions on the surface ($x_2 = 0$), and otherwise mixed boundary conditions with $g_\mathrm{R} = 0$. Absorbing boundary layers are included at the left and the right side for $x_1 < 1$ and $x_1 > 5$

1. We start with a coarse initial low order approximation choosing $p_R = q_R = 1$.
2. After solving the forward problem, the dual solution is approximated with respect to a goal functional measuring the error with respect to the receiver positions.
3. The error is estimated by a dual-weighted error indicator with a rough estimate for the interpolation error of the dual solution.
4. The polynomial degrees are refined and derefined with respect to the local weighted residuals of the error indicator.
5. The overall scheme is repeated up to a sufficiently small value of the error indicator.

The resulting seismograms are presented in Fig. 2 and for a single receiver in Fig. 3; time slices of the solution are shown in Fig. 4.

More details of the adaptive strategy and the parallel multigrid preconditioner using coarsening first in space and then in time are presented in [1, 2]. Moreover, the method extends to the full wave form inversion, where the full space-time solution is required for the adjoint problem backward in time, see [3].

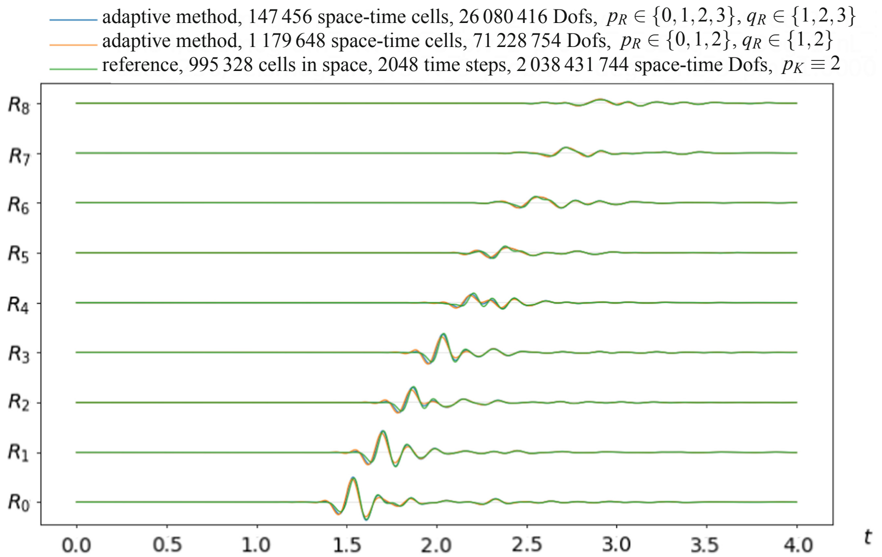


Fig. 2 Seismograms for the adaptive results compared with a reference solution on a very fine mesh computed with a time stepping scheme [6]. The wave is initiated by an impulse of wavelet form in space and time centered at $(t_S, x_S) = (0.15, (1, -0.25))$, the pressure is evaluated at the receiver positions between $x_{R,j} = (3, -0.25), \dots, (5, -0.25)$. The differences of the solutions are illustrated in more detail in Fig. 3

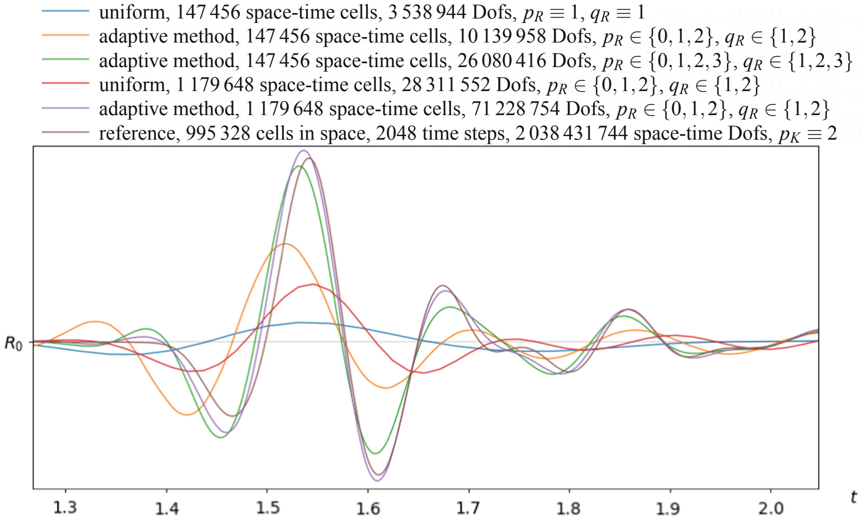


Fig. 3 Seismograms at the first receiver for different p-adaptive meshes, starting with the coarse problem results with a rough approximation and comparison with the reference solution. We observe that the adaptive results on level 3 (with maximal polynomial degree $p_R = 3$) and on level 4 (with maximal polynomial degree $p_R = 2$) are close to the reference solution with only a very small fraction of space-time degrees of freedom

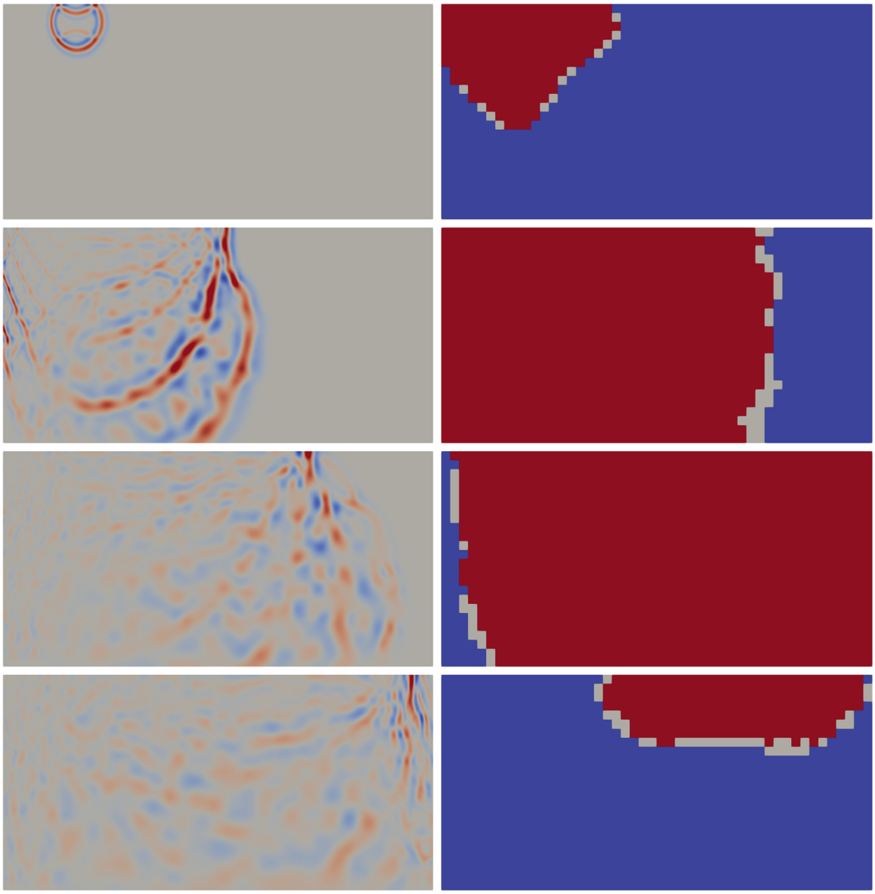


Fig. 4 Pressure distribution for the Marmousi benchmark and distribution of the polynomial degrees $q_R \in \{0, 1, 2, 3\}$ of the adaptive method at time $t = 0.4, 1.6, 2.4, 3.6$ [s]

Acknowledgements This work is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 258734477 – SFB 1173.

References

1. Dörfler, W., Findeisen, S., Wieners, C.: Space-time discontinuous Galerkin discretizations for linear first-order hyperbolic evolution systems. Comput. Methods Appl. Math. **16**(3), 409–428 (2016)
2. Dörfler, W., Findeisen, S., Wieners, C., Ziegler, D.: Parallel adaptive discontinuous Galerkin discretizations in space and time for linear elastic and acoustic waves. In: Langer, U., Steinbach, O. (eds.) Space-Time Methods. Applications to Partial Differential Equations, Radon Series on Computational and Applied Mathematics, vol. 25, pp. 61–88. Walter de Gruyter (2019)

3. Ernesti, J.: Space-time methods for acoustic waves with applications to full waveform inversion. Ph.D. thesis, Karlsruher Institut für Technologie (KIT) (2018)
4. Ernesti, J., Wieners, C.: Space-time discontinuous Petrov–Galerkin methods for linear wave equations in heterogeneous media. Comput. Methods Appl. Math. **19**(3), 465–481 (2019)
5. Fichtner, A.: Full Seismic Waveform Modelling and Inversion. Advances in Geophysical and Environmental Mechanics and Mathematics. Springer, Berlin Heidelberg (2011)
6. Hochbruck, M., Pazur, T., Schulz, A., Thawinan, E., Wieners, C.: Efficient time integration for discontinuous Galerkin approximations of linear wave equations. ZAMM Z. Angew. Math. Mech. **95**, 237–259 (2015)
7. Versteeg, R.: The Marmousi experience: velocity model determination on a synthetic complex data set. Lead. Edge **13**(9), 927–936 (1994)
8. Ziegler, D.: A parallel and adaptive space-time discontinuous Galerkin method for visco-elastic and visco-acoustic waves. Ph.D. thesis, Karlsruhe Institute of Technology (KIT) (2019)

A Hybrid Discontinuous Galerkin Method for Transport Equations on Networks

Herbert Egger and Nora Philippi

Abstract We discuss the mathematical modeling and numerical discretization of transport problems on one-dimensional networks. Suitable coupling conditions are derived that guarantee conservation of mass across network junctions and dissipation of a mathematical energy which allows us to prove existence of unique solutions. We then consider the space discretization by a hybrid discontinuous Galerkin method which provides a suitable upwind mechanism to handle the transport problem and allows to incorporate the coupling conditions in a natural manner. In addition, the method inherits mass conservation and stability of the continuous problem. Order optimal convergence rates are established and illustrated by numerical tests.

Keywords Hybrid discontinuous Galerkin methods · Transport problems · Partial differential equations on networks

MSC (2010) 65M08 · 65N08 · 35Q30

1 Introduction

Partial differential equations on networks arise in various applications including traffic flow, gas or water supply networks, and elastic multi-structures; see [6, 9, 10] for mathematical background, further applications, and references. In this paper, we study scalar conservation laws on one dimensional network structures describing, e.g., the transport of a chemical substance in a flow through a network of pipes. A linear advection equation is used to model the transport within the pipes and appropriate coupling conditions are formulated to describe the mixing of flows and the conservation of mass at network junctions. For the semi-discretization in space, we

H. Egger (✉) · N. Philippi
TU Darmstadt, Karolinenplatz 5, 64289 Darmstadt, Germany
e-mail: egger@mathematik.tu-darmstadt.de

N. Philippi
e-mail: philippi@mathematik.tu-darmstadt.de

R. Klöfkorn et al. (eds.), *Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples*, Springer Proceedings in Mathematics & Statistics 323,
https://doi.org/10.1007/978-3-030-43651-3_45

consider a hybrid discontinuous Galerkin method which turns out to be particularly well-suited for dealing with the hyperbolic nature of the problem as well as the coupling conditions at network junctions. Stability and conservation of the semi-discrete scheme and order optimal error estimates are established.

The rest of the paper is structured as follows: In Sect. 2, we introduce the basic notation and then give a complete formulation of the considered problem. A particular choice is made for the coupling conditions which allows us to prove conservation of mass and stability of the overall system. In Sect. 3, we introduce the discretization and establish conservation, discrete stability, and error estimates. Some numerical tests are presented in Sect. 4 for illustration of our results.

2 Notation and Problem Formulation

Following the notation of [3], the topology of the pipe network is described by a finite, directed, and connected graph $\mathscr{G} = (\mathscr{V}, \mathscr{E})$ with vertex set $\mathscr{V} = \{v_1, \dots, v_n\}$ and set of edges $\mathscr{E} = \{e_1, \dots, e_m\} \subset \mathscr{V} \times \mathscr{V}$. For any vertex $v \in \mathscr{V}$, let $\mathscr{E}(v)$ denote the set of edges having v as a vertex. We then distinguish between inner vertices, i.e. pipe junctions, $\mathscr{V}_0 = \{v \in \mathscr{V} : |\mathscr{E}(v)| \geq 2\}$ and boundary vertices $\mathscr{V}_\partial = \mathscr{V} \backslash \mathscr{V}_0$. For any edge $e = (v_i, v_j)$, we define $n^e(v_i) = -1$ and $n^e(v_j) = 1$ to indicate the start and the end point of the edge, and we set $n^e(v) = 0$ if $v \notin e$. We further identify e with the interval $(0, \ell^e)$ of positive length ℓ^e and denote by $L^2(e) = L^2(0, \ell^e)$ the space of square integrable functions on the edge $e \in \mathscr{E}$, and by

$$L^2(\mathscr{E}) = L^2(e_1) \times \cdots \times L^2(e_m) = \{u : u^e \in L^2(e) \text{ for all } e \in \mathscr{E}\}$$

the corresponding space on the network. Here and below, $u^e = u|_e$ denotes the restriction of a function defined over the network to a single edge e. We use

$$\|u\|^2_{L^2(\mathscr{E})} = \sum_{e \in \mathscr{E}} \|u^e\|^2_{L^2(e)} \qquad \text{and} \qquad (u, w)_{L^2(\mathscr{E})} = \sum_{e \in \mathscr{E}} (u^e, w^e)_{L^2(e)}$$

to denote the natural norm and scalar product of $L^2(\mathscr{E})$ and define by

$$H^s_{pw}(\mathscr{E}) = \{u \in L^2(\mathscr{E}) : u^e \in H^s(e) \text{ for all } e \in \mathscr{E}\}$$

the broken Sobolev spaces which are equipped with the canonical norms

$$\|u\|^2_{H^s_{pw}(\mathscr{E})} = \sum_{e \in \mathscr{E}} \|u^e\|^2_{H^s(e)}.$$

Let us note that $H^0_{pw}(\mathscr{E}) = L^2(\mathscr{E})$. For $s > 1/2$ the functions $u \in H^s_{pw}(\mathscr{E})$ are continuous along edges $e \in \mathscr{E}$ but may be discontinuous across junctions $v \in \mathscr{V}_0$.

On every edge (pipe) e of the network, the transport shall be described by

$$a^e(x)\partial_t u^e(x,t) + \partial_x(b^e u^e(x,t)) = 0, \qquad x \in e,\ t > 0, \tag{1}$$
$$u^e(x,0) = u_0^e(x), \qquad x \in e. \tag{2}$$

Here $u^e = u|_e$ is the concentration of the substance in pipe e, $a^e = a|_e$ represents the cross sectional area of the pipe, and $b^e = b|_e$ is the given volume flow rate.

Assumption 1 We assume $a,\ b \in H^1_{pw}(\mathscr{E})$ with $a^e(x) \geq a_0 > 0$ and b^e constant on every edge. Moreover, we require flow conservation at junctions, i.e.,

$$\sum_{e \in \mathscr{E}(v)} b^e n^e(v) = 0 \qquad \text{for all } v \in \mathscr{V}_0. \tag{C}$$

The conditions on b characterize an incompressible background flow. By the above assumption, we can associate a unique flow direction to every edge e. We then define for every vertex $v \in \mathscr{V}$ the sets of edges $\mathscr{E}^{\text{in}}(v) = \{e \in \mathscr{E}(v) : b^e n^e(v) > 0\}$ and $\mathscr{E}^{\text{out}}(v) = \{e \in \mathscr{E}(v) : b^e n^e(v) < 0\}$ having flow into or out of the vertex. We further split $\mathscr{V}_\partial$ into the sets of inflow and outflow vertices $\mathscr{V}_\partial^{\text{in}} = \{v \in \mathscr{V}_\partial : b^e n^e(v) < 0 \text{ for } e \in \mathscr{E}(v)\}$ and $\mathscr{V}_\partial^{\text{out}} = \{v \in \mathscr{V}_\partial : b^e n^e(v) > 0 \text{ for } e \in \mathscr{E}(v)\}$. The local transport problems (1)–(2) are then complemented by coupling and boundary conditions

$$u^e(v,t) = \hat{u}^v(t) \quad \text{for all } v \in \mathscr{V},\ e \in \mathscr{E}^{\text{out}}(v),\ t > 0 \tag{3}$$

with auxiliary vertex values $\hat{u}^v$ defined for $t \geq 0$ by the relations

$$\hat{u}^v(t) = g^v(t), \tag{4}$$

for all inflow vertices $v \in \mathscr{V}_\partial^{\text{in}}$. On the remaining vertices $v \in \mathscr{V}_0 \cup \mathscr{V}_\partial^{\text{out}}$, we set

$$\sum_{e \in \mathscr{E}^{\text{in}}(v)} b^e n^e(v)\hat{u}^v(t) = \sum_{e \in \mathscr{E}^{\text{in}}(v)} b^e n^e(v) u^e(v,t). \tag{5}$$

For convenience of notation, we write $\hat{u} = (\hat{u}^v)_{v \in \mathscr{V}}$ in the sequel.

Remark 1 From Assumption 1, we deduce that $\sum_{e \in \mathscr{E}^{\text{in}}(v)} b^e n^e(v) > 0$, so that $\hat{u}^v$ is well-defined for all $v \in \mathscr{V}$ and can thus be eliminated using conditions (4) and (5). Furthermore, the value $\hat{u}^v$ is a convex combination, i.e., a mixture, of the concentrations $u^e(v)$ in the flows entering the junction v. Using condition (C), one can also see that the mass at inner vertices is conserved, more precisely

$$\sum_{e \in \mathscr{E}^{\text{out}}(v)} b^e n^e(v)\hat{u}^v(t) = -\sum_{e \in \mathscr{E}^{\text{in}}(v)} b^e n^e(v) u^e(v,t) \quad \text{for all } v \in \mathscr{V}_0. \tag{6}$$

The transport problem on networks is now fully described by the system (1)–(5). The number and type of coupling and boundary conditions turns out to be appropriate to guarantee stability and well-posedness of the problem and to ensure conservation of mass across network junctions.

Theorem 1 *Let Assumption 1 hold and* $t_{\max} > 0$. *Then for any* $g \in W^{2,1}(0, t_{\max}; \mathscr{V}_\partial^{in})$ *and* $u_0 \in H^1_{pw}(\mathscr{E})$, *satisfying* (3)–(5) *for* $t = 0$ *with some* $\hat{u}(0) = \hat{u}_0 \in \mathbb{R}^{|\mathscr{V}|}$, *the problem* (1)–(5) *has a unique solution* $u \in C^1([0, t_{\max}]; L^2(\mathscr{E})) \cap C^0([0, t_{\max}]; H^1_{pw}(\mathscr{E}))$ *and* $\hat{u} \in C^0([0, t_{\max}]; \mathbb{R}^{|\mathscr{V}|})$. *Moreover, the conservation property*

$$\frac{d}{dt}\int_{\mathscr{E}} a(x)u(x,t)\,dx = -\sum_{v\in\mathscr{V}_\partial} b^e n^e(v) u^e(v,t)$$

holds as well as the energy identity

$$\begin{aligned}\frac{d}{dt}\|a^{1/2}u\|^2_{L^2(\mathscr{E})} = &-\sum_{v\in\mathscr{V}_\partial^{out}} |b^e n^e(v)||u^e(v)|^2 + \sum_{v\in\mathscr{V}_\partial^{in}} |b^e n^e(v)||g^v|^2 \\ &-\sum_{v\in\mathscr{V}_0}\sum_{e\in\mathscr{E}^{in}(v)} |b^e n^e(v)||u^e(v) - \hat{u}^v|^2.\end{aligned}$$

Proof The energy identity can be derived directly from (1)–(5) and establishes stability of the evolution problem. Existence of a unique solution then follows from the Lumer-Phillips theorem and semigroup theory [5, 11]; a detailed proof can be found in [12]. Related results can also be found in [2, 3, 8, 10]. □

Remark 2 Let us note that for junctions with more than two inflow pipes, the last term in the energy estimate does in general not vanish and represents physical dissipation, i.e., loss of information, due to mixing.

3 A Hybrid Discontinuous Galerkin Method

We now formulate a discontinuous Galerkin method for the semi-discretization of problem (1)–(5); see [1, 7] for a general introduction. Hybridization introduces additional unknowns $\hat{u}$ at the grid points of the mesh which play a similar role as the auxiliary mixing values in the coupling conditions (3). The spatial grid is defined by

$$\mathscr{T}_h = \{T_i^e = (x_{i-1}^e, x_i^e) : i = 1, \ldots, M^e,\ x_0^e = 0,\ x_{M^e}^e = \ell^e,\ e \in \mathscr{E}\}$$

with local and global mesh size denoted by $h_i^e = x_i^e - x_{i-1}^e$ and $h = \max h_i^e$. As approximation spaces for the concentration field, we choose

$$W_h = \{w_h \in L^2(\mathscr{E}) : w_h|_T \in P_k(T) \text{ for all } T \in \mathscr{T}_h\},$$

i.e., spaces of piecewise polynomials of degree $\leq k$, which may formally take multiple values at grid points. We introduce grid dependent scalar products

$$(u, w)_{\mathscr{T}_h} = \sum_{T \in \mathscr{T}_h} (u, w)_{L^2(T)}, \qquad \langle u, w \rangle_{\partial \mathscr{T}_h} = \sum_{T \in \mathscr{T}_h} u(x_{i-1}) w(x_{i-1}) + u(x_i) w(x_i),$$

where $T = (x_{i-1}, x_i)$, and associated norms $\|w\|^2_{\mathscr{T}_h} = (w, w)_{\mathscr{T}_h}$ and $\|w\|^2_{\partial\mathscr{T}_h} = \langle w, w \rangle_{\partial \mathscr{T}_h}$. The corresponding broken Sobolev spaces over the mesh $\mathscr{T}_h$ are denoted by

$$H^s_{pw}(\mathscr{T}_h) = \{w \in L^2(\mathscr{E}) : w|_T \in H^s(T) \text{ for all } T \in \mathscr{T}_h\}.$$

We further introduce the spaces of hybrid variables

$$\hat{W}_h = \mathbb{R}^{\hat{M}} \quad \text{and} \quad \hat{W}^0_h = \{\hat{w} \in \hat{W}_h : \hat{w}^v = 0 \text{ for all } v \in \mathscr{V}^{\text{in}}_\partial\}$$

with $\hat{M} = |\mathscr{V}| + \sum_{e \in \mathscr{E}} (M^e - 1)$ denoting the total number of grid points. Note that grid points associated to the same junction $v \in \mathscr{V}$ are identified. For the numerical approximation of (1)–(5), we then consider the following semi-discrete scheme.

Problem 1 Find $u_h \in C^1([0, t_{\max}]; W_h)$ and $\hat{u}_h \in C([0, t_{\max}]; \hat{W}_h)$ such that $(u_h(0), w_h)_{\mathscr{T}_h} = (u_0, w_h)_{\mathscr{T}_h}$ for all $w_h \in W_h$, and such that $\hat{u}^v_h(t) = g^v(t)$ for all $v \in \mathscr{V}^{\text{in}}_\partial$ as well as

$$(a \partial_t u_h(t), w_h)_{\mathscr{T}_h} + b_h(u_h(t), \hat{u}_h(t); w_h, \hat{w}_h) = 0 \tag{7}$$

holds for all $w_h \in W_h$ and $\hat{w}_h \in \hat{W}^0_h$ and all $0 \le t \le t_{\max}$, with bilinear form

$$b_h(u_h, \hat{u}_h; w_h, \hat{w}_h) = -(b u_h, \partial_x w_h)_{\mathscr{T}_h} + \langle bn\, u^*_h, w_h - \hat{w}_h \rangle_{\partial \mathscr{T}_h} + \langle bn\, \hat{u}_h, \hat{w}_h \rangle_{\mathscr{V}^{\text{out}}_\partial}, \tag{8}$$

and upwind value $bn\, u^*_h = \max(bn, 0) u_h + \min(bn, 0) \hat{u}_h$ in flow direction.

As noted in [4], the hybrid variable $\hat{u}_h$ can be eliminated from the system resulting in a standard discontinuous Galerkin discretization with upwind fluxes. At network junctions $v \in \mathscr{V}_0$, the hybrid variable $\hat{u}^v_h$ is determined by a discrete version of the coupling condition (5). Let us summarize some basic properties of the scheme.

Lemma 1 *The bilinear form b_h is semi-elliptic on the discrete spaces, i.e.,*

$$b_h(w_h, \hat{w}_h; w_h, \hat{w}_h) = \frac{1}{2} \big\| |b|^{1/2} (w_h - \hat{w}_h) \big\|^2_{\partial \mathscr{T}_h} + \frac{1}{2} \big\| |b|^{1/2} \hat{w}_h \big\|^2_{\mathscr{V}^{out}_\partial} \ \forall w_h \in W_h,\ \hat{w}_h \in \hat{W}^0_h.$$

As a consequence, Problem 1 is uniquely solvable. Moreover, the solution satisfies

$$\frac{d}{dt} \int_{\mathscr{E}} a(x) u_h(x, t)\, dx = -\sum_{v \in \mathscr{V}_\partial} b^e n^e(v) u^e_h(v, t)$$

for all $0 \le t \le t_{\max}$, as well as the discrete energy identity

$$\frac{d}{dt} \|a^{1/2} u_h\|^2_{\mathscr{T}_h} + \big\| |b|^{1/2} (u_h - \hat{u}_h) \big\|^2_{\partial \mathscr{T}_h} + \big\| |b|^{1/2} \hat{u}_h \big\|^2_{\mathscr{V}^{out}_\partial} = \big\| |b|^{1/2} g \big\|^2_{\mathscr{V}^{in}_\partial}.$$

Finally, let $(u, \hat{u})$ *be a sufficiently regular solution of* (1)–(5) *and set* $\hat{u}^{x_i} = u(x_i)$ *at grid points* x_i *in the interior of the edges. Then*

$$(a\partial_t u(t), w_h)_{\mathscr{T}_h} + b_h(u(t), \hat{u}(t); w_h, \hat{w}_h) = 0$$

for all $w_h \in W_h$, $\hat{w}_h \in \hat{W}_h^0$ *and all* $0 \le t \le t_{\max}$*, i.e., the method is consistent.*

Proof The semi-ellipticity of b_h follows by standard arguments; see e.g. [1, 3]. As a consequence of this identity and Assumption 1, $\hat{u}_h$ can be eliminated algebraically and the discrete problem can be turned into a linear ordinary differential equation. Existence of a unique solution then follows by the Picard-Lindelöf theorem. The conservation property and the energy identity follow by appropriate testing. □

Remark 3 The discretization inherits most of the properties from the continuous problem. The dissipation terms in the energy estimate are partly due to possible jumps across network junctions, which are present also on the continuous level, and partly due to jumps at interior vertices, which are caused by numerical dissipation due to the upwind mechanism in the discontinuous Galerkin method.

We are now in the position to establish order optimal a-priori error estimates.

Theorem 2 *Let* $(u, \hat{u})$ *denote a sufficiently regular solution of the system* (1)–(5) *and let* $(u_h, \hat{u}_h)$ *be the semi-discrete solution defined by Problem 1. Then*

$$\|u - u_h\|_{L^\infty(0,t_{\max};L^2(\mathscr{E}))} \le C_{\max} h^{k+1} \|u\|_{W^{1,1}(0,t_{\max};H^{k+1}_{pw}(\mathscr{T}_h))},$$

with constant $C_{\max}$ *only depending on the bounds for the coefficient a and* $t_{\max}$.

Proof As usual, the proof is based on an error splitting

$$\|u - u_h\|_{L^\infty(0,t_{\max};L^2(\mathscr{E}))} \le \|\eta_h\|_{L^\infty(0,t_{\max};L^2(\mathscr{E}))} + \|\epsilon_h\|_{L^\infty(0,t_{\max};L^2(\mathscr{E}))}$$

into projection error $\eta_h = u - \pi_h u$ and discrete error $\epsilon_h = \pi_h u - u_h$. Similar to [13], we use a particular projection $\pi_h : H^1_{pw}(\mathscr{E}) \to W_h$ defined for any $T_i^e \in \mathscr{T}_h$ by

$$\pi_h w(x^e_{\text{i,out}}) = w(x^e_{\text{i,out}}) \quad \text{and} \quad \int_{T_i^e} (w - \pi_h w) p \, dx = 0 \quad \forall p \in \mathscr{P}_{k-1}(T_i^e).$$

Here $x^e_{i,out}$ is the outflow point of the element $T_i^e = (x^e_{i-1}, x^e_i)$, i.e., $x^e_{i,out} = x^e_i$ if $b^e > 0$ and $x^e_{i,out} = x^e_{i-1}$ otherwise. By standard estimates for this projection, we obtain

$$\|\eta_h\|_{L^\infty(0,t_{\max};L^2(\mathscr{E}))} \le ch^{k+1}\|u\|_{L^\infty(0,t_{\max};H^{k+1}_{pw}(\mathscr{T}_h))} \le Ch^{k+1}\|u\|_{W^{1,1}(0,t_{\max};H^{k+1}_{pw}(\mathscr{T}_h))},$$

where we used the continuous embedding of $W^{1,1}$ into L^∞ for the second step. Further define $\hat{\pi}_h u^v = \hat{u}^v$ for vertices $v \in \mathscr{V}$ of the network and $\hat{\pi}_h u^{x_i^e} = u(x_i^e)$ for interior grid points x_i^e on edge e. We abbreviate $\hat{\epsilon}_h = \hat{\pi}_h u - \hat{u}_h$, $\hat{\eta}_h = \hat{u} - \hat{\pi}_h u$, and denote by η_h^* the upwind value as in the definition of the method. Note that $\hat{\eta}_h = 0$ and $\eta_h^* = 0$ by construction. Using consistency of the discrete problem, we get

$$\begin{aligned}(a\partial_t \epsilon_h(t), w_h)_{\mathcal{T}_h} &+ b_h(\epsilon_h(t), \hat{\epsilon}_h(t); w_h, \hat{w}_h) \\ &= (a\partial_t \eta_h(t), w_h)_{\mathcal{T}_h} + b_h(\eta_h(t), \hat{\eta}_h(t); w_h, \hat{w}_h)\end{aligned}$$

for all $w_h \in W_h$, $\hat{w}_h \in \hat{W}_h^0$, and $0 \le t \le t_{\max}$. Testing with $w_h = \epsilon_h$ and $\hat{w}_h = \hat{\epsilon}_h$ yields

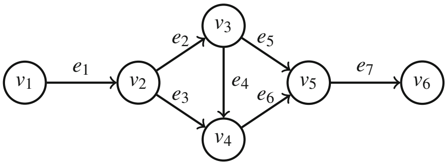
$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\|a^{1/2}\epsilon_h\|^2_{\mathcal{T}_h} &= \underbrace{-b_h(\epsilon_h, \hat{\epsilon}_h; \epsilon_h, \hat{\epsilon}_h)}_{\le 0} + (a\partial_t \eta_h, \epsilon_h)_{\mathcal{T}_h} + b_h(\eta_h, \hat{\eta}_h; \epsilon_h, \hat{\epsilon}_h) \\ &\le (a\partial_t \eta_h, \epsilon_h)_{\mathcal{T}_h} - \underbrace{(b\eta_h, \partial_x \epsilon_h)_{\mathcal{T}_h}}_{=0,\ \text{(proj.)}} + \langle bn \underbrace{\eta_h^*}_{=0}, \epsilon_h - \hat{\epsilon}_h\rangle_{\partial\mathcal{T}_h} + \langle bn \underbrace{\hat{\eta}_h}_{=0}, \hat{\epsilon}_h\rangle_{\mathcal{V}_\partial^{\text{out}}} \\ &\le c\|\partial_t \eta_h\|_{\mathcal{T}_h}\|a^{1/2}\epsilon_h\|_{\mathcal{T}_h}.\end{aligned}$$

Note that the constant c only depends on the bound for a. Integrating this inequality in time, using $\epsilon_h(0) = 0$, taking the maximum over all $0 \le t \le t_{\max}$ on the left hand side, and using Hölder and Young inequalities on the right hand side then allows to bound the $L^\infty(0, t_{\max}; L^2(\mathcal{E}))$ norm of the discrete error ϵ_h by the $W^{1,1}(0, t_{\max}; L^2(\mathcal{E}))$ norm of the projection error η_h. □

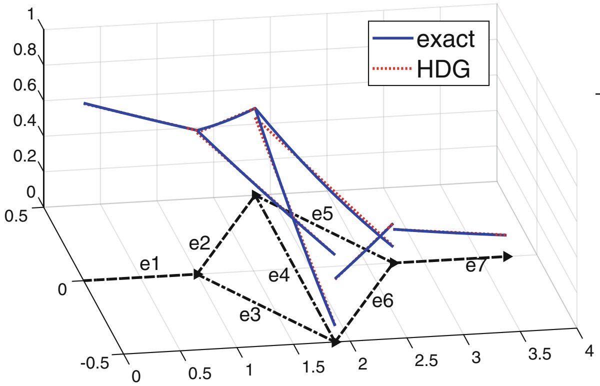
Remark 4 Using the semi-ellipticity of the discrete bilinear form, it is possible to obtain similar bounds also for the error $\hat{\epsilon}_h = \hat{\pi}_h u - \hat{u}_h = \hat{u} - \hat{u}_h$ at the grid points. A sub-sequent time discretization, e.g., by implicit Runge-Kutta methods, can also be analyzed with standard arguments; see [1, 13]. Since the problem is one-dimensional, the computational overhead of an implicit time integration scheme is negligible.

4 Numerical Tests

For our numerical tests, we consider the following network topology.



We set $\ell^e = 1$ and $a^e = 1$ for all edges, and choose $b^{e_1} = 2$, $b^{e_2} = b^{e_3} = 1$, $b^{e_4} = b^{e_5} = 0.5$, $b^{e_6} = 1.5$, and $b^{e_7} = 2$ for which condition (C) is satisfied. We further choose $u_0^e = 0$ as initial conditions and $g^{v_1}(t) = t^2/25$ as inflow boundary condition, such that the compatibility condition $u_0^{e_1}(0) = g^{v_1}(0)$ is satisfied. The



h	err	rate
2^0	0.0303	—
2^{-1}	0.0076	1.9948
2^{-2}	0.0019	2.0010
2^{-3}	0.0005	1.9774
2^{-4}	0.0001	1.9978
2^{-5}	0.0000	1.9725

Fig. 1 Left: Snapshot of the exact solution u (blue) and the hybrid dG solution u_h for mesh size $h = 1$ (red, dashed). The discontinuity at network junctions is clearly visible. Right: Error and convergence rates for time horizon $t_{\max} = 5$. As expected we observe second order convergence

solution for this problem can be computed analytically and one can verify that $u \in W^{1,1}(0, t_{\max}; H^2_{pw}(\mathscr{T}_h))$. From the estimates of Theorem 2, we therefore expect second order convergence when discretizing with piecewise polynomials of order $k = 1$. For time integration, we utilize an implicit Euler method with sufficiently small step size $\tau \leq h^2$, such that time discretization errors are negligible, and we use

$$err = \max_{0 \leq t^n \leq t_{\max}} \|I_h u(t_n) - u_h^n\|_{L^2(\mathscr{E})}$$

as a measure for the error, where $I_h u$ denotes the element-wise linear interpolation. As expected, the convergence rates observed in our numerical tests coincide with predictions from Theorem 2. The solution plot in Fig. 1 clearly illustrates the discontinuity of the analytical solution at network junctions.

Acknowledgements This work was supported by the German Research Foundation (DFG) via grants TRR 154 C4 and the “Center for Computational Engineering” at TU Darmstadt.

References

1. Di Pietro, D.A., Ern, A.: Mathematical Aspects of Discontinuous Galerkin Methods. Springer Science & Business Media (2011)
2. Dorn, B.: Semigroups for flows on infinite networks. M.Sc. thesis, Eberhard Karls Universität Tübingen (2005)
3. Egger, H., Kugler, T.: Damped wave systems on networks: exponential stability and uniform approximations. Numer. Math. **138**, 839–867 (2018)
4. Egger, H., Schöberl, J.: A hybrid mixed discontinuous Galerkin finite element method for convection-diffusion problems. IMA J. Num. Anal. **30**, 1206–1234 (2009)

5. Engel, K.J., Nagel, R.: One-Parameter Semigroups for Linear Evolution Equations, 1st edn. Springer, New York (2000)
6. Garavello, M., Piccoli, B.: Traffic flow on networks. In: AIMS Series on Applied Mathematics, vol. 1. American Institute of Mathematical Sciences (AIMS), Springfield, MO (2006)
7. Johnson, C.: Numerical Solution of Partial Differential Equations by the Finite Element Method. Dover Publications (2009)
8. Kramar, M., Sikolya, E.: Spectral properties and asymptotic periodicity of flows in networks. Mathematische Zeitschrift **249**, 139–162 (2005)
9. Lagnese, L.E., Leugering, G., Schmidt, E.J.P.G.: Modeling, Analysis and Control of Dynamic Elastic Multi-link Structures. Systems & Control: Foundations & Applications. Springer Science+Business Media, New York (1994)
10. Mugnolo, D.: Semigroup Methods for Evolution Equations on Networks. Springer (2014)
11. Pazy, A.: Semigroups of Linear Operators and Applications to Partial Differential Equations, 1st edn. Springer, New York (1983)
12. Philippi, N.: Analysis and numerical approximation of transport equations on networks. M.Sc. thesis, TU Darmstadt (2019). https://opus4.kobv.de/opus4-trr154/
13. Thomée, V.: Galerkin Finite Element Methods for Parabolic Problems. Springer (1984)

MUSCL Discretization for the Fluid Flow Convection Operator on Staggered Meshes

A. Brunel, R. Herbin, and J.-C. Latché

Abstract We propose in this paper a second order discretization of the momentum convection operator for fluid flow simulation on staggered quadrangular or hexahedral meshes. The velocity is approximated by the Rannacher-Turek finite element. The implemented MUSCL-like approach is of algebraic type, in the sense that the limitation procedure does not invoke any slope reconstruction, and is independent from the geometry of the cells. The derived discrete convection operator applies both to constant or variable density flows; we perform here numerical tests for the barotropic and incompressible Navier-Stokes equations.

Keywords Convection operator · Fluid flow · MUSCL · Staggered grid

MSC (2010) 65M08 · 76M12

1 Introduction

Several works combining a finite element approximation of diffusion terms with a finite volume discretization for the convection operator may be found in the literature. The implementation of such a technique to obtain monotone schemes for convection-diffusion equations may be found for instance in [1–3]. Since finite-volume convection operators (with suitable upwinding) also enjoy desirable L^2-stability properties, they have been used for the discretization of the Navier-Stokes equations, preferably for compatible accuracy, i.e., low order, approximations. An

A. Brunel · R. Herbin
Aix Marseille University, CNRS, Marseille, France
e-mail: raphaele.herbin@univ-amu.fr

A. Brunel (✉) · J.-C. Latché
Institut de Sûreté et de Radioprotection Nucléaire (IRSN), Fontenay aux Roses, France
e-mail: aubin.brunel@univ-amu.fr; aubin.brunel@irsn.fr

J.-C. Latché
e-mail: jean-claude.latche@irsn.fr

R. Klöfkorn et al. (eds.), *Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples*, Springer Proceedings in Mathematics & Statistics 323,
https://doi.org/10.1007/978-3-030-43651-3_46

application of this strategy for the discretization of stationary incompressible Navier-Stokes equations by Crouzeix-Raviart finite elements may be found in [4]; extension to quasi-incompressible unsteady flows, both with the Crouzeix-Raviart and Rannacher-Turek finite elements, is performed in [5]. These two works only consider a first order upwinding technique, and our aim in this paper is to develop a second order convection operator, based on an algebraic MUSCL-like technique [6]. The obtained operator is quite general, in the sense that it may be applied to incompressible constant or variable density flows as well as to compressible flows, either for Navier-Stokes or Euler equations. We show here some numerical applications for the barotropic and the incompressible Navier-Stokes equations.

The continuous momentum convection operator that we consider here takes the following generic form $\partial_t(\rho \boldsymbol{u}) + \mathrm{div}(\rho \boldsymbol{u} \otimes \boldsymbol{u})$ where ρ is the fluid density and $\boldsymbol{u}$ the velocity. It may be recast under the form of a transport operator (which is central for its stability) provided that a mass balance equation holds, that is

$$\partial_t \rho + \mathrm{div}(\rho \boldsymbol{u}) = 0, \tag{1}$$

which we suppose here. The problem in which the convection operator is involved is supposed to be posed over $\Omega \times [0, T)$ where $\Omega \subset \mathbb{R}^d$ (with $d = 2, 3$) is an open bounded domain of boundary $\partial\Omega$ and $[0, T)$ is a finite time interval.

2 Space and Time Discretizations

We first define a primal mesh $\mathcal{M}$ by splitting Ω into a finite family of disjoint quadrangles (if $d = 2$) or hexahedra (if $d = 3$) denoted by K and called control volumes or cells. We then denote by $\mathcal{E}$ the set of faces of the mesh $\mathcal{M}$; for $K \in \mathcal{M}$, $\mathcal{E}(K)$ stands for the set of faces of K and we thus have $\partial K = \cup_{\sigma\in\mathcal{E}(K)}\sigma$. Any face $\sigma \in \mathcal{E}$ is either a part of the boundary of Ω, i.e., $\sigma \subset \partial\Omega$, in which case σ is said to be an external face, or there exists $(K, L) \in \mathcal{M}^2$ with $K \neq L$ such that $\overline{K} \cap \overline{L} = \sigma$: we denote in this case $\sigma = K|L$ and σ is said to be an internal face. We denote by $\mathcal{E}_{\mathrm{ext}}$ and $\mathcal{E}_{\mathrm{int}}$ the set of external and internal faces. For $K \in \mathcal{M}$ and $\sigma \in \mathcal{E}$, we denote by $|K|$ the measure of K and $|\sigma|$ the $(d-1)$-measure of the face σ.

The discretization is staggered in the sense that the scalar and vector unknowns are not colocated. Indeed, the pressure and density unknowns are associated with the cells of the primal mesh $\mathcal{M}$ and denoted by p_K, ρ_K while the degrees of freedom for the velocity are defined on a dual mesh using the Rannacher-Turek non-conforming low-order finite element approximation [7] and are denoted $\boldsymbol{u}_\sigma = (u_{\sigma,1}, \dots, u_{\sigma,d})$. The dual mesh is constructed as follows: if $K \in \mathcal{M}$ is a rectangle or a rectangular cuboid, we denote by x_K the mass center of K and we construct $D_{K,\sigma}$ as the cone with basis σ and with vertex x_K; this definition is extended to a general cell K, by supposing that K is split in the same number of sub-cells (the geometry of which does not need to be specified) and with the same connectivity. We now define D_σ, the dual cell of basis σ, as $D_\sigma = D_{K,\sigma} \cup D_{L,\sigma}$ if $\sigma = K|L \in \mathcal{E}_{\mathrm{int}}$ and $D_\sigma = D_{K,\sigma}$ if

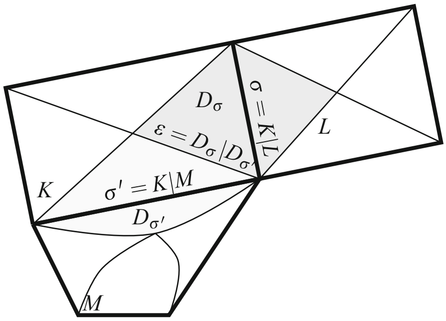


Fig. 1 Primal and dual meshes for the Rannacher-Turek elements

$\sigma \in \mathcal{E}(K) \cap \mathcal{E}_{\text{ext}}$; its measure is denoted by $|D_\sigma|$. We then denote by $\tilde{\mathcal{E}}(D_\sigma)$ the set of dual faces of D_σ, and by $\varepsilon = D_\sigma|D_{\sigma'}$ the face separating two dual cells D_σ and $D_{\sigma'}$. All the components of the velocity are then approximated on each face of the mesh, and their degrees of freedom are identified to the mean value of the velocity component over the face. An example of the discretization with a few control volumes is given on Fig. 1.

Finally, a constant time step denoted by δt is used for the time discretization, and for $0 \le n \le N = T/\delta t$, we define $t^n = n\,\delta t$.

3 A Second Order Discrete Convection Operator

Let us first address the discretization of the mass balance equation (1) over the primal mesh. Mimicking the divergence theorem, the discrete divergence reads (dropping the time exponents for short), for $K \in \mathcal{M}$:

$$\operatorname{div}(\rho \boldsymbol{u})_K = \frac{1}{|K|} \sum_{\sigma \in \mathcal{E}(K)} F_{K,\sigma} \tag{2}$$

where $F_{K,\sigma}$ stands for the (primal) numerical mass flux across σ outward K and is defined by:

$$\forall \sigma = K|L \in \mathcal{E}_{\text{int}}, \quad F_{K,\sigma} = |\sigma| \rho_\sigma \boldsymbol{u}_\sigma \cdot \boldsymbol{n}_{K,\sigma} \tag{3}$$

with $\boldsymbol{n}_{K,\sigma}$ the normal vector to the face σ outward K.

The dual mass fluxes $F_{\sigma,\varepsilon}$ for $\sigma \in \mathcal{E}$ and $\varepsilon \in \tilde{\mathcal{E}}(D_\sigma)$, $\varepsilon \subset K$ are constructed from these primal fluxes so as to ensure that a discrete mass balance holds over the dual mesh D_σ; this is obtained by computing the face densities ρ_{D_σ} and the mass fluxes $F_{\sigma,\varepsilon}$ as a linear combination of the densities in the primal cells adjacent to σ and the mass fluxes through the primal faces of K, respectively; we refer to [8] for the exact expressions.

These dual fluxes are then used for the definition of the discrete momentum convection operator, i.e., the discretization of the continuous term $\mathrm{div}(\rho u_i \boldsymbol{u})$. For $1 \le i \le d$ and $\sigma \in \mathcal{E}$, the term $\mathrm{div}(\rho u_i \boldsymbol{u})_\sigma$ reads:

$$\mathrm{div}(\rho u_i \boldsymbol{u})_\sigma = \frac{1}{|D_\sigma|} \sum_{\varepsilon \in \tilde{\mathcal{E}}(D_\sigma)} F_{\sigma,\varepsilon} u_{\varepsilon,i} \tag{4}$$

where $u_{\varepsilon,i}$ is an approximation of u_i over the face ε; it is obtained by the algebraic MUSCL-like technique introduced in [6], which implements the following procedure. Let us consider the explicit part of the convection term

$$T_{\sigma,i} = \rho^n_{D_\sigma} u^n_{\sigma,i} - \delta t\ \mathrm{div}(\rho^n u^n_i \boldsymbol{u}^n)_\sigma.$$

The discrete convection operator is said to be monotone if the term $T_{\sigma,i}$ can be written as a convex combination of degrees of freedom of u^n_i; for instance, such a property would ensure a discrete maximum principle for the transport equation, or a convection-diffusion equation with a suitable (only available on specific meshes) discretization of the diffusion term. Such a formulation of the term $T_{\sigma,i}$ is possible by using the discrete mass balance over D_σ if the following condition holds for each $\varepsilon \in \tilde{\mathcal{E}}_{\mathrm{int}}$ such as $\varepsilon = D_\sigma | D_{\sigma'}$:

$$\exists \alpha^\sigma_\varepsilon \in [0,1], \exists \tilde{\sigma} \in \mathcal{E} \text{ such that } \quad u_{\varepsilon,i} - u_{\sigma,i} = \left| \begin{array}{l} \alpha^\sigma_\varepsilon (u_{\sigma,i} - u_{\tilde{\sigma},i}) \text{ if } F_{\sigma,\varepsilon} \ge 0, \\ \alpha^\sigma_\varepsilon (u_{\tilde{\sigma},i} - u_{\sigma,i}) \text{ otherwise,} \end{array} \right. \tag{5}$$

together with the following CFL condition:

$$\mathrm{CFL} = \max_{\sigma \in \mathcal{E}} \Big\{ \frac{\delta t}{|D_\sigma|} \sum_{\varepsilon \in \tilde{\mathcal{E}}(D_\sigma)} |F_{\sigma,\varepsilon}| \Big\} \le 1. \tag{6}$$

We now deduce from the relation (5) a constructive process to compute the quantities $u_{\varepsilon,i}$. Let ε be an internal face separating an upwind dual cell D_{σ^-} from the downstream dual cell D_{σ^+} (i.e., $F_{\sigma^-,\varepsilon} \ge 0$). Let us now choose two sets $\mathcal{N}_\varepsilon(D_{\sigma^-})$ and $\mathcal{N}_\varepsilon(D_{\sigma^+})$ of neighbouring dual cells of D_{σ^-} and D_{σ^+} respectively. The following assumptions are then a transcription of Condition (5):

$$\exists\ D_{\overline{\sigma}} \in \mathcal{N}_\varepsilon(D_{\sigma^+}) \text{ such that } u_{\varepsilon,i} \in I_1 = [u_{\overline{\sigma},i}, u_{\overline{\sigma},i} + \frac{\xi^+}{2}(u_{\sigma^+,i} - u_{\overline{\sigma},i})]; \tag{7a}$$

$$\exists\ D_{\overline{\sigma}} \in \mathcal{N}_\varepsilon(D_{\sigma^-}) \text{ such that } u_{\varepsilon,i} \in I_2 = [u_{\sigma^-,i}, u_{\sigma^-,i} + \frac{\xi^-}{2}(u_{\sigma^-,i} - u_{\overline{\sigma},i})]; \tag{7b}$$

where ξ^+ and ξ^- are two numerical parameters lying in the interval [0, 2].

Here we choose $\mathcal{N}_\varepsilon(D_{\sigma^+}) = \{D_{\sigma^-}\}$. Concerning $\mathcal{N}_\varepsilon(D_{\sigma^-})$, several choices are possible; we choose here the opposite cell to D_{σ^+} with regard to D_{σ^-}, which means that if D_{σ^-} and D_{σ^+} share the primal face ε, we choose the cell $D_{\sigma'}$ such that $D_{\sigma'}$

shares a face ε' with D_{σ^-} and ε and ε' share no vertex. With these choices, the upwind value is always admissible, and in fact it is the only admissible cell if the two parameters are chosen equal to 0.

We are now in position to give the algorithm used to compute the quantities $u_{\varepsilon,i}$:

1. Compute a tentative value $\overline{u}_{\varepsilon,i}$ with a convex combination of the values (e.g. the centered choice) in the surrounding faces.
2. Evaluate $F_{\sigma,\varepsilon}$ to determine the upwind face D_{σ^-} and the downwind face D_{σ^+}, and choose accordingly the neighbouring sets $\mathcal{N}_\varepsilon(D_{\sigma^-})$ and $\mathcal{N}_\varepsilon(D_{\sigma^+})$.
3. Compute an admissible interval $I_1 \cap I_2$ for $u_{\varepsilon,i}$ by (7).
4. Compute $u_{\varepsilon,i}$ by projecting the tentative value $\overline{u}_{\varepsilon,i}$ into the interval obtained in the previous step.

Since this procedure is not linear, we cannot expect to derive an explicit formula to compute the values of the coefficients a_ε^σ. Their evaluation is however useless for an explicit scheme: indeed, the presented algorithm univocally defines the value $u_{\varepsilon,i}$. However, for this reason, we cannot easily define an implicit-in-time MUSCL scheme (this would require an iterative process for each time step).

4 Numerical Tests

The computations presented here are performed with the open-source CALIF^{3}S software developed at IRSN [9].

4.1 *Compressible Navier-Stokes Equations*

We first show an application to the barotropic compressible case:

$$\partial_t(\rho u_i) + \mathrm{div}(\rho u_i \boldsymbol{u}) + \partial_i p - \mathrm{div}(\mu(\nabla \boldsymbol{u} + \nabla \boldsymbol{u}^t))_i = 0 \qquad (i \in [\![1, d]\!]) \tag{8a}$$

$$\partial_t(\rho) + \mathrm{div}(\rho \boldsymbol{u}) = 0 \tag{8b}$$

$$p = a\rho^\gamma, \quad a > 0,\ \gamma \geq 1. \tag{8c}$$

where p stands for the pressure.

The scheme—A first-order forward Euler time-discretization of System (8) reads:

$$\forall K \in \mathcal{M}, \quad \frac{1}{\delta t}(\rho_K^{n+1} - \rho_K^n) + \mathrm{div}(\rho^n \boldsymbol{u}^n)_K = 0, \tag{9a}$$

For $1 \leq i \leq d,\ \forall \sigma \in \mathcal{E}$,

$$\frac{1}{\delta t}(\rho_{D_\sigma}^{n+1} u_{\sigma,i}^{n+1} - \rho_{D_\sigma}^n u_{\sigma,i}^n) + \mathrm{div}(\rho^n u_i^n \boldsymbol{u}^n)_\sigma$$

$$+ (\nabla p)^n_{\sigma,i} - \mathrm{div}(\mu(\nabla \boldsymbol{u}^n + \nabla(\boldsymbol{u}^n)^t))_{\sigma,i} = 0, \tag{9b}$$

$$\forall K \in \mathcal{M},\ p_K^{n+1} = a\ (\rho_K^{n+1})^\gamma, \tag{9c}$$

with the previously described discrete convection terms and with the discrete pressure gradient and momentum diffusion term as given in [10]. Second order in time is obtained by using the Heun scheme, which reads

$$\boldsymbol{W}^{n+\frac{1}{3}} = S(\boldsymbol{W}^n), \quad \boldsymbol{W}^{n+\frac{2}{3}} = S(\boldsymbol{W}^{n+\frac{1}{3}}), \quad \boldsymbol{W}^{n+1} = \frac{1}{2}(\boldsymbol{W}^n + \boldsymbol{W}^{n+\frac{2}{3}}),$$

where $\boldsymbol{W}^n = (\rho^n, u^n, p^n)$ is the vector of unknowns at step n and $S(\boldsymbol{W}^n)$ is obtained by one step of the forward Euler (9) The Heun scheme is used in the following numerical test.

Translated standing vortex—Here we assess the convergence rate of the proposed scheme on a test case built to this purpose. We first derive an analytical solution of the steady barotropic Euler equations consisting in a standing vortex; then this solution is made unsteady by adding a constant velocity translation. A solution for the Navier-Stokes equations is finally derived by compensating the viscous forces (that appear on the left hand side of equation (8a)) with a source term. We refer to [11] for the expression of this solution.

The viscosity μ is chosen so that the Reynolds number is equal to 50, $a = 9.81/2$ and $\gamma = 2$. The domain is the square $\Omega = [-1.2, 2]^2$ and the computation is run on the time interval $[0, 0.8]$. Uniform $n \times n$ grids are used, starting from $n = 32$ and doubling the number of control volumes in each direction up to $n = 256$ mesh. The time step is set to $0.03125 \times h$, with $h = 3.2/n$, which yields a CFL number with respect to the celerity of the fastest wave close to 0.01 (the material velocity and the speed of sound are in the range of 1.45 and 0.76 respectively); this low value of the CFL number is imposed by the explicit discretization of the diffusion term (the constraint stems from the necessity to be stable up to the finest mesh).

On Fig. 2, we draw the L^1 norm of the numerical error for the velocity and the density as a function of the mesh step. This error is obtained by taking the difference between the computed velocity or density at the final time and the piecewise constant function defined by taking on each dual cell the value of the continuous solution at the cell center. The measured convergence orders are close to 1.8 and 2 for the velocity and the density respectively.

4.2 Incompressible Navier-Stokes Equation

We now turn to the incompressible Navier-Stokes equations (with $\rho = 1$):

$$\partial_t u_i + \mathrm{div}(u_i \boldsymbol{u}) + \partial_i p - \mathrm{div}(\mu(\nabla \boldsymbol{u} + \nabla \boldsymbol{u}^t))_i = 0 \qquad (i \in [\![1, d]\!]), \tag{10a}$$

$$\mathrm{div}(\boldsymbol{u}) = 0. \tag{10b}$$

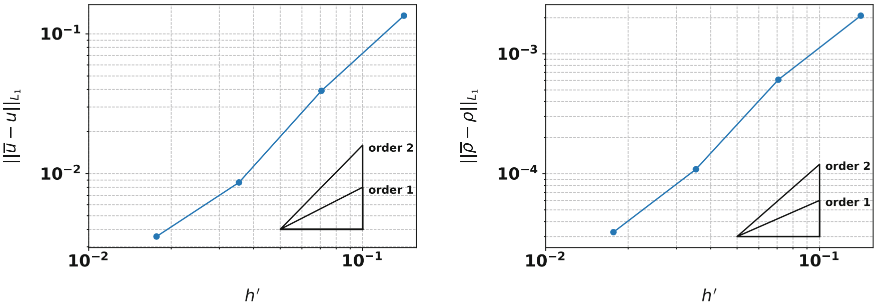


Fig. 2 L^1 norm error for the MUSCL scheme for the velocity and the density. Here $h' = \max_{K\in\mathcal{M}} \mathrm{diam}(K) = \sqrt{2}h$

The scheme—This system is solved using a projection scheme, which consists in the two following steps:

Prediction step—Solve for $\tilde{\boldsymbol{u}}^{n+1}$:

$$\begin{aligned}\text{For } 1 \le i \le d,\ \forall\sigma \in \mathcal{E},\quad & \frac{1}{\delta t}\left(\tilde{\boldsymbol{u}}^{n+1}_{\sigma,i} - \boldsymbol{u}^{n}_{\sigma,i}\right) + \mathrm{div}(\tilde{\boldsymbol{u}}^{n}_{i}\boldsymbol{u}^{n})_\sigma \\ & +(\nabla p)^n_{\sigma,i} - \mathrm{div}(\mu(\nabla\tilde{\boldsymbol{u}}^{n+1} + (\nabla\tilde{\boldsymbol{u}}^{n+1})^t))_{\sigma,i} = 0.\end{aligned} \tag{11a}$$

Correction step—Solve for p^{n+1} and $\boldsymbol{u}^{n+1}$:

$$\text{For } 1 \le i \le d,\ \forall\sigma \in \mathcal{E},\quad \frac{1}{\delta t}(\boldsymbol{u}^{n+1}_{\sigma,i} - \tilde{\boldsymbol{u}}^{n+1}_{\sigma,i}) + (\nabla p^{n+1})_{\sigma,i} - (\nabla p^{n})_{\sigma,i} = 0, \tag{11b}$$

$$\forall K \in \mathcal{M},\quad \mathrm{div}(\boldsymbol{u}^{n+1})_K = 0. \tag{11c}$$

The convection terms are defined in the previous section, with the density set to 1 in the mass flux. We refer once again for short to [10] for the definition of the discrete pressure gradient and the momentum diffusion term.

Flow past a cylinder—We consider here a flow past a cylinder studied in a literature benchmark [12]. The geometry of the domain is given in [12, Fig. 1]. The present test corresponds to the **Test Case 2D-2** of [12]. The viscosity μ is chosen so that the Reynolds number is equal to 5000, so the convection is strongly dominant. The computations are performed using a very coarse grid with 4033 cells, representative of what may be encountered in very complex 3D industrial simulations. We compare the results with the present convection scheme with the results obtained with (implicit-in-time) upwind and centered convection operators.

The main quantities of interest are the pressure difference ΔP between the front and end points of the cylinder, the Strouhal number, the maximum drag coefficient and the maximal and minimal lift coefficients. They are gathered in Table 1 and Table 2, together with reference values obtained with a converged-in-space computation. Even if the convergence is far from being reached with the (intentionally) very

Table 1 ΔP and Strouhal number

Reference	
$\Delta P_{\max}$	S_t
3.33080	0.3371

Scheme	$\Delta P_{\max}$	S_t
Upwind	2.37020	0.2304
Centered	2.00400	0.2591
MUSCL	2.63470	0.2660

Table 2 Lift and drag coefficients

Reference		
$c_{d,\max}$	$c_{l,\max}$	$c_{l,\min}$
3.46088	2.8479	–2.78536

Scheme	$c_{d,\max}$	$c_{l,\max}$	$c_{l,\min}$
Upwind	3.23544	0.52157	–0.51295
Centered	3.18728	0.10793	–0.16129
MUSCL	3.42478	1.23162	–1.16792

coarse mesh used in this study, the MUSCL scheme seems able to capture at least the order of magnitude of the reported quantities. This is supported by an examination of the computed flow structure: the vortex shedding phenomenon is qualitatively reproduced by the MUSCL scheme while the upwind and centered ones yield, respectively, unrealistic small and large recirculation zones.

References

1. Ohmori, K., Ushijima, T.: A technique of upstream type applied to a linear nonconforming finite element approximation of convective diffusion equations. RAIRO. Anal. Numér. **18**, 309–332 (1984)
2. Angermann, L.: Numerical solution of second-order elliptic equations on plane domains. Math. Model. Numer. Anal. **25**, 169–191 (1991)
3. Eymard, R., Hilhorst, D., Vohralík, M.: A combined finite volume-nonconforming/mixed-hybrid finite element scheme for degenerate parabolic problems. Numer. Math. **105**, 73–131 (2006)
4. Schieweck, F., Tobiska, L.: An optimal order error estimate for an upwind discretization of the Navier-Stokes equations. Numer. Methods Part.L Differ. Equ.S **12**, 407–421 (1996)
5. Ansanay-Alex, G., Babik, F., Latché, J., Vola, D.: An L2-stable approximation of the Navier-stokes convection operator for low-order non-conforming finite elements. Int. J. Numer. Methods Fluids **66**, 555–580 (2011)

6. Piar, L., Babik, F., Herbin, R., Latché, J.-C.: A formally second-order cell centred scheme for convection-diffusion equations on general grids. Int. J. Numer. Methods Fluids **71**, 873–890 (2013)
7. Rannacher, R., Turek, S.: Simple nonconforming quadrilateral stokes element. Numer. Methods Part.L Differ. Equ.S **8**, 97–111 (1992)
8. Herbin, R., Latché, J.-C., Nguyen, T.: Consistent segregated staggered schemes with explicit steps for the isentropic and full Euler equations. Math. Model. Numer. Anal. **52**, 893–944 (2018)
9. CALIF^{3}S: A software components library for the computation of fluid flows. https://gforge.irsn.fr/gf/project/califs
10. Grapsas, D., Herbin, R., Kheriji, W., Latché, J.-C.: An unconditionally stable staggered pressure correction scheme for the compressible Navier-stokes equations. SMAI J. Comput. Math. **2**, 51–97 (2016)
11. Gallouët, T., Herbin, R., Latché, J.-C., Nasseri, Y.: A second order consistent MAC scheme for the shallow water equations. *This conference* (2020)
12. Schäfer, M., Turek, S., Durst, F., Krause, E., Rannacher, R.: Benchmark computations of laminar flow around a cylinder. Flow Simul. High-Perform. Comput. **II**, 547–566 (1996)

An Active Flux Method for Cut Cell Grids

Christiane Helzel and David Kerkmann

Abstract We present recent work in progress towards the development of a third order accurate Cartesian grid cut cell method for the approximation of hyperbolic conservation laws in complex geometries. Our cut cell method is based on the *Active Flux* method of Eymann and Roe, a new finite volume method, which evolves both cell average values and point values of the conserved quantities. The evolution of the point values leads to an automatic stabilisation of the cut cell update, i.e. the method is stable for time steps that are appropriate for the regular cells. While most of the existing cut cell stabilisation methods lead to a loss of accuracy, we show that it is possible to obtain third order accurate results. In this contribution we restrict our considerations to the linear transport equation in one and two space dimensions.

Keywords Cartesian cut cell method · Finite volume method · Active Flux method · High-order methods

MSC (2010) 65M08 · 65M12 · 65M25 · 35L65 · 35L04

1 Introduction

The Active Flux method, first introduced by Eymann and Roe [3], is a new finite volume method that offers efficient third order computations. Its very local stencil is a very attractive feature that might simplify the treatment of the boundary of a complex geometry. Previous works have defined Active Flux methods in one and two space dimensions [3, 4]. Two-dimensional methods have been derived both for triangular [4] and for rectangular grids [1, 5]. The latter form the basis for our cut

C. Helzel · D. Kerkmann (✉)
Mathematisches Institut, Lehrstuhl für Angewandte Mathematik, Heinrich-Heine-Universität Düsseldorf, Düsseldorf, Germany
e-mail: david.kerkmann@hhu.de

C. Helzel
e-mail: helzel@hhu.de

R. Klöfkorn et al. (eds.), *Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples*, Springer Proceedings in Mathematics & Statistics 323,
https://doi.org/10.1007/978-3-030-43651-3_47

cell method. To lay the foundations, we focus on the advection equation in one and two space dimensions.

This work is organised as follows: In Sect. 2 we discuss the Active Flux method for cut cells in one dimension. We provide accuracy and stability results. Section 3 is concerned with the method in two dimensions. We adapt the Cartesian grid method to cut cell grids and have a first look at accuracy and stability.

2 Active Flux for Cut Cells in One Space Dimension

2.1 Regular Grid

The Active Flux method in one space dimension uses not only the cell averages, but also point values of the conserved quantity at the interface as the degrees of freedom. Therefore, not only a definition of the numerical flux, but also the update of the point values is required for the construction of one time step. Let Q_i^n denote the cell average and $Q_{i-\frac{1}{2}}^n$ denote the left interface point value of cell i at time t_n. For this work, the advection equation $q_t + f(q)_x = 0$, $f(q) = aq$ for $a \in \mathbb{R}$ will be our test model. Without loss of generality let $a > 0$. The Active Flux method is defined by the following steps:

1. Reconstruct a uniquely defined parabola q_{rec} in each cell with the cell average and the two interface values that border the cell.
2. Define the interface updates through an evolution operator that approximates the exact evolution. In our case, by the method of characteristics, the exact solution at time level $t_{n+1} = t_n + \Delta t$ is $q(x, t_n + \Delta t) = q(x - a\Delta t, t_n)$. The new interface value $Q_{i-\frac{1}{2}}^{n+1}$ will be the evaluation of the reconstruction at the corresponding position.
3. Approximate the flux over an interface $x_{i-\frac{1}{2}}$ at time t_n with Simpson's rule:

$$\frac{1}{\Delta t}\int_0^{\Delta t} f(q(x_{i-\frac{1}{2}}, t_n + t))dt \approx \frac{1}{6}\left(f(Q_{i-\frac{1}{2}}^n)) + 4f(Q_{i-\frac{1}{2}}^{n+\frac{1}{2}})) + f(Q_{i-\frac{1}{2}}^{n+1}))\right) =: F_{i-\frac{1}{2}}^n \tag{1}$$

The new cell interface value $Q_{i-\frac{1}{2}}^{n+1}$ is used in the flux formulation. The middle value $Q_{i-\frac{1}{2}}^{n+\frac{1}{2}}$ is obtained in the same way as $Q_{i-\frac{1}{2}}^{n+1}$ by replacing Δt with $\frac{\Delta t}{2}$ in step 2. Notice that for one-dimensional linear problems, the flux approximation by Simpson's rule is exact for quadratic reconstructions, as it can be transformed to an integral in space [5]:

$$\frac{1}{\Delta t}\int_0^{\Delta t} f(q(x_{i+\frac{1}{2}}, t_n + t))dt = \frac{1}{\Delta t}\int_{x_{i+\frac{1}{2}}-a\Delta t}^{x_{i+\frac{1}{2}}} q_{rec}(x, t_n)dx \tag{2}$$

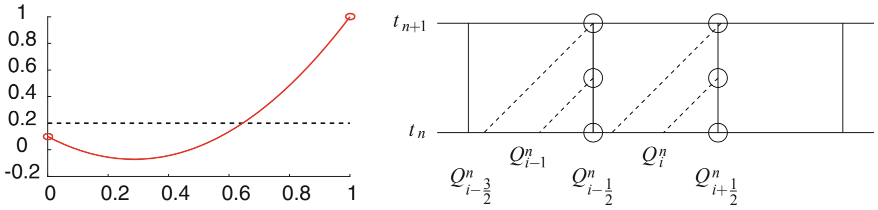


Fig. 1 Active Flux method. Left: Reconstruction in one cell (here in a reference cell). Right: Point evaluation and flux computation

4. Perform the usual finite volume update

$$Q_i^{n+1} = Q_i^n - \frac{\Delta t}{\Delta x}\left(F^n_{i+\frac{1}{2}} - F^n_{i-\frac{1}{2}}\right). \tag{3}$$

Figure 1 illustrates the method. This method is third order accurate. The local truncation error is given in Sect. 2.3. The time step restriction for this method reads

$$a\Delta t \leq \Delta x. \tag{4}$$

2.2 Artificial Cut Cell

Consider a one-dimensional grid $x_{i-\frac{1}{2}}$, $i = 0, \dots, n+1$, with cell interface values $Q_{i-\frac{1}{2}}$ and cell average values Q_i. Let $x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}} = \Delta x \ \forall i \in \{0, \dots, n+1\} \setminus \{k\}$. Let $x_{k+\frac{1}{2}} - x_{k-\frac{1}{2}} = \alpha \Delta x$ for $\alpha \in (0, 1)$. Cell k will be referred to as the *small cell*. As cells in one dimension can only be cut into smaller cells, this corresponds to the only cut cell situation. Following the steps from above, we naturally obtain a cut cell method by defining the treatment of steps 3 and 4 for the cut cell. The natural time step restriction if we maintain the local stencil becomes

$$a\Delta t < \alpha \Delta x. \tag{5}$$

Depending on the signs of $a\Delta t - \alpha\Delta x$ and $a\Delta t - 2\alpha\Delta x$, the evaluations of the reconstruction for the required flux quadrature points at interface $k + \frac{1}{2}$ will be done in cell k or $k - 1$. The case $a\Delta t - \alpha\Delta x < 0$ satisfies condition (5) and can be described as an irregular grid case. In cut cell applications, the size of the *small cell* can be orders of magnitude smaller than the standard cell size, i.e. $\alpha \ll 1$. In that case, the update of the interface as well as the middle value needed for the flux will be taken from the reconstruction in cell $k - 1$. The third possibility combines both previous cases in having one evaluation in each cell. Figure 2 shows all cases.

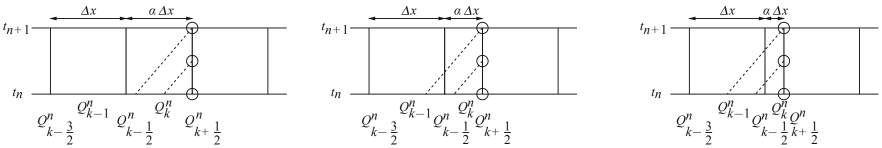


Fig. 2 Extracts from one-dimensional cut cell grids. Left: both characteristics origin in the small cell. Center: one of the characteristics origins in the small cell. Right: no characteristic origins in the small cell

In the interesting cases, where $a\Delta t - 2\alpha\Delta x > 0$, the approximation to the flux at the interface $k + \frac{1}{2}$ is no longer exact since the integrand consists of a piecewise quadratic function that is continuous, but not differentiable at $x_{k-\frac{1}{2}}$. To study the accuracy in more detail, we state the local truncation errors.

2.3 Local Truncation Error

Let

$$q_i^n = \frac{1}{\Delta x_i}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} q(x, t_n)dx$$

be the cell average of the exact solution at time t_n, $q^n_{i-\frac{1}{2}} = q(x_{i-\frac{1}{2}}, t_n)$ and $\tilde{F}_{i-\frac{1}{2}}$ the numerical flux obtained by replacing the numerical values Q_i^n and $Q^n_{i-\frac{1}{2}}$ by the exact values q_i^n and $q^n_{i-\frac{1}{2}}$ for all i. Let $\gamma = \frac{a\Delta t}{\Delta x}$ be the CFL number for a grid of some standard cell size Δx.

Definition 1 The local truncation error in cell i is defined as

$$\tau_i := \frac{q_i^{n+1} - q_i^n + \frac{\Delta t}{\Delta x_i}\left(\tilde{F}_{i+\frac{1}{2}} - \tilde{F}_{i-\frac{1}{2}}\right)}{\Delta t}. \tag{6}$$

Lemma 1 *The local truncation error of the Active Flux method on a regular sized grid reads*

$$\tau_i = \frac{1}{24}a\Delta x^3\gamma(1-\gamma)^2\frac{\partial^4}{\partial x^4}q(x_i, t_n) + \mathscr{O}(\Delta x^4). \tag{7}$$

On a cut cell grid, as described in Sect. 2.2, one obtains one out of three possible schemes (cf. Fig. 2). We will only inspect the case $\Delta x > a\Delta t > 2\alpha\Delta x$ and give an argument for the other two cases.

Lemma 2 *The local truncation error of the Active Flux method in the small cell k, as defined above, in the case* $\Delta x > a\Delta t > 2\alpha\Delta x$ *reads*

$$\tau_k = \left(-\frac{5}{72} - \frac{5}{24}\alpha - \frac{5}{36}\alpha^2 + \frac{1}{4}(1+\alpha)\gamma - \frac{1}{6}\gamma^2\right) a\Delta x^2 \frac{\partial^3}{\partial x^3} q(x_k, t_n) + \mathcal{O}(\Delta x^3). \quad (8)$$

The local truncation error in cell $k+1$ *in the same case reads*

$$\tau_{k+1} = \left(\frac{5}{72} + \frac{5}{24}\alpha + \frac{5}{36}\alpha^2 - \frac{1}{4}(1+\alpha)\gamma + \frac{1}{6}\gamma^2\right) a\alpha\Delta x^2 \frac{\partial^3}{\partial x^3} q(x_{k+1}, t_n) + \mathcal{O}(\Delta x^3). \quad (9)$$

Proof The proofs of Lemmata 1 and 2 are done by using the Taylor series expansion and are ommited due to their lengths.

It is clearly visible that the method reduces to second order in the *small cell k* and its right neighbour $k+1$. This reduces the convergence order to two if measured in the L_∞ norm, but not if measured in the L_1 norm. This loss of accuracy was also observed in numerical simulations.

Remark 1 Notice that the error in cell $k+1$ will influence the flux at the interface $x_{k+\frac{3}{2}}$ and will thus propagate to cell $k+2$ by a certain percentage that depends on γ. The same effect will happen for cell $k+2$ and the flux at $x_{k+\frac{5}{2}}$ one time step later. This means that the error in cell $k+1$ will spread out to all other $n-k = \mathcal{O}(\frac{1}{\Delta x})$ cells over time. Even though the one-step-error $\Delta t\tau_{k+1} = \mathcal{O}(\Delta x^3)$ is created in every time step, the error in cell $k+1$ (and all subsequent cells) will be bounded through this harmonic property and one can confirm third order convergence even in cell $k+1$. The same effect does not appear in the cut cell in this situation because the cell average Q_k is never used as can be seen from the right plot of Fig. 2. For the other cases (left and center plot of Fig. 2) the effect takes place and third order is obtained.

To obtain third order in the L_∞ norm, one can replace Simpson's rule by an exact integration that can for example be carried out by an iterative Simpson's rule.

2.4 Stability

In order to study the linear stability of the one-dimensional Active Flux method in the presence of small cells we write the method in the matrix-vector form

$$Q^{n+1} = AQ^n,$$

where the vector Q^n contains all degrees of freedom at time level t_n. The method is Lax-Richtmyer stable iff $\|A^n\|$ is bounded independently of n. Using the Jordan

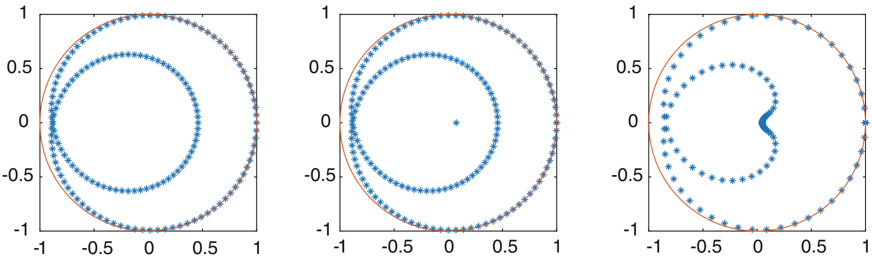


Fig. 3 Results of a linear stability analysis of the one-dimensional Active Flux method with regular cells (left), one small cell (center) and an alternation of small and regular cells (right). The CFL number used is 0.9. The cut cell size is $\alpha = 0.05$ for every cut cell

decomposition of A, one can show that asymptotic stability is equivalent to the condition $|\lambda| \leq 1$ for all eigenvalues λ of A and if $|\lambda| = 1$, then the geometric and algebraic multiplicity need to match [6]. While we do not have an analytical formula for the eigenvalues, we compute eigenvalues (using Python) for different grids. Results are shown in Fig. 3. In the left plot we show eigenvalues of the matrix A which describes the regular Active Flux method without small cells. In the central plot we show the situation with one small cell of size $\alpha \Delta x < \frac{1}{2} a \Delta t$. In the right plot we show the eigenvalues for the resulting method on a grid for which every second grid cell is a small cell. In all situations Simpson's rule provides a stable method for time steps that correspond to the usual CFL condition (4) of the regular part of the grid.

Note that linear stability of the Active Flux method with exact integration is easy to show, since exact evolution does not increase the total variation. However, in more general situations, it might not be possible to use exact integration. Therefore, it is important to know that Simpson's rule, while degrading the accuracy in the small cell, leads to a stable method.

3 Active Flux for Cut Cells in Two Space Dimensions

In the original work by Eymann and Roe, a triangular grid was used for the two-dimensional Active Flux method [4]. Recently, a Cartesian grid version was developed by Barsukow et. al [1] as well as Helzel, Kerkmann and Scandurra [5]. The latter will be the basis for our cut cell approach. This method follows the same procedure as already explained in the one-dimensional approach. We will briefly cover all changes. Details will be found in future publications.

We study the advection equation in two dimensions

$$q_t + aq_x + bq_y = 0, \quad (x, y) \in \Omega, \ t > 0, \tag{10}$$

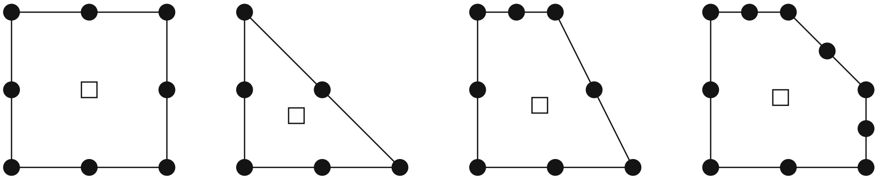


Fig. 4 Degrees of freedom in the two-dimensional Active Flux method for all possible cut cells

for a spatial domain Ω, $a, b \in \mathbb{R}$. Inflow and outflow boundary conditions are imposed.

1. In two space dimensions, cut cells can have various shapes. In this work, we restrict ourselves to cut cells with straight boundary segments. We also require that our grid is fine enough so each cell is only crossed by a maximum of one connected boundary path. Figure 4 shows a representative for each possible shape. The degrees of freedom are placed in a natural way along the boundaries. The solid dots indicate point values sitting on the interface. As the cells have a different amount of degrees of freedom, different reconstructions have to be used. All reconstructions use only the degrees of freedom that belong to the respective cell and give a third order approximation of the exact solution.
2. The exact evolution leads to the formula $q(x, y, t_n + \Delta t) = q(x - a\Delta t, y - b\Delta t, t_n)$ which is used for the update of the degrees of freedom on the interface.
3. The flux computation now uses a two-dimensional version of Simpson's rule. The required point values are found in the same way as in the one-dimensional case through the formula given in 2.

Remark 2 A similar transformation to Eq. (1) can be performed in two dimensions. Except for $a = 0$ or $b = 0$, the resulting spatial area will overlap multiple cells. Therefore, Simpson's rule is now no longer exact even for regular Cartesian grid cells. To perform an exact integration, one will have to iteratively integrate over the correct cell parts.

4. The finite volume update reads

$$Q_{i,j}^{n+1} = Q_{i,j}^{n} - \frac{\Delta t}{|\Omega_{i,j}|} \sum_{m=1}^{s} F_m^n \cdot \nu_m \tag{11}$$

where s denotes the number of cell boundary faces and ν_m denotes the outer normal.

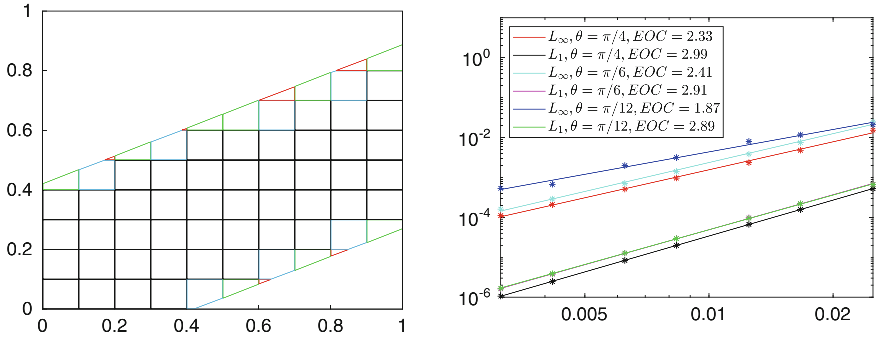


Fig. 5 Left: Coarse cut cell grid. Right: Estimated error of convergence for $\theta = \pi/12, \pi/6, \pi/4$

3.1 Accuracy Study

Let $\theta \in (0, \pi/4]$ be an angle, $\delta \in (0, 1)$ an offset, $\Omega = [0, 1]^2 \cap \{(x, y) \mid y - \tan(\theta)(x - \delta) > 0 \wedge y - \delta - \tan(\theta)x < 0\}$ a channel in two dimensions and $a = \cos(\theta)$ and $b = \sin(\theta)$ velocities parallel to the channel walls. The setup and an example of a cut cell grid are shown in Fig. 5 (left).

We impose inflow boundary conditions for $x = 0$ and $y = 0$ and outflow boundary conditions for $x = 1$ and $y = 1$. Since the flow is parallel to the channel walls, there is no flow across them. The initial condition reads

$$q_0(x, y) = 5 \exp(-100(x + y - 0.7)^2). \tag{12}$$

We use $\Delta t = 0.7 \max\{\frac{\Delta x}{a}, \frac{\Delta y}{b}\}$ and the final time $T = 0.4$. A similar test using a discontinuous Galerkin cut cell method is performed in [2].

We estimate the order of convergence by the solution to the least square problem that is given by fitting a straight line to the logarithmic errors. We test for various offsets δ and angles θ. The results for some values of θ and $\delta = 0.2001$ using exact integration are shown in Fig. 5 (right). The results for all other tested values look very similar but cannot be presented here due to the limited amount of space. The method remains stable and accurate for any cut cell size. Third order is achieved in the L_1 norm and second order or better is achieved in the L_∞ norm. In these tests, the size of the smallest cut cell varied between a factor of 10^{-3} to 10^{-8} compared to the regular cells.

4 Conclusions

We present a third order accurate finite volume cut cell method for the linear transport equation in one and two space dimensions. It is based on the Active Flux method and appears to be uniformly stable with regard to a time step determined by the size of the

Cartesian grid cells. More details of the method, in particular for the two-dimensional case, will be presented in a later publication. We plan to extend the method to more complicated geometries and equations, such as the linear transport equation with spatially varying velocity field, the linear acoustic equations and ultimately non-linear equations.

Acknowledgements This work was supported by the DFG through HE 4858/4-1.

References

1. Barsukow, W., Hohm, J., Klingenberg, C., Roe, P.L.: The Active Flux scheme on cartesian grids and its low mach number limit. J. Sci. Comput. **81**, 594–622 (2019)
2. Engwer, C., May, S., Nüssing, C., Streitbürger, F., A stabilized discontinuous Galerkin cut cell method for discretizing the linear transport equation. (2019). arXiv:1906.05642
3. Eymann, T.A., Roe, P.: Active Flux schemes. In: 49th AIAA Aerospace Science meeting (2011)
4. Eymann, T.A., Roe, P.L.: Multidimensional Active Flux schemes. In: AIAA Conference Paper (2013)
5. Helzel, C., Kerkmann, D., Scandurra, L.: A new ADER method inspired by the Active Flux method. J. Sci. Comput. **80**, 1463–1497 (2019)
6. van Drosselaer, J.L.M., Kraaijevanger, J.F.B.M., Spijker, M.N.: Linear stability analysis in the numerical solution of initial value problems. Acta Numer. **2**, 199–237 (1993)

Practical Examples

Finite Volume Discretisation of Fracture Deformation in Thermo-poroelastic Media

Ivar Stefansson, Inga Berre, and Eirik Keilegavlen

Abstract This paper presents a model where thermo-hydro-mechanical processes are coupled to a deformation model for preexisting fractures. The model is formulated within a discrete-fracture-matrix framework where the rock matrix and the fractures are considered as individual subdomains, and interaction between them takes place on the matrix-fracture interfaces. A finite volume discretisation implemented in the simulation toolbox PorePy is presented and applied in a simulation showcasing the effects of the different mechanisms on fracture deformation governed by contact mechanics, as well as their different timescales.

Keywords Thermo-poroelasticity · Porous media · Contact mechanics · Fractures · Mixed-dimensional

1 Introduction

We consider the simulation of fully coupled thermo-hydro-mechanical (THM) dynamics in fractured porous media, where the fractures can undergo sliding if the shear forces on the fracture planes are sufficient to overcome frictional resistance. These processes are highly relevant for several subsurface applications, including geothermal energy extraction, storage of CO_2 and energy and groundwater management. Our simulation approach is based on three main ingredients: First, conservation of mass, energy and momentum is preserved under discretisation by the employment of a fully coupled finite volume (FV) approach approach for the governing equations. Second, the network of fractures, which act as main conduits for fluid flow and energy transport, is explicitly represented in the simulation model. Specifically, the fractures are represented as lower-dimensional manifolds that are embedded in the host porous medium, thus the simulation model is defined on a mixed-dimensional geometry. Third, the sliding of fractures is modelled as a frictional contact problem, which is

I. Stefansson (✉) · I. Berre · E. Keilegavlen
University of Bergen, Bergen, Norway
e-mail: ivar.stefansson@uib.no

R. Klöfkorn et al. (eds.), *Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples*, Springer Proceedings in Mathematics & Statistics 323,
https://doi.org/10.1007/978-3-030-43651-3_48

solved by an active set approach. The discretisation of the contact problem benefits from the finite volume approach, which directly provides discrete representations of displacements as well as of mechanical, fluid and thermal forces on the fracture surfaces. Furthermore, the explicit degrees of freedom for fluid pressure inside the fractures allow us to capture the critical interplay between elevated fluid pressures and fracture deformation.

2 Model

We consider a mixed-dimensional geometry which is decomposed in subdomains of different dimensions representing the host porous medium and the lower-dimensional planar fractures, and separated by interfaces, see [5] for details. Variables and governing equations are defined on subdomains and interfaces, with full flexibility to vary the type and number of variables and equations between geometric objects. The framework accommodates heterogeneous and multiphysics models, with a natural treatment of modelling and discretisation of mixed-dimensional problems.

We denote a subdomain by Ω_i and its boundary by $\partial\Omega_i$, and identify the variables defined within it by subscript i. Where convenient, we will denote the higher-dimensional matrix domain by Ω_h and lower-dimensional fracture domains by Ω_l, as indicated in Fig. 1.

$\partial\Omega_i$ may be divided into the external boundary $\partial\Omega_i^e$ and the internal fracture boundary $\partial\Omega_i^f$, which coincides geometrically with both the immersed fracture domain Ω_l and the interface between Ω_h and Ω_l. This interface is denoted by Γ_j, and the associated variables identified by subscript j. The two sides of the fracture are denoted by $+$ and $-$, as shown in Fig. 1. Projection of variables from the interface to the subdomains is performed by Ξ_h^j and Ξ_l^j, respectively, whereas Π_h^j and Π_l^j project from the subdomains to the interface.

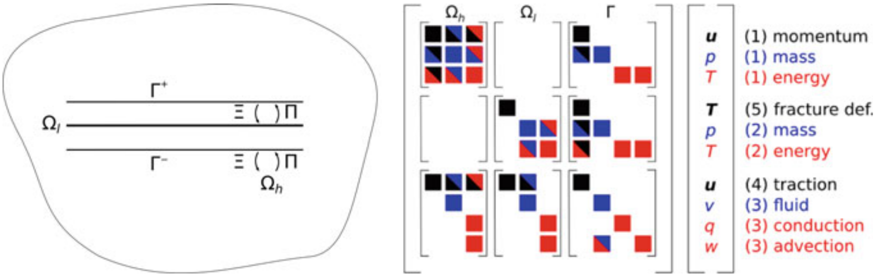


Fig. 1 Schematic representation of a fracture Ω_l and a matrix subdomain Ω_h to the left. The two subdomains are separated by the interface, whose two sides are denoted by Γ^+ and Γ^-. Also shown are the projection operators used for transfer of variables between the subdomains and the interface. To the right, we show the block structure of the matrix resulting from a fully implicit discretisation of Eqs. 1–5

For Ω_h, the primary variables are displacement $\boldsymbol{u}$, fluid pressure p and temperature T. The fluid flux and the advective and conductive heat fluxes are denoted by $\boldsymbol{v}$, $\boldsymbol{w}$ and $\boldsymbol{q}$, respectively. The definition of the parameters used in the following may be found in the repository at [8].

Assuming all external source and sink terms to be zero, conservation of momentum, mass and energy in Ω_h can be written as [9]

$$\begin{aligned}\nabla \cdot \left[\frac{\mathbf{D}}{2}(\nabla \boldsymbol{u} + \nabla \boldsymbol{u}^T) - \alpha p \mathbf{I} - \beta_s K (T - T_0)\mathbf{I}\right] = 0,\\ \left(\phi c + \frac{\alpha - \phi}{K}\right)\frac{\partial p}{\partial t} + \alpha \frac{\partial (\nabla \cdot \boldsymbol{u})}{\partial t} - \phi \beta_f \frac{\partial T}{\partial t} - \nabla \cdot \frac{\mathcal{K}}{\mu}\nabla p = 0,\\ \rho_e C_e \frac{\partial T}{\partial t} + \beta_s K T_0 \frac{\partial (\nabla \cdot \boldsymbol{u})}{\partial t} - \phi \beta_f T_0 \frac{\partial p}{\partial t} + \rho_f C_f \boldsymbol{v} \cdot \nabla T - \nabla \cdot \kappa_e \nabla T = 0.\end{aligned} \tag{1}$$

Similarly, conservation of mass and energy in Ω_l and fracture intersection domains Ω_x is given by

$$\begin{aligned}a^* c_f \frac{\partial p}{\partial t} + \frac{\partial a^*}{\partial t} - a^* \beta_f \frac{\partial T}{\partial t} - \nabla \cdot a^* \frac{\mathcal{K}}{\mu}\nabla p = \Xi_i^j v_j,\\ a^* \rho_f C_f \frac{\partial T}{\partial t} - a^* \beta_f T_0 \frac{\partial p}{\partial t} + a^* \rho_f C_f \boldsymbol{v} \cdot \nabla T - \nabla \cdot a^* \kappa_f \nabla T = \Xi_i^j (q_j + w_j).\end{aligned} \tag{2}$$

Here, the specific volume a^* accounts for the extension in the collapsed dimension(s), while subscript 0 denotes the initial value and f, s and e indicate fluid, solid and effective parameters, respectively.

Denoting the trace operator by $\mathtt{tr}(\cdot)$, the conditions on Γ_j are the three flux relationships:

$$\begin{aligned}v_j &= -\mathcal{K}_j(\Pi_l^j p_l - \Pi_h^j \mathtt{tr}(p_h)),\\ q_j &= -\kappa_j(\Pi_l^j T_l - \Pi_h^j \mathtt{tr}(T_h)),\\ w_j &= \begin{cases} \rho_f C_f v_j \Pi_h^j \mathtt{tr}(T_h) & \text{if } v_j > 0,\\ \rho_f C_f v_j \Pi_l^j T_l & \text{if } v_j \leq 0.\end{cases}\end{aligned} \tag{3}$$

Eqs. 1–3 are complemented by the internal boundary conditions $\mathtt{tr}(\boldsymbol{u}_h) = \Xi_h^j \boldsymbol{u}_j$, $\boldsymbol{v}_h \cdot \boldsymbol{n} = \Xi_h^j v_j$, $\boldsymbol{q}_h \cdot \boldsymbol{n} = \Xi_h^j q_j$ and $\boldsymbol{w}_h \cdot \boldsymbol{n} = \Xi_h^j w_j$ on $\partial\Omega_h^f$, and standard Dirichlet and Neumann conditions on the external boundaries.

The fracture deformation is described by relations between the contact traction on the fracture surface, $\boldsymbol{T}$, and the jump in displacement over the fracture. Denoting the displacements on the two sides of the interface by $\boldsymbol{u}_j^+$ and $\boldsymbol{u}_j^-$, the displacement jump is $[\![\boldsymbol{u}_j]\!] = \Xi_l^j \left(\boldsymbol{u}_j^+ - \boldsymbol{u}_j^-\right)$, and $[\![\delta \boldsymbol{u}]\!]$ denotes its increment. $[\![\boldsymbol{u}]\!]$ is also related to the aperture a and specific volume: for Ω_l, we set $a^* = a = [\![\boldsymbol{u}]\!]_n + a_0$, with a_0 denoting the initial aperture. When computing a for Ω_x, we use the mean of a for the adjacent cells of the intersecting fractures. Similarly, a^* is the product of the adjacent fracture apertures.

Since the fracture deformation depends on the traction caused by the *contact* between the two surfaces, we subtract the contribution from the pressure p_l on the fracture surfaces. Thus, the interface tractions on the two fracture surfaces and the traction balance are

$$\begin{aligned} \boldsymbol{T}_j^+ &= \Pi_h^j \sigma \cdot \boldsymbol{n}|_{\partial\Omega_h^+}, \\ \boldsymbol{T}_j^- &= \Pi_h^j \sigma \cdot \boldsymbol{n}|_{\partial\Omega_h^-}, \\ \Pi_l^j(\boldsymbol{T} - p_l \boldsymbol{n}_l) &= \boldsymbol{T}_j^+ = -\boldsymbol{T}_j^-, \end{aligned} \tag{4}$$

with $\boldsymbol{n}_l$ equalling the outward normal on the $+$ side. Denoting tangential and normal components of vectors on the fracture by subscripts τ and n, respectively, the fracture deformation for Ω_l is governed by three non-penetration relations and three friction law constraints:

$$\begin{array}{rll} [\![\boldsymbol{u}]\!]_n \leq 0 & ||\boldsymbol{T}_\tau|| \leq -FT_n & \\ T_n[\![\boldsymbol{u}]\!]_n = 0 & ||\boldsymbol{T}_\tau|| < -FT_n & \rightarrow\ [\![\delta\boldsymbol{u}]\!]_\tau = 0 \\ T_n \leq 0 & ||\boldsymbol{T}_\tau|| = -FT_n & \rightarrow\ \exists\, \zeta \in \mathbb{R}^- : \boldsymbol{T}_\tau = \zeta[\![\delta\boldsymbol{u}]\!]_\tau, \end{array} \tag{5}$$

with F denoting the friction coefficient of the fracture. Further detail on the fracture deformation is found in [2, 4].

3 Discretisation

Applying Implicit Euler for the temporal discretisation, the scalar conservation Eqs. 1 and 2 are discretised using a Multi-Point Flux Approximation [1] for the conductive terms and a first order upwind scheme for the advective term. The momentuum conservation equation and the $\nabla \cdot\, \boldsymbol{u}$ terms in the scalar conservation laws are discretised using the FV scheme introduced in [6]. The scheme, termed Multi-Point Stress Approximation (MPSA), is based on local momentum conservation and is formulated in terms of discrete cell centred pressures and displacement unknowns. Originally developed for the pure hydro-mechanical problem, the coupled discretisation approach can readily be extended to the THM case [7].

Thanks to the structure provided by the mixed-dimensional framework, the discretisation of the coupling fluxes of Eq. 3 consists of two simple tasks. Discrete projection operators transfer variables from higher-dimensional faces and lower-dimensional cells to the interface cells. The interface fluxes are then discretised directly using the projected variables, see [5].

The traction balance and fracture deformation relations of Eqs. 4 and 5 are formulated in terms of displacement and traction on the fracture surfaces. The former is included as a primary interface variable, and thus directly available. While the latter is not a primary variable, the FV framework is formulated in terms of face tractions, implying that in the discrete setting, the surface traction can be readily reconstructed from the primary variables. Specifically, we apply available discretisation operators to get contributions to the stress from displacements, pressures and temperatures in Ω_h, the interface variable $\boldsymbol{u}_j$, and conditions on external boundaries.

For the discretisation of the fracture deformation relations we first reformulate Eq. 5 as two nonlinear complementary functions and compute their derivatives. A semismooth Newton scheme is applied on the basis of the three sets

$$\begin{aligned} \mathcal{I}_n &= \{b \leq 0\} \\ \mathcal{I}_\tau &= \left\{ || - \boldsymbol{T}_\tau + c^* [\![\delta \boldsymbol{u}]\!]_\tau || < b \right\} \\ \mathcal{A} &= \left\{ || - \boldsymbol{T}_\tau + c^* [\![\delta \boldsymbol{u}]\!]_\tau || \geq b > 0 \right\}, \end{aligned} \tag{6}$$

which correspond to fracture cells which are open, sticking and sliding, respectively. c^* denotes a numerical parameter and $b = F\,(-T_n + c^* [\![\boldsymbol{u}]\!]_n)$ is the friction bound. For each fracture cell ν, this results in the following cell-wise constraints when computing iterate $k+1$ from the current iterate k:

$$\begin{aligned} \boldsymbol{T}^{k+1} &= \boldsymbol{0} && \nu \in \mathcal{I}_n \\ [\![\boldsymbol{u}^{k+1}]\!]_n &= 0 && \nu \in \mathcal{I}_\tau \cup \mathcal{A} \\ [\![\delta \boldsymbol{u}^{k+1}]\!]_\tau + (F [\![\delta \boldsymbol{u}^k]\!]_\tau / b^k) T_n^{k+1} &= [\![\delta \boldsymbol{u}^k]\!]_\tau && \nu \in \mathcal{I}_\tau \\ \boldsymbol{T}_\tau^{k+1} + L^k [\![\delta \boldsymbol{u}^{k+1}]\!]_\tau + F \boldsymbol{s}^k T_n^{k+1} &= \boldsymbol{r}^k + b^k \boldsymbol{s}^k && \nu \in \mathcal{A}. \end{aligned} \tag{7}$$

The coefficients L, $\boldsymbol{s}$ and $\boldsymbol{r}$ are functions of $[\![\delta \boldsymbol{u}^k]\!]_\tau$, $[\![\boldsymbol{u}^k]\!]_n$ and $\boldsymbol{T}^k$, and can thus be computed from the previous iterate and time step. For further details of the discretisation and implementation of the fracture deformation equations, we refer to [2].

In terms of implementation, we mention that the mixed-dimensional framework allows us to discretise each term for each subdomain or interface independently. We may thereby break the highly complex task of discretising the contact conditions with a coupled THM stress down in manageable tasks. For the global discretisation matrix, this manifests as a two-level block structure. The first level has the subdomains on the diagonal and interfaces on the off-diagonals. The second level corresponds to the primary variables, and has coupling terms between different variables on the off-diagonals, see Fig. 1.

The model is implemented for two- and three-dimensional problems in the open source simulation toolbox PorePy presented in [5], and run scripts for the example simulation presented in the following section may be found in the repository [8]. The simplicial spatial grid is constructed such that the lower-dimensional cells coincide with higher-dimensional faces through a back-end to Gmsh [3].

4 Results

To demonstrate the applicability of the model and discretisation, we present simulations of THM and fracture deformation effects for a 2d domain containing seven fractures, see Fig. 2 for the geometry and numbering of the fractures. The setup is designed to expose the method to a wide range of physical driving forces and thus probe the stability and performance of the simulation model.

Starting out from a homogeneous initial state for all primary variables, the simulation consists of four phases, where we study the effect of sequentially adding different driving forces. In phase I, the deformation is caused by a boundary displacement of $(0.002, -0.005)^T$ m applied at the top. To allow the system to reach equilibrium, this phase lasts from $t = -10\,000$ s to $t = 0$ s. In phase II, a pressure gradient is applied from left to right. At the end of the phase, at $t = 0.02$ s, the pressure has virtually reached a steady state, see Fig. 2. In phase III, we reduce the temperature at the left boundary of from 0 to $-100\,°$C and in phase IV we increase it to $100\,°$C, thus exploring both thermal expansion and contraction. At the end of each of the two last phases, at $t = 2.5$ s and $t = 5$ s, the domain has reached a close to uniform temperature.

For the end of each of the simulation phases, Fig. 3 shows the deformation state, i.e. whether $[\![\boldsymbol{u}]\!]_n$ and $[\![\boldsymbol{u}]\!]_\tau$ are nonzero for each fracture cell. These show that the state changes for all fractures, and that for each phase, at least three fractures change their state. Figure 4 shows the norm of $[\![\boldsymbol{u}]\!]_n$ and $[\![\boldsymbol{u}]\!]_\tau$ on each fracture throughout the simulation. Because of the time scale difference, the pressure phase is shown in the left plot and the temperature phase in the right one. The former shows gradual and moderate deformation for the fractures which have nonzero jumps at the end of phase I, and onset of sliding for fracture 3. The latter shows more complex deformation, displaying non-monotone jump evolution for several fractures, e.g. fracture 3 first opening and subsequently closing, and fracture 7 undergoing the reverse process.

Figure 4 also displays the number of Newton iterations required for convergence for each of the time steps. While fairly stable results are demonstrated, some increase may be observed whenever the fracture state changes more markedly, e.g. when the cooling is introduced at the onset of phase IV.

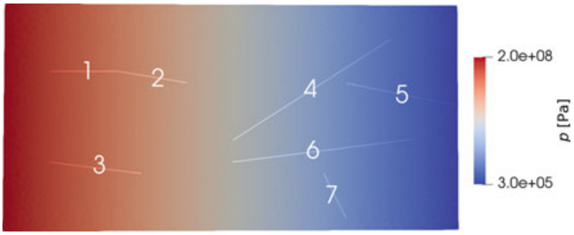


Fig. 2 The equilibrated pressure state at the end of phase II

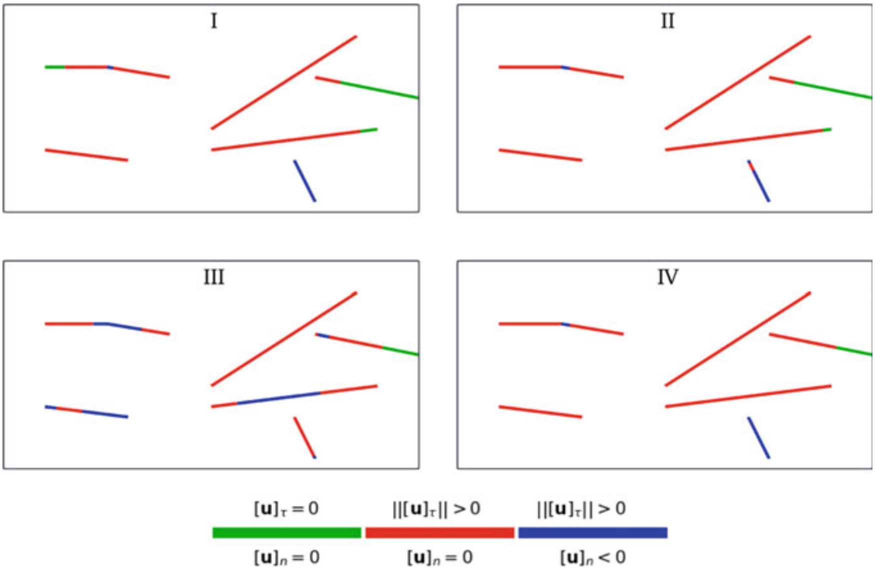


Fig. 3 The deformation state of the fractures at the end of the four phases

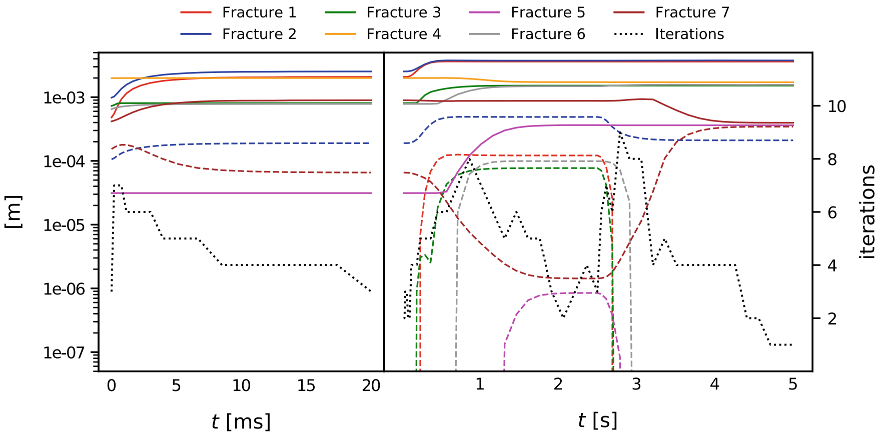


Fig. 4 Norm of displacement jumps at each fracture and number of Newton iterations for each time step. Phase II is shown to the left, with the state at the end of phase I shown at $t = 0$, and phases III and IV to the right. Normal jumps are shown in dashed lines and tangential jumps in solid lines

5 Conclusion

A FV framework for simulation of thermo-poroelasticity fully coupled to fracture deformation is presented. The contact mechanics problem is naturally discretised in terms of displacement and traction at the fracture faces, exploiting the availability of these in the FV formulation of the THM problem. A numerical example exhibiting complex THM interactions and fracture dynamics demonstrates both the range of the processes captured by the model, and its applicability for challenging problems.

References

1. Aavatsmark, I.: An introduction to multipoint flux approximations for quadrilateral grids. Comput. Geosci. (2002)
2. Berge, R.L., Berre, I., Keilegavlen, E., Nordbotten, J.M., Wohlmuth, B.: Finite volume discretization for poroelastic media with fractures modeled by contact mechanics. Int. J. Numer. Methods Eng. (2020)
3. Geuzaine, C., Remacle, J.-F.: Gmsh: A 3-D finite element mesh generator with built-in pre- and post-processing facilities. Int. J. Numer. Methods Eng. (2009)
4. Hüeber, S., Stadler, G., Wohlmuth, B.I.: A primal-dual active set algorithm for three-dimensional contact problems with coulomb friction. SIAM J. Sci. Comput. (2018)
5. Keilegavlen, E., Berge, R., Fumagalli, A., Starnoni, M., Stefansson, I., Varela, J., Berre, I.: PorePy: an open-source software for simulation of multiphysics processes in fractured porous media. arXiv:1908.09869 (2019)
6. Nordbotten, J.M.: Stable cell-centered finite volume discretization for Biot equations. SIAM J. Numer. Anal. (2016)
7. Nordbotten, J.M., Keilegavlen, E.: An introduction to multi-point flux (MPFA) and stress (MPSA) finite volume methods for thermo-poroelasticity. arXiv:2001.01990 (2020)
8. Run scripts at https://github.com/IvarStefansson/Finite-volume-discretisation-of-fracture-deformation-in-thermo-poroelastic-media.git
9. Salimzadeh, S., Paluszny, A., Nick, Hamidreza M., Zimmerman, R.W.: A three-dimensional coupled thermo-hydro-mechanical model for deformable fractured geothermal systems. Geothermics (2018)

A Control Volume Finite Element Formulation with Subcell Reconstruction for Phase-Field Fracture

Juan Michael Sargado

Abstract We present a control volume finite element formulation for phase-field brittle fracture based on unstructured simplex meshes. Linear finite elements are employed for the linear momentum equation, while a cell-centered finite volume method based on the two-point flux approximation scheme is used to discretize the phase-field equation. Additionally, we perform linear reconstruction of the phase-field variable over subcells of the original control volumes in order to better model gradient discontinuities. This yields a higher order scheme that also gives more conservative predictions of critical loads compared to existing low-order methods.

Keywords Damage · Brittle fracture · Finite volume · Finite element

MSC (2010) 74R05 · 65N08 · 65N30

1 Introduction

The simulation of fracture evolution in solids presents unique challenges due to the generally a priori unknown nature of crack propagation paths, and the possible occurrence of topological changes in the form of crack branching and coalescence. Variational phase-field models [2] address these issues by modeling fractures as diffuse entities through a scalar phase-field variable $\phi \in [0, 1]$, where 0 and 1 represent respectively the fully intact and broken states. This results in the regularization of displacement discontinuities across fractures and eliminates the need to explicitly track the evolution of lower-dimensional entities (i.e. crack surfaces), at the cost of introducing an additional equation governing phase-field evolution. The latter is in turn parametrized by a length scale that controls the amount of crack regularization.

J. M. Sargado (✉)
Department of Energy Resources, University of Stavanger, Stavanger, Norway
e-mail: juan.m.sargado@uis.no

NORCE Norwegian Research Centre AS, Bergen, Norway

R. Klöfkorn et al. (eds.), *Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples*, Springer Proceedings in Mathematics & Statistics 323,
https://doi.org/10.1007/978-3-030-43651-3_49

In the classical second-order formulation of Bourdin-Francfort-Marigo [2], the phase-field profile contains gradient discontinuities or kinks at locations corresponding to fully developed cracks. Such features are difficult to reproduce with conventional low-order finite element schemes and may lead to inaccurate solutions unless aggressive mesh refinement is performed with respect to the phase-field length scale. On the other hand, a discretization framework combining P_1 finite elements for the linear momentum balance and cell-centered (P_0) finite volumes for the phase-field equation was proposed in a recent work [12]. This can be interpreted as a variant of the control volume finite element method in the sense of [9], and is based on the idea that phase-field regularization of fractures eliminates crack tip stress singularities [13]. Consequently, mesh size restrictions related to the accuracy of numerical results become tied to the problem of resolving gradient discontinuities in the phase-field. As the two-point flux approximation scheme implicitly allows for gradient discontinuities within the control volumes themselves, the presence of kinks in the phase-field profile is naturally handled by the cell-centered FV formulation in conjunction with TPFA. This in turn leads to better accuracy of numerical solutions over the same mesh compared to a linear FE approximation of both the displacement and phase-field. Furthermore the formulation is computationally cheap, as global matrices arising from the former scheme are considerably sparser then their FE counterparts.

Previous works have shown that FV schemes of higher order can be obtained by means of reconstruction techniques [3, 5, 6]. In this study, we seek to improve upon the formulation given in [12] by performing linear reconstructions of the phase-field over subcells of the original control volumes. While this generally yields a discontinuous representation of the phase-field across subcells within a given control volume, said approach nevertheless better models kinks at locations corresponding to fully developed fractures resulting in more accurate predictions with regard to critical loads.

2 Model Equations

In the phase-field approach to brittle fracture, a regularized total energy functional pertaining to a homogeneous body Ω is assumed to exist which consists of bulk and surface terms together with the external work due to applied forces, i.e.

$$\begin{aligned}\Pi_\ell\left(\mathbf{u},\phi\right) = \int_\Omega \psi\left(\varepsilon\left(\mathbf{u}\right),\phi\right)\mathrm{d}\Omega + \int_\Omega \mathscr{G}_c\left(\frac{\phi^2}{2\ell} + \frac{\ell}{2}\nabla\phi\cdot\nabla\phi\right)\mathrm{d}\Omega \\ - \int_\Omega \mathbf{b}\cdot\mathbf{u}\,\mathrm{d}\Omega - \int_{\partial\Omega^t} \mathbf{t}\cdot\mathbf{u}\,\mathrm{d}\partial\Omega\end{aligned} \tag{1}$$

wherein $\mathbf{u}$ and ϕ are respectively the displacement and phase-field variables, $\psi\left(\varepsilon,\phi\right)$ is the damage-dependent bulk energy density with ε denoting the symmetric small-strain tensor, $\mathscr{G}_c$ is the critical energy release rate and ℓ is the phase-field regularization parameter. In particular we adopt the following form for ψ from Amor et al. [1] that approximates unilateral contact of fracture surfaces:

$$\psi\left(\varepsilon, \phi\right) = \psi_0^-\left(\varepsilon\right) + g\left(\phi\right)\psi_0^+\left(\varepsilon\right) = \frac{\kappa}{2}\langle \operatorname{tr}\varepsilon\rangle_-^2 + g\left(\phi\right)\left[\frac{\kappa}{2}\langle \operatorname{tr}\varepsilon\rangle_+^2 + \mu\,\varepsilon_D : \varepsilon_D\right]. \tag{2}$$

In the above expression, $\langle \bullet \rangle_{\pm} = \frac{1}{2}\left(\bullet \pm |\bullet|\right)$, tr denotes the trace and ε_D is the deviatoric strain. Material constants κ and μ refer respectively to the bulk and shear moduli, while $g\left(\phi\right)$ is an energy degradation function that is monotonically decreasing with $g\left(0\right) = 1$ and $g\left(1\right) = g'\left(1\right) = 0$. We make use of the quadratic degradation function $g\left(\phi\right) = \left(1-\phi\right)^2$ in the current study due to its simplicity and general familiarity with the research community, however alternative forms have been recently introduced in order to improve accuracy of numerical simulations [11].

Energy balance implies $\delta\Pi_\ell = 0$ for any variation $\delta\mathbf{u}$ and $\delta\phi$ in the arguments of Π_ℓ. This leads to a strongly coupled system of equations, given in weak form as

$$\int_\Omega \left[\sigma_0^-\left(\varepsilon\right) + g\left(\phi\right)\sigma_0^+\left(\varepsilon\right)\right] : \delta\varepsilon\,\mathrm{d}\Omega = \int_\Omega \mathbf{b}\cdot\delta\mathbf{u}\,\mathrm{d}\Omega + \int_{\partial\Omega^t} \mathbf{t}\cdot\delta\mathbf{u}\,\mathrm{d}\Omega \tag{3}$$

$$\int_\Omega g'\left(\phi\right)\psi_0^+\left(\varepsilon\right)\mathrm{d}\Omega + \int_\Omega \mathscr{G}_c\left(\frac{1}{\ell}\phi\,\delta\phi + \ell\nabla\phi\cdot\nabla\delta\phi\right)\mathrm{d}\Omega = 0. \tag{4}$$

Equation (3) corresponds to the balance of linear momentum, where the positive and negative elastic stress projections are derived from (2) and given by

$$\sigma_0^+ = \kappa\langle \operatorname{tr}\varepsilon\rangle_+\mathbf{I} + \mu\,\varepsilon_D \qquad \sigma_0^- = \kappa\langle \operatorname{tr}\varepsilon\rangle_-\mathbf{I}, \tag{5}$$

where $\mathbf{I}$ is the 2nd order identity tensor. On the other hand, (4) governs the evolution of ϕ by imposing incremental energy balance with respect to crack growth.

It has been pointed out in [2] that while (1) is non-convex with respect to the pair $(\mathbf{u}, \phi)$, fixing either $\mathbf{u}$ or ϕ restores convexity of the energy with respect to the remaining argument which allows the coupled system of equations to be solved via an alternate minimization strategy. In order to enforce irreversibility of crack growth, we use a history field $\mathscr{H}\left(t\right) = \max_{s\in[0,t]}\psi_0^+\left(\varepsilon\left(s\right)\right)$ in place of ψ_0^+ in the phase-field equation [7]. Recasting (4) in strong form yields the PDE

$$\mathscr{G}_c\ell\nabla^2\phi - \frac{\mathscr{G}_c}{\ell}\phi = g'\left(\phi\right)\mathscr{H} \quad \text{in } \Omega\,, \tag{6}$$

together with the Neumann boundary condition $\nabla\phi\cdot\mathbf{n} = 0$ on $\partial\Omega$. By integrating the above expression over some internal subdomain Ω_k and applying the divergence theorem, we obtain the control volume form

$$\int_{\Omega_k} g'\left(\phi\right)\mathscr{H}\,\mathrm{d}\Omega + \int_{\Omega_k}\frac{\mathscr{G}_c}{\ell}\phi\,\mathrm{d}\Omega - \int_{\partial\Omega_k}\mathscr{G}_c\ell\,\nabla\phi\cdot\mathbf{n}\,\mathrm{d}\Gamma = 0. \tag{7}$$

3 Numerical Discretization

We employ a finite element discretization of the linear momentum equation in which displacements are approximated using piecewise linear basis functions over unstructured simplex meshes. The discrete residual corresponding to (3) is then given by

$$\mathbf{r}^{u}\left(\varepsilon,\phi\right)=\int_{\Omega}\mathbf{B}^{\mathrm{T}}\left[\sigma_0^{-}\left(\varepsilon\right)+g\left(\phi\right)\sigma_0^{+}\left(\varepsilon\right)\right]\mathrm{d}\Omega-\int_{\Omega}\mathbf{N}^{\mathrm{T}}\mathbf{b}\,\mathrm{d}\Omega-\int_{\partial\Omega^t}\mathbf{N}^{\mathrm{T}}\mathbf{t}\,\mathrm{d}\Omega, \tag{8}$$

in which the matrices **N** and **B** are defined as

$$\mathbf{N}_I=\begin{bmatrix} N_I & 0\\ 0 & N_I\end{bmatrix} \qquad \mathbf{B}_I=\begin{bmatrix} N_{I,x} & 0\\ 0 & N_{I,y}\\ N_{I,y} & N_{I,x}\end{bmatrix}, \quad I=1,\ldots,m \tag{9}$$

with m being the number of nodes in the mesh, and N_I the basis function associated with node I. Meanwhile, the control volume form of the phase-field equation is approximated via a cell-centered FV scheme, with the surface integral in (7) discretized using classical TPFA. That is, the normal flux q_k^i through face Γ_k^i of an interior cell Ω_k is constructed using the phase-field value ϕ_k at the cell center Ω_k, and its value ϕ_k^i at the center of adjoining control volume Ω_k^i across Γ_k^i. Denoting by $\bar{\phi}_k^i$ the value of ϕ on Γ_k^i and assuming that $\phi_k^i > \bar{\phi}_k^i > \phi_k$, the flux can be approximated as

$$q_k^i=\mathscr{G}_{ck}\ell_k\frac{\bar{\phi}_k^i-\phi_k}{\|\bar{\mathbf{x}}_k^i-\mathbf{x}_k\|}\|\Gamma_k^i\|=\mathscr{G}_{c\,k}^{\,i}\ell_k^i\frac{\phi_k^i-\bar{\phi}_k^i}{\|\mathbf{x}_k^i-\bar{\mathbf{x}}_k^i\|}\|\Gamma_k^i\|, \tag{10}$$

wherein $\mathbf{x}_k$, $\mathbf{x}_k^i$ and $\bar{\mathbf{x}}_k^i$ are respectively the centers of Ω_k, Ω_k^i and Γ_k^i, and $\|\Gamma_k^i\|$ is the measure of Γ_k^i. Combining the two expressions for q_k^i in (10) yields

$$q_k^i=T_k^i\left(\phi_k^i-\phi_k\right), \tag{11}$$

in which the transmissibility coefficient associated with the surface Γ_k^i is given by

$$T_k^i=\frac{\|\Gamma_k^i\|}{\dfrac{\|\bar{\mathbf{x}}_k^i-\mathbf{x}_k\|}{\mathscr{G}_{ck}\ell_k}+\dfrac{\|\mathbf{x}_k^i-\bar{\mathbf{x}}_k^i\|}{\mathscr{G}_{c\,k}^{\,i}\ell_k^i}}. \tag{12}$$

3.1 Discontinuous Representation of ϕ Over Ω_k

In classical cell-centered FV, the primary unknown is usually assumed to be piecewise constant over the control volumes. Such assumption is however not optimal when dealing with the fracture phase-field model used in the current study, as it is known that the solution to (6) for a fully developed crack contains gradient discontinuities in the form of cusps. Instead, we incorporate the latter knowledge into the numerical formulation by assuming ϕ to be linear over subcells of the original control volume. This implies that ϕ is generally discontinuous across said subcells.

Let $\hat{\Omega}_k^i$ be a subcell of Ω_k (triangle in 2D, tetrahedron in 3D) with apex at $\mathbf{x}_k$ and base coinciding with face Γ_K^i so that $\bigcup_{i=1}^N \hat{\Omega}_k^i = \Omega_k$. By combining (10) and (12), we obtain the phase-field value at Γ_k^i as

$$\bar{\phi}_k^i = \phi_k + R_k^i \left(\phi_k^i - \phi_k\right), \tag{13}$$

where the factor R_k^i is given by

$$R_k^i = \frac{\|\bar{\mathbf{x}}_k^i - \mathbf{x}_k\|}{\mathscr{G}_{ck}\ell_k \|\Gamma_k^i\|} T_k^i. \tag{14}$$

As ϕ is assumed linear over $\hat{\Omega}_k^i$, its value at the center of $\hat{\Omega}_k^i$ can be calculated as

$$\hat{\phi}_k^i = w\phi_k + (1-w)\,\bar{\phi}_k^i \tag{15}$$

in which $w = 1/\left(D+1\right)$, with D being the problem dimensionality. Combining the above expression with (13) yields

$$\hat{\phi}_k^i = \phi_k + (1-w)\, R_k^i \left(\phi_k^i - \phi_k\right). \tag{16}$$

Using the above approximation for volume integrals, the residual corresponding to the discrete form of (7) for an interior cell Ω_k^i becomes

$$r_k = \sum_{i=1}^N \left[g'\left(\hat{\phi}_k^i\right) \psi_0\left(\varepsilon_k\right) + \frac{\mathscr{G}_{ck}}{\ell_k} \hat{\phi}_k^i \right] \|\hat{\Omega}_k^i\| + \sum_{i=1}^N T_k^i \left(\phi_k - \phi_k^i\right) = 0. \tag{17}$$

As the linear reconstruction of ϕ over subcell $\hat{\Omega}_k^i$ involves the same set of unknowns ϕ_k and ϕ_k^i that are used to approximate the flux through Γ_k^i, the sparsity profile for the resulting global coefficient matrix is identical to the case where ϕ is assumed constant over Ω_k. However said coefficient matrix becomes non-symmetric owing to the factor $\|\bar{\mathbf{x}}_k^i - \mathbf{x}_k\|$ being present in R_k^i.

4 Numerical Results

We consider two numerical examples. The first involves solution of the phase-field equation in 1D with an internal condition corresponding to a fully developed crack, whereas the second example deals with the well-known Miehe shear benchmark

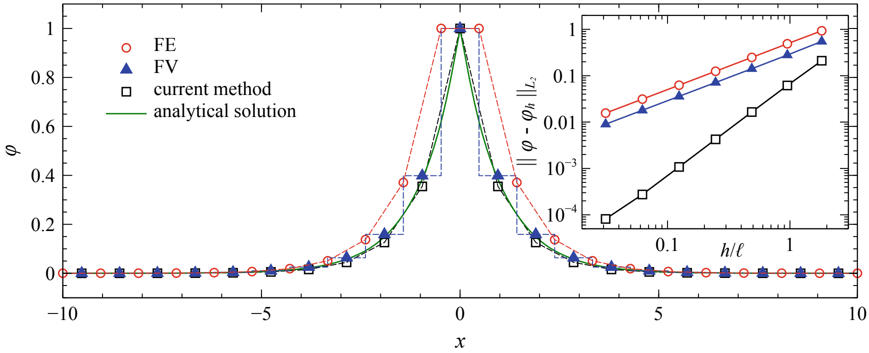


Fig. 1 Phase-field profile for bar with crack at $x = 0$, obtained with linear finite elements, classical cell-centered finite volumes and the current method for a mesh consisting of 21 cells. Inset shows convergence behavior of the aforementioned methods with respect to mesh refinement

problem. The proposed formulation is implemented in the open-source multiphysics framework BROOMStyx [10] together with other methods in order to compare results and run times. All simulations are carried out on a shared-memory parallel computer equipped with a 6-core processor running at 3.20 GHz base frequency.

4.1 Stationary Crack in 1D

For a uniform cylindrical bar having endpoints at $x \pm 10$ and fully cut by a crack at $x = 0$, the corresponding phase-field profile for $\ell = 1$ is given by $\phi(x) = \exp(-|x|)$, which solves the homogeneous BVP

$$\phi''(x) - \phi(x) = 0 \; \forall x \in (-10, 10), \quad \phi'(\pm 10) = 0, \quad \phi(0) = 1. \tag{18}$$

Following [12], we run simulations on a sequence of mesh refinements to determine convergence rates. L_2-norms of errors with respect to the analytical solution are plotted in Fig. 1, along with the superposed solutions from P_1-FE, cell-centered FV and the current method for a particular mesh refinement. We can observe that the proposed formulation displays a higher rate of convergence compared to classical P_1-FE and cell-centered FV methods.

4.2 Miehe Shear Benchmark

We investigate performance of the proposed method in simulating a benchmark problem from [8] involving fracture propagation in a notched square specimen subjected

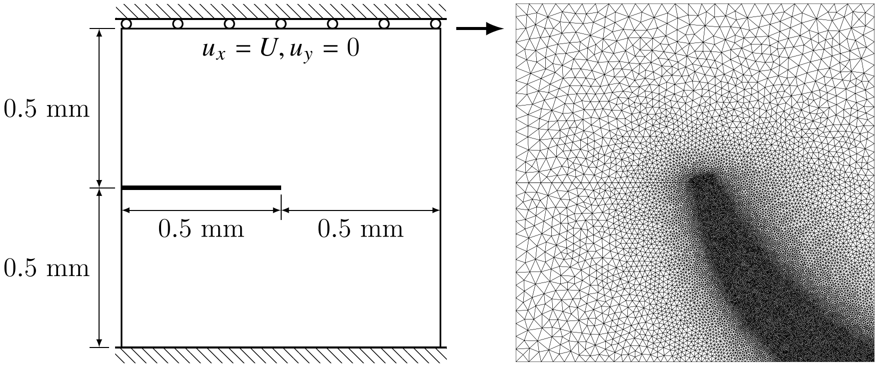


Fig. 2 Miehe shear benchmark problem showing geometry and boundary conditions (left), and typical unstructured mesh (right)

to shearing boundary conditions. The relevant geometry is shown in Fig. 2 along with a typical discretization used for analysis. The problem domain is discretized using unstructured simplex meshes that constitute admissible FV discretizations in the sense of [4]. Said meshes incorporate the initial fracture and are locally pre-refined around the expected crack path such that cells in the refined region have sides of length ℓ/n, with $n \in \{1, 2, 4, 8\}$. The following material parameters are used: $E = 210$ GPa, $\nu = 0.3$, $\mathscr{G}_c = 2.7$ N/mm and $\ell = 0.0075$ mm. Furthermore, the prescribed horizontal displacement at the top surface is applied in increments of 1×10^{-4} mm up to $u_x = 0.0085$ mm, and thereafter in smaller increments of 1×10^{-5} mm up to the final displacement of $U = 0.0125$ mm. For comparison, we also carry out simulations using an equal-order (P_1) FE discretization of the coupled system (3)–(4) in addition to the FE-FV formulation discussed in [12] on the same set of discretizations.

Run times and resulting peak loads for different mesh refinements in conjunction with the aforementioned methods are shown in Table 1, where it can be seen that for a given value of ℓ/h the proposed method (designated CVFE-LR) gives consistently lower predictions of peak loads compared to both a pure FE discretization of the coupled system and the FE-FV formulation described in [12]. The same trend can be observed in the post-peak behavior of the resulting load-displacement curves that are plotted in Fig. 3. In particular the post-peak curves for FE-FE at $\ell/h = 8$, FE-FV at $\ell/h = 4$ and CVFE-LR at $\ell/h = 1$ are in near vicinity of each other, notwithstanding over-/undershoots in the latter due to relative mesh coarseness. This result is in line with behavior predicted in Fig. 1, which shows similar error norms for the respective combination of ℓ/h values and formulations.

Table 1 Summary of results for the Miehe shear benchmark

Formulation	ℓ/h	nDOF	Run time (h : min : s)	Peak load (N)
FE-FE	2	47,941	1 : 13 : 49	408.86
FE-FE	4	143,198	5 : 25 : 43	394.56
FE-FE	8	475,827	26 : 44 : 26	387.90
FE-FV	1	24,445	0 : 20 : 35	401.46
FE-FV	2	63,696	1 : 05 : 49	389.00
FE-FV	4	190,641	4 : 35 : 34	377.44
FE-FV	8	634,050	20 : 56 : 22	377.37
CVFE-LR	1	24,445	0 : 21 : 44	380.97
CVFE-LR	2	63,696	1 : 00 : 17	374.73
CVFE-LR	4	190,641	4 : 17 : 43	368.96
CVFE-LR	8	634,050	19 : 31 : 18	371.20

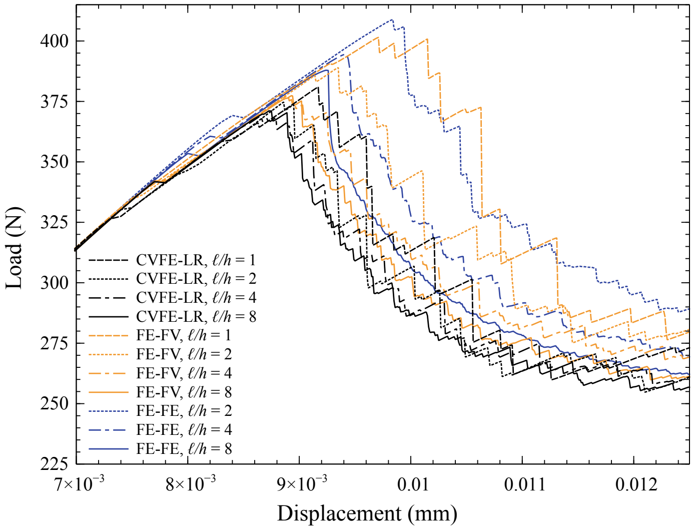


Fig. 3 Truncated load displacement curves for the Miehe shear benchmark with varying levels of mesh refinement for a fixed value of ℓ

End-of-simulation phase-field profiles for the three methods at the aforementioned ℓ/h ratios are displayed in Fig. 4. The simulated crack paths are nearly identical for all three cases, however a finer discretization of the domain with respect to the phase-field length scale results in a smoother crack trajectory. In particular the crack path predicted by the current method using a mesh refinement of $\ell/h = 1$ contains a lot of small twists, resulting in oscillatory behavior of the corresponding load-displacement curve in Fig. 3.

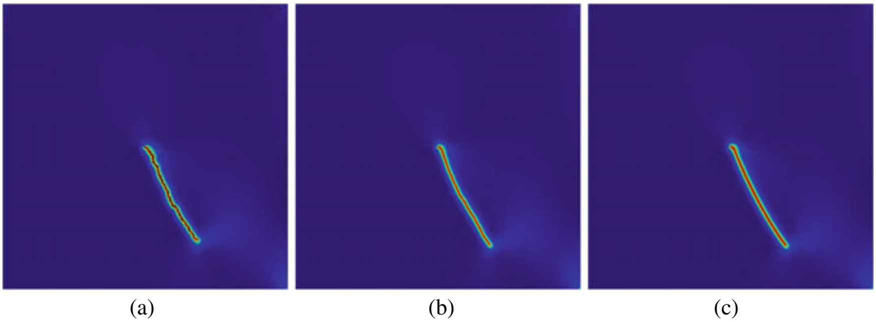


Fig. 4 Final crack trajectories corresponding to an upper boundary displacement of $u_x = 0.0125$ mm for **a** the current method with $\ell/h = 1$ **b** cell-centered FV with $\ell/h = 4$, and **c** P_1-FE with $\ell/h = 8$

Acknowledgements The author acknowledges the Research Council of Norway and the industry partners, ConocoPhillips Skandinavia AS, Aker BP ASA, Vår Energi AS, Equinor ASA, Neptune Energy Norge AS, Lundin Norway AS, Halliburton AS, Schlumberger Norge AS, and Wintershall DEA, of The National IOR Centre of Norway for support.

References

1. Amor, H., Marigo, J.J., Maurini, C.: Regularized formulation of the variational brittle fracture with unilateral contact: numerical experiments. J. Mech. Phys. Solids **57**, 1209–1229 (2009)
2. Bourdin, B., Francfort, G.A., Marigo, J.J.: Numerical experiments in revisited brittle fracture. J. Mech. Phys. Solids **48**, 797–826 (2000)
3. Dumbser, M., Balsara, D.S., Toro, E.F., Munz, C.D.: A unified framework for the construction of one-step finite volume and discontinuous galerkin schemes on unstructured meshes. J. Comput. Phys. **227**, 8209–8253 (2008)
4. Eymard, R., Gallouët, T., Herbin, R.: Finite volume methods. In: Handbook of Numerical Analysis, vol. 7, pp. 713–1018. Elsevier (2000)
5. Klöfkorn, R., Kvashchuk, A., Nolte, M.: Comparison of linear reconstructions for second-order finite volume schemes on polyhedral grids. Comput. Geosci. **21**(5–6) (2017)
6. Kučera, V.: Higher-order reconstruction: from finite volumes to discontinuous galerkin. In: Fořt, J., Fürst, J., Halama, J., Herbin, R., Hubert, F. (eds.) Finite Volumes for Complex Applications VI—Problems & Perspectives, pp. 613–621. Springer, Berlin Heidelberg (2011)
7. Miehe, C., Hofacker, M., Welschinger, F.: A phase field model for rate-independent crack propagation: robust algorithmic implementation based on operator splits. Comput. Methods Appl. Mech. Engrg. **199**, 2765–2778 (2010)
8. Miehe, C., Welschinger, F., Hofacker, M.: Thermodynamically consistent phase-field models of fracture: variational principles and multi-field FE implementations. Int. J. Numer. Methods Eng. **83**(10), 1273–1311 (2010)
9. Salinas, P., Pavlidis, D., Xie, Z., Jacquemyn, C., Melnikova, Y., Jackson, M.D., Pain, C.C.: Improving the robustness of the control volume finite element method with application to multiphase porous media flow. Int. J. Numer. Meth. Fluids **85**, 235–246 (2017)
10. Sargado, J.M.: A new object-oriented framework for solving multiphysics problems via combination of different numerical methods. arXiv:1905.00104 [cs.MS] (2019)

11. Sargado, J.M., Keilegavlen, E., Berre, I., Nordbotten, J.M.: High-accuracy phase-field models for brittle fracture based on a new family of degradation functions. J. Mech. Phys. Solids **111**, 458–489 (2018)
12. Sargado, J.M., Keilegavlen, E., Berre, I., Nordbotten, J.M.: A combined finite element-finite volume framework for phase-field fracture. arXiv:1904.12395 [math.NA] (2019)
13. Sicsic, P., Marigo, J.J.: From gradient damage laws to griffith's theory of crack propagation. J. Elast. **113**, 55–74 (2013)

A Conservative Phase-Field Model for Reactive Transport

Carina Bringedal

Abstract We present a phase-field model for single-phase flow and reactive transport where ions take part in mineral precipitation/dissolution reactions. The evolving interface between fluid and mineral is approximated by a diffuse interface, which is modeled using an Allen–Cahn equation. As the original Allen–Cahn equation is not conservative, we apply a reformulation ensuring conservation of the phase-field variable and address the sharp-interface limit of the reformulated model. This model is implemented using a finite volume scheme and the discrete conservation of the reformulated Allen–Cahn equation is shown. Numerical examples show how the discrete phase-field variable is conserved up to the chemical reaction.

Keywords Allen–Cahn · Reactive transport · Discrete conservation

MSC (2010) 76D05 · 74N20 · 65M08

1 Introduction

We consider single-phase flow with solute transport, where ions can form a mineral and hence leave the fluid phase. Also, minerals in the mineral phase can dissolve, releasing ions to the fluid phase. We account for the time evolution of the mineral and fluid phases. This evolution is not known a-priori as it depends on the considered reactions, hence we obtain a free-boundary problem. In [10], existence and uniqueness of a weak solution for such a free-boundary model is proved in a one-dimensional domain. In [8] a free-boundary model for precipitation and dissolution is included in a two-dimensional model using a level-set formulation.

C. Bringedal (✉)
Stuttgart Center for Simulation Technology (SimTech),
Institute for Modelling Hydraulic and Environmental Systems (IWS),
University of Stuttgart, Pfaffenwaldring 5a, 70569 Stuttgart, Germany
e-mail: carina.bringedal@iws.uni-stuttgart.de

Computational Mathematics (CMAT), Hasselt University, Hasselt, Belgium

R. Klöfkorn et al. (eds.), *Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples*, Springer Proceedings in Mathematics & Statistics 323,
https://doi.org/10.1007/978-3-030-43651-3_50

These approaches apply a sharp interface between the mineral and fluid. Alternatively, the transition between mineral and fluid can be considered as diffuse. Diffuse-interface models for reactive and diffusive transport have been formulated and analyzed in [9, 11], and later extended to fluid flow in [2]. In these papers the original Allen–Cahn equation [1] describes the evolution of the diffuse interface.

The Allen–Cahn equation is derived from mean-curvature flow. Hence, the interface can evolve due to curvature effects, which may be desirable from a chemical point of view [13]. It fulfills the max/min principle, but is in its original form not conservative meaning that the volume of the considered phases will not remain constant. Reformulations to ensure conservation of the phases for other applications have been suggested and analyzed in the form of a nonlocal term or a Lagrange multiplier [4, 7, 12]. These papers do not consider mineral precipitation/dissolution, but we will base on their work to find a reformulation applicable to these processes.

The structure of the paper is as following: In Sect. 2 we present the original phase-field model from [2] that we will build upon, while in Sect. 3 the conservative reformulation and its sharp-interface limit is addressed. Section 4 formulates a finite volume scheme and we show that the reformulated Allen–Cahn equation is discretely conservative. Finally we show some numerical examples in Sect. 5.

2 The Original Phase-Field Model and Its Sharp-Interface Limit

The original model, as formulated in [2] is, for $\mathbf{x} \in \Omega$, $t > 0$:

$$\lambda^2 \partial_t \phi + \gamma P'(\phi) = \gamma \lambda^2 \nabla^2 \phi - 4\lambda \phi (1 - \phi) \frac{1}{u^*} f(u), \tag{1a}$$

$$\nabla \cdot (\phi \mathbf{v}) = 0, \tag{1b}$$

$$\rho_f \partial_t (\phi \mathbf{v}) + \rho_f \nabla \cdot (\phi \mathbf{v} \otimes \mathbf{v}) = -\phi \nabla p + \mu_f \phi \nabla^2 (\phi \mathbf{v}) - \frac{1}{\lambda} g(\phi) \mathbf{v} + \frac{1}{2} \rho_f \mathbf{v} \partial_t \phi, \tag{1c}$$

$$\partial_t \left(\phi (u - u^*) \right) + \nabla \cdot (\phi \mathbf{v} u) = D \nabla \cdot (\phi \nabla u). \tag{1d}$$

Here, Ω is the combined fluid and mineral domain and hence constant in time. The phase field ϕ approaches 1 in the fluid and 0 in the mineral while λ denotes the width of the diffuse zone separating the two phases. The double-well potential is $P(\phi) = 8\phi^2(1 - \phi)^2$ and γ is the diffusivity of the interface. Further, $\mathbf{v}$ is the fluid velocity and p the pressure, while ρ_f and μ_f are the constant fluid density and viscosity. Since the flow equations are solved also for the mineral part of the domain, the monotonously decreasing interpolation term $g(\phi)$ fulfilling $g(1) = 0$ and $g(0) > 0$ is included to ensure zero flow in the mineral. The solute concentration is denoted as u, D is its diffusivity, and the constant mineral concentration is u^*. The mineral precipitation and dissolution reaction rate is $f(u) = k(u^2/u_{\mathrm{eq}}^2 - 1)$, where u_{eq} is a given equilibrium concentration and k a reaction constant.

The sharp-interface limit of the model (1) is derived by matched asymptotic expansions in [2]. The background of this procedure can be found in [3]. We let $\Omega_f(t)$ denote the domain where $\phi \to 1$ and $\Omega_m(t)$ where $\phi \to 0$ as $\lambda \to 0$. By separating between these two regions, the model (1) reduces to

$$\begin{aligned} \nabla \cdot \mathbf{v} &= 0 && \text{in } \Omega_f(t), && \text{(2a)}\\ \rho_f \partial_t \mathbf{v} + \rho_f \nabla \cdot (\mathbf{v} \otimes \mathbf{v}) + \nabla p &= \mu_f \nabla^2 \mathbf{v} && \text{in } \Omega_f(t), && \text{(2b)}\\ \partial_t u + \nabla \cdot (\mathbf{v} u) &= D \nabla^2 u && \text{in } \Omega_f(t), && \text{(2c)}\\ \mathbf{v} &= \mathbf{0}, && \text{in } \Omega_m(t), && \text{(2d)} \end{aligned}$$

as $\lambda \to 0$. Through inner expansions and hence investigating the behavior near the diffuse transition zone, it is found that, as $\lambda \to 0$ [2]:

$$\begin{aligned} v_n &= -\gamma\kappa - \frac{1}{u^*} f(u) && \text{on } \Gamma(t), && \text{(3a)}\\ \mathbf{v} &= \mathbf{0} && \text{on } \Gamma(t), && \text{(3b)}\\ v_n(u^* - u) &= \mathbf{n} \cdot D \nabla u && \text{on } \Gamma(t), && \text{(3c)} \end{aligned}$$

where v_n is the normal velocity of the interface $\Gamma(t)$ and the curvature κ introduces the curvature-driven motion. The normal vector $\mathbf{n}$ points into the mineral.

However, it is clear from (3a) that the interface evolution is not conservative as the curvature-driven motion will alter the size of the fluid/mineral domains. By applying homogeneous Neumann boundary conditions on $\partial\Omega$ we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t} \int_\Omega \phi d\mathbf{x} = \int_\Omega \left(-\frac{1}{\lambda^2} \gamma P'(\phi) - \frac{4}{\lambda} \phi (1 - \phi) \frac{1}{u^*} f(u) \right) d\mathbf{x},$$

which is non-zero even without chemical reactions.

3 Conservative Phase-Field Model

We now formulate a conservative phase-field model based on the reformulation considered in [12] for phase separation, where also well-posedness of the reformulation was assessed. We replace the phase-field equation (1a) by

$$\lambda^2 \partial_t \phi + \gamma P'(\phi) = \gamma \lambda^2 \nabla^2 \phi - 4\lambda \phi (1 - \phi) \frac{1}{u^*} f(u) + \frac{\gamma}{|\Omega|} \int_\Omega P'(\phi) d\mathbf{x}, \quad (4)$$

where $|\Omega|$ is the size of the considered domain. The Eqs. (1b)–(1d) are left unchanged. The reformulated Eq. (4) fulfills the global conservation property:

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{\Omega}\phi d\mathbf{x} = \int_{\Omega}\left(-\frac{4}{\lambda}\phi(1-\phi)\frac{1}{u^*}f(u)\right)d\mathbf{x}. \tag{5}$$

Now the curvature-driven motion does not affect the total amount of ϕ anymore. Interpreting the integrated phase field as porosity, means that porosity can only change due to the mineral precipitation and dissolution.

We address the sharp-interface limit of the reformulated phase-field equation (4) by following similar steps as in [4], where a conservative Allen–Cahn equation without chemical reactions was addressed. The Eq. (4) is first split in two equations:

$$\lambda^2\partial_t\phi + \gamma P'(\phi) = \gamma\lambda^2\nabla^2\phi - 4\lambda\phi(1-\phi)\frac{1}{u^*}f(u) + \gamma\lambda\xi(t), \tag{6a}$$

$$\xi(t) = \frac{1}{\lambda}\frac{1}{|\Omega|}\int_{\Omega}P'(\phi)d\mathbf{x}. \tag{6b}$$

The integral in (6b) is small and will decrease as λ decreases. Hence, it is shown in [4] that $\xi(t) = O(\lambda^0)$ as $\lambda \to 0$. The lowest order terms of the outer expansions of (6a) still lead to ϕ approaching 0 and 1 in the mineral and fluid domains as earlier.

For the inner expansions [3], following steps similar as in [2, 4], we arrive at

$$v_n = -\gamma(\kappa - \overline{\kappa}) - \frac{1}{u^*}f(u) \quad \text{on } \Gamma(t),$$

where $\overline{\kappa}$ is the average curvature along $\Gamma(t)$. Hence, the interface velocity is still driven by both the chemical reaction and by curvature, but where the curvature-driven movement now fulfills conservation of the phase-field parameter. This motion redistributes the mineral towards constant curvature; that is, towards bubbles [13].

4 Conservative Numerical Discretization

We apply a standard finite-volume scheme on an admissible mesh $\mathscr{E}$ [5], and forward or backward Euler in time with constant time step size Δt. For each element $K \in \mathscr{E}$, the discretization for the phase-field equation (4) reads

$$\begin{aligned}\lambda^2|K|\frac{\phi_K^{n+1}-\phi_K^n}{\Delta t} + |K|\gamma P'(\phi_K^\ell) &= \gamma\lambda^2\sum_{L\in\mathscr{N}(K)}|\sigma_{K,L}|F_{K,L}^\ell \\ &\quad -4\lambda|K|\phi_K^\ell(1-\phi_K^\ell)\frac{f(u_K^\ell)}{u^*} + |K|\frac{\gamma}{|\Omega|}\sum_{J\in\mathscr{E}}|J|P'(\phi_J^\ell),\end{aligned} \tag{7}$$

where $|K|$ is the measure of element K. Further, $\mathscr{N}(K)$ refers to the neighboring elements of K and $|\sigma_{K,L}|$ is the measure of the edge $\sigma_{K,L}$ between element K and a neighbor L. The integral $\int_\Omega P'(\phi)d\mathbf{x}$ is approximated by the sum $\sum_{J\in\mathscr{E}} |J|P'(\phi_J^\ell)$. The superscript ℓ is either n or $n+1$ when forward or backward Euler is applied, respectively. The fluxes $F_{K,L}^\ell$ approximate the diffusive flux $\nabla^2\phi$ and are given by

$$F_{K,L}^\ell = \frac{\phi_L^\ell - \phi_K^\ell}{d_{K,L}},$$

where $d_{K,L}$ is the Euclidean distance between the points $x_K \in K$ and $x_L \in L$. Obviously we have $F_{K,L}^\ell = -F_{L,K}^\ell$ on interior edges. As we will apply homogeneous Neumann boundary conditions for ϕ, $F_\sigma \equiv 0$ for edges $\sigma \in \partial\Omega$.

Theorem 1 *The scheme* (7) *is globally conservative (up to the chemical reaction) under homogeneous Neumann boundary conditions on ϕ when the two terms concerning $P'(\phi)$ are either both solved explicitly or both implicitly.*

Proof We sum over all $K \in \mathscr{E}$. Since $F_{K,L}^\ell = -F_{L,K}^\ell$ on internal edges and $F_\sigma = 0$ on boundary edges, the contribution from the diffusive flux vanishes. Hence,

$$\sum_{K\in\mathscr{E}} |K|\phi_K^{n+1} = \sum_{K\in\mathscr{E}} |K|\phi_K^n + \frac{\Delta t\gamma}{\lambda^2}\sum_{K\in\mathscr{E}} |K| \left(\frac{1}{|\Omega|}\sum_{J\in\mathscr{E}} |J|P'(\phi_J^\ell) - P'(\phi_K^\ell)\right)$$
$$-\frac{4\Delta t}{\lambda}\sum_{K\in\mathscr{E}} |K|\phi_K^\ell(1-\phi_K^\ell)\frac{f(u_K^\ell)}{u^*}.$$

Since $\sum_{K\in\mathscr{E}} |K|\frac{1}{|\Omega|}\sum_{J\in\mathscr{E}} |J|P'(\phi_J^\ell) = \sum_{J\in\mathscr{E}} |J|P'(\phi_J^\ell)$ as $\sum_{K\in\mathscr{E}} |K| = |\Omega|$, the two terms concerning $P'(\phi)$ cancel each other when they are evaluated at the same time level t^n or t^{n+1}. Hence

$$\sum_{K\in\mathscr{E}} |K|\phi_K^{n+1} = \sum_{K\in\mathscr{E}} |K|\phi_K^n - \frac{4\Delta t}{\lambda}\sum_{K\in\mathscr{E}} |K|\phi_K^\ell(1-\phi_K^\ell)\frac{f(u_K^\ell)}{u^*},$$

which means that the scheme is globally conservative in case of $f(u) = 0$, and the integrated value of ϕ can only change due to the chemical reactions. □

Remark 1 We here only address the non-linear terms fully explicit and implicit, but mention that also a convex-concave splitting would be discretely conservative when the same elements in the two terms concerning $P'(\phi)$ are chosen as explicit/implicit.

5 Numerical Examples

We consider two numerical examples in 2D: In the first the Allen–Cahn equation is applied to a circular mineral, while in the second example we consider also (1b)–(1d), where flow through a channel with a dissolving mineral layer is addressed.

Equations (1b)–(1d) are discretized with a FV scheme similar as in (7) with u and p at nodes x_K and velocity on a dual mesh for the edge midpoints. The full system is also conservative. The numerical examples use uniform, rectangular meshes. The meshes fulfill $\max\{\Delta x, \Delta y\} < \lambda/4$ to ensure proper resolution of the diffuse interface. For all presented results we use $\gamma = 1$ and $\lambda = 0.1$, but the results are qualitatively the same for other choices. All non-linear systems of equations are solved iteratively using Newton's method. Note that for the implicit phase-field equation, the Jacobian is full since every element depends on all the other elements (c.f. (7)).

5.1 Circular Mineral

The unwanted behavior of the non-conservative Allen–Cahn equation is especially visible for a circular mineral, as the constant curvature of the diffuse interface zone causes the mineral to shrink without the presence of any chemical reaction.

We initialize the square $\Omega = [0, 1]^2$ with a phase field depicting a mineral of radius 0.4 centered in the middle of the square. Homogeneous Neumann conditions are used on all sides. We use the strategy described in [2] to initialize the phase field. No chemical reactions are included.

Figure 1 shows the integrated phase field over time. In the standard Allen–Cahn model the mineral disappears. For the conservative formulation, the changes in porosity are 5.2×10^{-9} and 4.9×10^{-11} for the explicit and implicit formulation, respectively. The changes for the implicit formulation can be connected to the tolerance of the Newton iterations, while the changes for the explicit are mainly an artifact of the explicit time stepping as instabilities evolved.

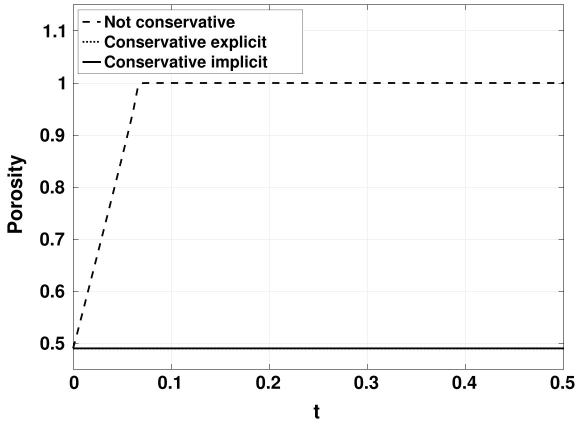


Fig. 1 The integral of the phase field; i.e., porosity, as a function of time. The two lines for conservative implicit and explicit are lying on top of each other

5.2 *Flow Through a Dissolving Channel*

We consider a channel $\Omega = [0, 1] \times [0, 0.1]$ with a prescribed parabolic inflow profile with $v_{\max} = 1$ on the left side and a constant pressure $p = 0$ on the right side. Initially, a mineral layer of width 0.025 is at the top and bottom of the channel, and the fluid is saturated with the equilibrium concentration $u_{\mathrm{eq}} = 0.5$. Due to symmetry we only consider the lower half of the channel. At the left inlet a Dirichlet condition of $u = 0.25$ is applied, triggering mineral dissolution. We consider three cases:

(1) The original model (1), solved fully coupled with backward Euler.
(2) The conservative model (4), (1b)–(1d), solved fully coupled with backward Euler.
(3) The conservative model (4), (1b)–(1d), solved fully coupled with backward Euler except for the two terms concerning $P'(\phi)$, which are both solved explicitly.

The resulting non-linear system of equations is solved with Newton's method. In the second case the Jacobian for the phase-field equation is full. Although the third case gives cheaper Newton iterations, a small time-step size is needed for stability.

In all three cases the mineral dissolves. However, the speed and location of the dissolution vary due to differences in curvature behavior. Figure 2 (left) shows the porosity minus the accumulated reactive term ((5) integrated in time) as function of time for each of the three cases, which should be (close to) zero. Figure 2 (right) shows the across-channel integral of 1-ϕ, giving the mineral width, at $t = 0.3$.

From Fig. 2 (left) it is clear that the non-conservative formulation (1) gradually gives a nonphysical porosity. The conservative implicit formulation (2) has a nonphysical change in porosity of 6.5×10^{-11} throughout the simulation, which can be connected to the tolerance used for the Newton iterations. Despite applying $\Delta t = 10^{-5}$ for the conservative explicit formulation (3), only solutions up to

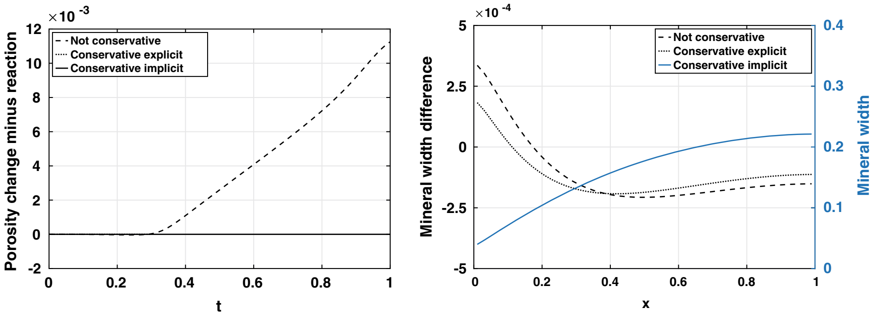


Fig. 2 Left: Changes in porosity not coming from the chemical reaction over time. The dotted line is hidden behind the solid line. Right: Mineral width along the x-axis of case (2) in blue; black lines show difference between case (1) or (3) and (2), at time $t = 0.3$

$t = 0.35$ could be obtained due to instabilities, and a nonphysical change in porosity of 7.9×10^{-8} is observed at this time. The conservative explicit case (3) also shows some difference in curvature behavior compared to case (2) (Fig. 2, right).

6 Discussion and Conclusion

We have formulated a conservative phase-field model for flow and reactive transport with a mineral precipitation and dissolution reaction. The sharp-interface limit shows how the interface evolution still includes curvature-driven motion as well as reaction-driven motion, but where the curvature-driven motion is now conservative.

A standard FV scheme on an admissible mesh ensures the discrete conservation of the phase-field variable (up to the chemical reaction) as long as a consistent explicit/implicit choice is made for the non-linear terms. However, the explicit choice needs a very small time step to avoid numerical instabilities. For the implicit choice the Jacobian is full, giving expensive Newton iterations. It would be beneficial to rather use an iterative scheme like the L-scheme [6] for the non-linear solving steps. A concave-convex splitting of the non-linearity could also be beneficial.

Finally, we note that the reformulated phase-field model can be upscaled similarly as in [2], hence giving a two-scale (pore-Darcy) model where the phase field is updated locally. The conservation property is then achieved for each local pore.

Acknowledgements The author would like to thank Florian Frank (FAU Erlangen-Nürnberg) for useful discussions on the phase-field formulation. The computational resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Research Foundation—Flanders (FWO) and the Flemish Government—department EWI. Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—Project Number 327154368—SFB 1313.

References

1. Allen, S.M., Cahn, J.W.: A microscopic theory for antiphase boundary motion and its application to antiphase domain coarsening. Acta Metall. **27**(6), 1085–1095 (1979). https://doi.org/10.1016/0001-6160(79)90196-2
2. Bringedal, C., von Wolff, L., Pop, I.S.: Phase field modeling of precipitation and dissolution processes in porous media: Upscaling and numerical experiments. Multiscale Model. Simul. http://www.uhasselt.be/Documents/CMAT/Preprints/2019/UP1901.pdf (2019)
3. Caginalp, G., Fife, P.: Dynamics of layered interfaces arising from phase boundaries. SIAM J. Appl. Math. **48**(3), 506–518 (1988). https://doi.org/10.1137/0148029
4. Chen, X., Hilhorst, D., Logak, E.: Mass conserving Allen-Cahn equation and volume preserving mean curvature flow. Interface Free. Bound **12**, 527–549 (2010). https://doi.org/10.4171/IFB/244
5. Eymard, R., Gallouët, T., Herbin, R.: Finite volume method, vol. 7. Elsevier, Amsterdam (2000)
6. List, F., Radu, F.A.: A study on iterative methods for solving Richards' equation. Comput. Geosci. **20**(2), 341–353 (2016). https://doi.org/10.1007/s10596-016-9566-3

7. Mu, X., Frank, F., Riviere, B., Alpak, F.O., Chapman, W.G.: Mass-conserved density gradient theory model for nucleation process. Ind. Eng. Chem. Res. **57**(48), 16476–16485 (2018). https://doi.org/10.1021/acs.iecr.8b03389
8. van Noorden, T.L.: Crystal precipitation and dissolution in a porous medium: effective equations and numerical experiments. Multiscale Model. Simul. **7**(3), 1220–1236 (2009)
9. van Noorden, T.L., Eck, C.: Phase field approximation of a kinetic moving-boundary problem modelling dissolution and precipitation. Interface Free. Bound **13**(1), 29–55 (2011). https://doi.org/10.4171/IFB/247
10. van Noorden, T.L., Pop, I.S.: A Stefan problem modelling crystal dissolution and precipitation. IMA J. Appl. Math. **73**(2), 393–411 (2008). https://doi.org/10.1093/imamat/hxm060
11. Redeker, M., Rohde, C., Pop, I.S.: Upscaling of a tri-phase phase-field model for precipitation in porous media. IMA J. Appl. Math. **81**(5), 898–939 (2016). https://doi.org/10.1093/imamat/hxw023
12. Rubinstein, J., Sternberg, P.: Nonlocal reaction-diffusion equations and nucleation. IMA J. Appl. Math. **48**(3), 249–264 (1992). https://doi.org/10.1093/imamat/48.3.249
13. Schlögl, F.: Chemical reaction models for non-equilibrium phase transitions. Zeitschrift für Physik **253**(2), 147–161 (1972). https://doi.org/10.1007/BF01379769

A Fully Conforming Finite Volume Approach to Two-Phase Flow in Fractured Porous Media

Samuel Burbulla and Christian Rohde

Abstract In many natural and technical applications in porous media fluid's flow behavior is highly affected by fractures. Many approaches employ mixed-dimensional models that model thin features as dimension-reduced manifolds. Following this idea, we consider porous media where dominant heterogeneities are geometrically represented by sharp interfaces. We model incompressible two-phase flow in porous media both in the bulk porous medium and within the fractures. We present a reliable and geometrically flexible implementation of a fully conforming finite volume approach within the DUNE framework for two and three spatial dimensions. The implementation is based on the new `dune-mmesh` grid implementation that manages bulk and surface triangulation simultaneously. The model and the implementation are extended to handle fracture junctions. We apply our scheme to benchmark cases with complex fracture networks to show the reliability of the approach.

Keywords Flows in porous media · Two-phase flows · Finite volume methods

MSC (2010) 76S05 · 76T99 · 74S10

1 Introduction

The numerical simulation of flow in fractured porous media is a challenging task. Hence, many approaches use discrete fracture-matrix models where fractures are represented by lower-dimensional interfaces. However, the representation of lower-dimensional computational grids that are coupled with a surrounding bulk grid raises a lot of geometrical difficulties. For instance, in conforming approaches, the lower-dimensional grid has to coincide with facets of the bulk mesh.

S. Burbulla (✉) · C. Rohde
Institute of Applied Analysis and Numerical Simulation, University of Stuttgart,
Pfaffenwaldring 57, D-70569 Stuttgart, Germany
e-mail: samuel.burbulla@mathematik.uni-stuttgart.de

R. Klöfkorn et al. (eds.), *Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples*, Springer Proceedings in Mathematics & Statistics 323,
https://doi.org/10.1007/978-3-030-43651-3_51

We follow a conforming approach, because the implementation of the coupling is restricted only to the facets that coincide with the lower-dimensional grid. In order to provide a tool for conforming discretizations in a general framework, we introduce the new DUNE grid implementation `dune-mmesh`. It is a grid manager for triangulations in 2D and 3D and can export a predescribed set of facets as a separate network grid. We have implemented a model for incompressible two-phase flow in fractured porous media on the basis of a conforming discretization using `dune-mmesh` and give some details about this method in the following.

2 Governing Equations

We consider convection-dominated two-phase flow in porous media in the so-called fractional flow formulation, see Eq. (1) below, [12]. This formulation of the two-phase flow model leads to a nonlinear mixed hyperbolic-elliptic system of equations. It allows to exploit numerical stabilization techniques as developed for nonlinear conservation laws.

Fractional Flow Formulation for Two-Phase Flow in Porous Media

Let $\Omega \in \mathbb{R}^d$ be an open and bounded domain and $t_{end} > 0$. Neglecting capillary pressure, two-phase flow in porous media is governed for the unknowns (wetting phase) saturation $S : \Omega \times (0, t_{end}) \to [0, 1]$, total velocity $v : \Omega \times (0, t_{end}) \to \mathbb{R}^d$ and global pressure $P : \Omega \times (0, t_{end}) \to \mathbb{R}$ by

$$\left.\begin{aligned} (\phi S)_t + \operatorname{div} F(S, v) &= q_w, \\ v + \lambda(S)\mathbf{K}(\nabla P - G(S)g) &= 0, \\ \operatorname{div}(v) &= q_w + q_n, \end{aligned}\right\} \quad \text{in } \Omega \times (0, t_{end}). \tag{1}$$

The flux function is defined by

$$F(S, v) := f(S)v - f(S)\lambda_n(S)\mathbf{K}(\rho_n - \rho_w)g.$$

The gravity term in front of the acceleration vector $g \in \mathbb{R}^d$ is given by $G(S) := (\lambda_w(S)\rho_w + \lambda_n(S)\rho_n)/\lambda(S)$, $f(S) := \lambda_w(S)/\lambda(S)$ is the fractional flow function and $\lambda(S) := \lambda_w(S) + \lambda_n(S)$ is the total mobility. The phase mobility function is $\lambda_\alpha(S_\alpha) = k_\alpha(S_\alpha)/\mu_\alpha$, where $k_\alpha(S_\alpha)$ is the relative permeability of phase $\alpha \in \{w, n\}$. Further physical parameters are the porosity $\phi = \phi(x) \in (0, 1]$, the constant phase densities $\rho_\alpha > 0$, the dynamic viscosities μ_α and the symmetric and positive definite intrinsic permeability tensor $\mathbf{K} = \mathbf{K}(x) \in \mathbb{R}^{d\times d}$. The function $q_\alpha : \Omega \times (0, t_{end}) \to \mathbb{R}$ is a source or sink term. Appropriate initial and boundary conditions have to be added.

Fractional Flow Formulation in a Fractured Porous Medium

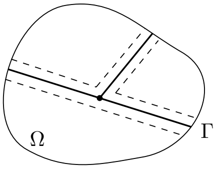
If the fractures's apertures are small compared to the overall size of Ω, it is justified to model them as lower dimensional manifolds to reduce the computational effort for representing them [9]. A mathematical model to describe two-phase flow in fractured porous media on the basis of a discrete fracture network approach can be formulated as follows [8]. Let $t_{end} > 0$, $\Omega \subset \mathbb{R}^d$ open and bounded and $\Gamma \subset \Omega$ open with $\mathcal{H}^d(\Gamma) = 0$, $\mathcal{H}^{d-1}(\Gamma) > 0$, where $\mathcal{H}^d$ is the d-dimensional Hausdorff measure. We look for S, v, P and S_Γ, v_Γ, P_Γ such that S, v, P satisfy (1) in $(\Omega \setminus \Gamma)^\circ$ and S_Γ, v_Γ, P_Γ satisfy

$$\left.\begin{aligned}(\phi^f \omega S_\Gamma)_t + \operatorname{div}_\tau F_\Gamma(S_\Gamma, v_\Gamma) &= [\![F(S, v) \cdot n]\!] + q_w^\Gamma,\\ v_\Gamma + \lambda^f(S_\Gamma)\mathbf{K}_\tau^f \omega(\nabla_\tau P_\Gamma - G^f(S_\Gamma) g_\tau) &= 0,\\ \operatorname{div}_\tau(v_\Gamma) &= [\![v \cdot n]\!] + q_w^\Gamma + q_n^\Gamma,\end{aligned}\right\} \quad \text{in } \Gamma \times (0, t_{end}), \tag{2}$$

where $F_\Gamma(S_\Gamma, v_\Gamma) := f^f(S_\Gamma)v_\Gamma - f^f(S_\Gamma)\lambda_n^f(S_\Gamma)\mathbf{K}_\tau^f \omega(\rho_n - \rho_w)g$. The superscript ^{f} indicates the physical parameters defined on Γ, the subscript $_\tau$ the tangential projection onto Γ, n denotes a normal to the fracture and $\omega : \Gamma \to \mathbb{R}^{>0}$ defines the aperture of the fracture at a given position. The systems (1), (2) are closed by the transmission conditions

$$\left.\begin{aligned}F(S, v) \cdot n &= F^f(S_\Gamma, v) \cdot n,\\ \eta\{\!\{v \cdot n\}\!\} &= [\![P]\!] + \omega G^f(S_\Gamma)(g \cdot n),\\ \{\!\{P\}\!\} - P_\Gamma &= \frac{\eta}{12}[\![v \cdot n]\!]\end{aligned}\right\} \quad \text{at } \Gamma \tag{3}$$

with $\eta := \frac{\omega}{\lambda^f(S_\Gamma)}(\mathbf{K}_n^f)^{-1}$ and appropriate initial data and boundary conditions. For $x \in \Gamma$ we define $[\![\phi]\!] := \lim_{\varepsilon\downarrow 0} \phi(x + \varepsilon n) - \lim_{\varepsilon\downarrow 0} \phi(x - \varepsilon n)$ and $\{\!\{\phi\}\!\} := \frac{1}{2}\lim_{\varepsilon\downarrow 0}$ $\phi(x + \varepsilon n) + \frac{1}{2}\lim_{\varepsilon\downarrow 0} \phi(x - \varepsilon n)$.



At junctions of the fracture network we suppose

$$\sum_{k=1}^{n} F_{\Gamma_k}(S_{\Gamma_k}, v_{\Gamma_k}) \cdot n_{\Gamma_k} = 0, \tag{4}$$

$$\sum_{k=1}^{n} v_{\Gamma_k} \cdot n_{\Gamma_k} = 0, \tag{5}$$

$$P_{\Gamma_i} = P_{\Gamma_j} \quad \text{for } 1 \le i < j \le n, \tag{6}$$

where $\Gamma_k,\ 1 \leq k \leq n$, are the fracture angles at the junction. The assumption of the continuity of phase pressure is reasonable if all fractures have the same high permeability compared to the bulk medium.

3 Discretization

We discretize the mixed-dimensional equations in (1)–(6) by a fully conforming finite volume approach. Thereby, we are able to ensure basic properties like mass conservation and we can take care of the hyperbolicity in the Eqs. (1a), (2a). The conforming discretization allows to not only model open fractures, but also barriers for the fluid flow that have a lower permeability than the bulk medium. The numerical method is extended to handle intersections of multiple fractures.

Elliptic Equations

In order to discretize the elliptic part of the model we plug together pressure and velocity equations. The finite volume approach for the divergence constraint (1c) reads

$$\int_T q_w + q_n = \sum_{e \in \mathcal{E}(T)} |e| v_e \cdot n_e^T$$

where v_e is a suitable approximation of v on e. A common choice for v_e is the two point flux approximation (TPFA) [4] defined by Eq. (1b). We extend the standard formulation of the normal velocity by the gravity terms. Then, it reads

$$v_e \cdot n_e^T := -\mathbb{T}_{T,T'}\big((P_{T'} - P_T) - \mathbb{G}_{T,T'} \cdot g\big).$$

Here, the transmissibility $\mathbb{T}_{T,T'}$ is given by

$$\mathbb{T}_{T,T'} := \frac{\mathbb{T}_T \mathbb{T}_{T'}}{\mathbb{T}_T + \mathbb{T}_{T'}} \qquad \text{with} \qquad \mathbb{T}_i := \frac{n_e^i \lambda(S_i) \mathbf{K}_i d_i}{\|d_i\|^2}$$

where $d_i := m_e - m_i$ is the distance vector between the center of the facet e and the cell centers, $i \in \{T, T'\}$, and S_T denotes the saturation value in cell T. The gravitational influence $\mathbb{G}_{T,T'}$ is

$$\mathbb{G}_{T,T'} := \mathbb{G}_T - \mathbb{G}_{T'}, \qquad \mathbb{G}_T := d_T G(S_T).$$

This TPFA-based discretization is consistent for isotropic intrinsic permeabilities if we locate the pressure values P_T at the circumcenters of the tetrahedral cells [4]. This is still valid for the coupling to the fracture network as the circumcenters of the lower-dimensional grid elements are located at the orthogonal connection line of the circumcenters of the two adjacent bulk cells.

At facets that coincide with a lower-dimensional fracture element T_Γ, we include the coupling conditions of the reduced model. Therefore, we introduce intermediate pressure values $P_1|_{\gamma_1}$ and $P_2|_{\gamma_2}$ at the boundaries of the bulk medium next to the fracture. A condition for these values can be stated by

$$v_i \cdot n_i|_{\gamma_i} = -\mathbb{T}_i\big((P_i|_{\gamma_i} - P_i) - \mathbb{G}_i \cdot g\big), \qquad i = 1, 2.$$

The coupling conditions for velocity and pressure in (3) can be used to eliminate the intermediate values. Defining $\mathbb{G}_\Gamma := \frac{\omega}{2} G^f(S_\Gamma)n$ and the fracture transmissibility $\alpha_f := 2/\eta = 2\lambda^f(S_\Gamma)\mathbf{K}_n^f/\omega$ we obtain

$$\begin{aligned} v_1 \cdot n_1|_{\gamma_1} = &\frac{\mathbb{T}_1\alpha_f}{\mathbb{T}_1\mathbb{T}_2 + \alpha_f(2(\mathbb{T}_1 + \mathbb{T}_2) + 3\alpha_f)} \\ &\times \left[\begin{pmatrix} 2\mathbb{T}_2 + 3\alpha_f \\ \mathbb{T}_2 \\ -3(\mathbb{T}_2 + \alpha_f) \end{pmatrix} \cdot \begin{pmatrix} P_1 \\ P_2 \\ P_\Gamma \end{pmatrix} + \begin{pmatrix} \mathbb{T}_2 + 3\alpha_f \\ 2\mathbb{T}_2 \\ \mathbb{T}_2 + 3\alpha_f \end{pmatrix} \cdot \begin{pmatrix} \mathbb{G}_1 \cdot g \\ \mathbb{G}_2 \cdot g \\ \mathbb{G}_\Gamma \cdot g \end{pmatrix} \right]. \end{aligned}$$

Hyperbolic Saturation Equations

In order to solve the hyperbolic saturation equation appropriately, we employ a finite volume scheme with a suitable numerical flux.

For an initial value problem in conservation like (1a) the corresponding (implicit) finite volume scheme can be written as

$$\begin{aligned} S_T^{(n+1)} &= S_T^{(n)} - \frac{\Delta t_n}{|T|} \sum_{e\in\mathcal{E}(T)} |e| g_e(S_T^{(n+1)}, S_{T'}^{(n+1)}) + \int_T q_w, \qquad (n \geq 0), \\ S_T^{(0)} &= \frac{1}{|T|} \int_T S_0, \end{aligned}$$

where $g_e(\cdot,\cdot)$ is a numerical flux that is consistent with the flux function at the corresponding intersection e, i.e. $g_e(S, S) = F(S, v_e) \cdot n_e$. Because the flux function F is non-convex in the first argument, a simple upwinding is not sufficient. We choose the Godunov flux that results from an exact solution of the Riemann problem.

Generalization of the Scheme to Networks

The discretizations presented so far are applicable to the bulk problem and the surface fracture problem. As we consider networks of fractures we generalize the finite volume scheme for the network situation. Therefore, we use the assumptions we made in (4)–(6).

When we consider two subdomains with different rock types, e.g. with different permeability, the flux function becomes discontinuous at the interface. We define the flux as the unique solution $\bar{g}(S_l, S_r)$ satisfying

$$\bar{g}(S_l, S_r) := g^l(S_l, S^*) = g^r(S^*, S_r) \tag{7}$$

for some intermediate value S^* [11]. Here, $g^\circ(S, S) = F_\circ(S, v_e) \cdot n_e$ for $\circ = l, r$. In case of a junction we propose a new idea analogously to the idea in (7). We introduce an intermediate value S^* and fix its value by the mass conservation condition in (4) on a discrete level by

$$\sum_{i=1}^{k} g_i(S_{T_i}, S^*) = 0,$$

where $g_i(S, S) = F_i(S, v_i) \cdot n_e^{T_i}$. For this definition, we can show the existence of $S^* \in [0, 1]$ and the uniqueness of the resulting fluxes.

The pressure continuity (6) allows to assume that there is a unique pressure value P^* at the intersection. We require that the discrete normal velocity $v_k \cdot n_e^k$ satisfies

$$v_k \cdot n_e^k = -\mathbb{T}_k\big((P^* - P_k) - \mathbb{G}_k \cdot g\big), \qquad k = 1, \dots, n,$$

and using the second equation of (5) we obtain

$$P^* = \frac{\sum_{k=1}^{n} \mathbb{T}_k(P_k + \mathbb{G}_k \cdot g)}{\sum_{k=1}^{n} \mathbb{T}_k}.$$

4 Implementation

We implemented our method from Sect. 3 within the software framework DUNE on the basis of the discretization module DuMu$^\text{x}$ [5]. The implementation applies to 2D and 3D domains. All equations are solved monolithically by a fully implicit time discretization.

The New DUNE Grid Implementation `dune-mmesh`

We have initiated the implementation of an own DUNE [1] grid module which is tailored for mixed-dimensional, conforming discretizations. Thus, we present the new DUNE grid implementation `dune-mmesh` [3] which is a wrapper of CGAL [13] triangulations acting as a DUNE grid. It is implemented for spatial dimension $d = 2, 3$ and gives access to all the capabilities of the underlying CGAL triangulations. The grid implementation is extended by an interface grid implementation that exports a predescribed set of facets as a separate DUNE grid. With this approach we can use `dune-mmesh` as computational grid for both the bulk and the surface problem. The interface grid is directly integrated in the CGAL grid wrapper implementation and provides access to all neighbor relationships easily. One of the main advantages of the strong coupling of the two grids is the simultaneous remeshing of the bulk and the interface grid, e.g. during adaptation.

`dune-mmesh` is the first DUNE grid implementation that is capable of exporting a predescribed set of facets as a separate surface grid. Furthermore, with our interface grid implementation, we add a new network grid implementation to the DUNE framework where cells can have multiple neighbors at the same facet. The implementation will be further improved, but we can already imagine `dune-mmesh` to be used in several applications where mixed-dimensional problems are solved on the basis of various conforming discretizations. A first open-source release of `dune-mmesh` is available implementing the grid wrapper for CGAL triangulations and the interface grid [3].

Numerical Experiments

We apply our implementation to complex networks showing the reliability of the approach. We performed various test cases in two and three space dimensions for several fracture configurations.

Benchmark Test From [6]

To validate the numerical solver we use test case 5.1 from [6] applied to a 3D domain. Consider a rectangular domain $\Omega = (0, 1) \times (-4, 4) \times (0, 1)$ that is cut by a diagonal fracture $\Gamma = \{x = -0.1y + 0.5\} \cap \Omega$ with aperture $\omega = 0.01$. We choose $\mathbf{K} = \mathbf{I}$, $S_0 = 0.8$ for $y < 0$ and $\mathbf{K} = 2\mathbf{I}$, $S_0 = 0.1464$ for $y > 0$ both in the bulk and the fracture. Furthermore, we set $\phi = 1$ everywhere, $k_n(S) = S$, $k_w(S) = 1 - S$, $\rho_w = 2$, $\rho_n = 1$, $\mu_w = \mu_n = 1$, $g = (0, -1, 0)^T$ and $v \equiv 0$. The solution shows the expected rarefaction wave between $S = 0.8$ to $S = 0.5$ and a steady discontinuity at $y = 0.5$ (Fig. 1).

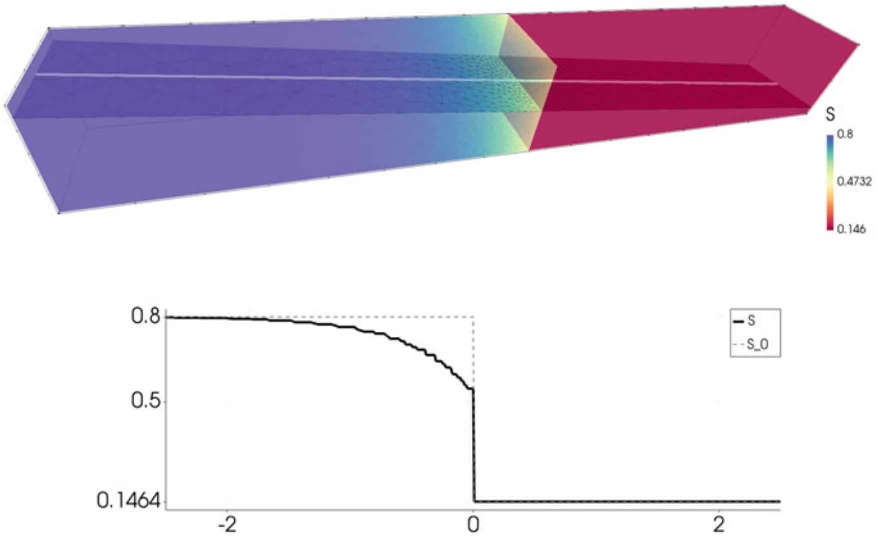


Fig. 1 We use a Riemann problem as benchmark with a diagonal fracture cutting the domain. The solution at $t = 1.5$ shows the expected rarefaction wave and the steady discontinuity. The plot over line shows the saturation along the fracture's centerline

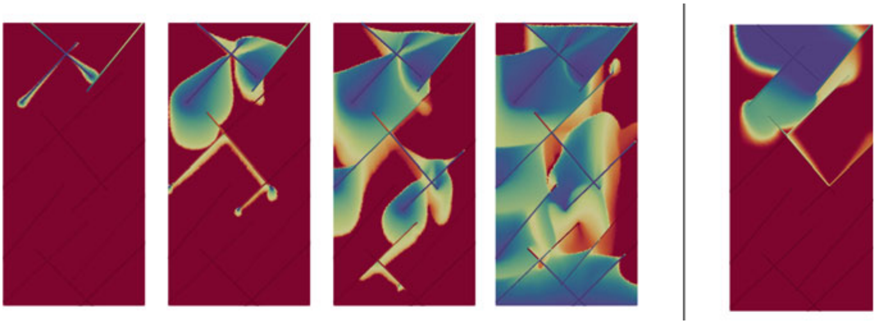


Fig. 2 Gravity-induced wetting of a fracture network for $t \in \{10, 320, 800, 2220\}$ s. In comparison, on the right-hand side, the same setup with Richards equation for $t = 2220$ s

Realistic Test Case with Complex Fracture Network

In Fig. 2 we consider the wetting of a fracture network with the dynamics driven by the density difference of the two phases for realistic physical parameters. Eleven fractures with apertures between 1.46×10^{-6} m and 1.61×10^{-6} m are placed in a cubic domain of size 1.5 cm × 3 cm. We choose $\rho_w = 1000\,\mathrm{kgm}^{-3}$, $\rho_n = 100\,\mathrm{kgm}^{-3}$, $\mu_w = \mu_n = 1 \times 10^{-3}\,\mathrm{kgm}^{-1}\,\mathrm{s}^{-1}, \mathbf{K} = 1 \times 10^{-13}\,\mathrm{m}^2, \mathbf{K}_f = 5 \times 10^{-9}\mathrm{m}^2, \phi = 0.2, \phi_f = 0.8$ and a quadratic relative permeability law. We set no-flow boundary conditions everywhere except at the two fracture endings at the top where we prescribe a Dirichlet boundary value $(S, P) = (1, 100\,\mathrm{Pa})$ at the left and $(S, P) = (0, 0\,\mathrm{Pa})$ at the right tip. For comparison, we add the result for the same setup with a linear relative permeability law and $\rho_n = \rho_w$ where the system collapses to the Richards equation.

5 Outlook

We plan to integrate the expansion of fractures by coupling the scheme with the moving mesh concept developed in [2]. The movement of the fracture tip will be obtained by the integration of a phase-field model on the microscale locally around the fracture tips [7, 10].

Acknowledgements Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—Project Number 327154368—SFB 1313.

References

1. Blatt, M., Burchardt, A., Dedner, A., Engwer, C., Fahlke, J., Flemisch, B., Gersbacher, C., Gräser, C., Gruber, F., Grüninger, C., Kempf, D., Klöfkorn, R., Malkmus, T., Müthing, S., Nolte, M., Piatkowski, M., Sander, O.: The distributed and unified numerics environment, version 2.4. Arch. Numer. Softw. **4**, 13–29 (2016)
2. Chalons, C., Rohde, C., Wiebe, M.: A finite volume method for undercompressive shock waves in two space dimensions. ESAIM Math. Mod. Num. Anal. **51**, 1987–2015 (2017)
3. Dune-MMesh: A DUNE grid implementation based on CGAL Delaunay triangulations. Release 1.1. https://gitlab.dune-project.org/samuel.burbulla/dune-mmesh
4. Eymard, R., Gallouët, T., Herbin, R.: Finite volume methods. Techniques of Scientific Computing, Part III, Handbook of Numerical Analysis VII (2000)
5. Flemisch, B., Darcis, M., Erbertseder, K., Faigle, B., Lauser, A., Mosthaf, K., Müthing, S., Nuske, P., Tatomir, A., Wolff, M., Helmig, R.: DuMux: DUNE for multi-phase, component, scale, physics, ... flow and transport in porous media. Adv. Water Resour. **34**, 1102–1112 (2011)
6. Fumagalli, A., Scotti, A.: A numerical method for two-phase flow in fractured porous media with non-matching grids. Adv. Water Resour. **62**, 454–464 (2013)
7. Giovanardi B., Scotti A., Formaggia L.: A hybrid XFEM–Phase field (Xfield) method for crack propagation in brittle elastic materials. Comput. Methods Appl. Mech. Eng. **320** (2017)
8. Jaffré, J., Mnejja, M., Roberts, J.E.: A discrete fracture model for two-phase flow with matrix-fracture interaction. Procedia Comput. Sci. **4** (2011)
9. Martin, V., Jaffré, J., Roberts, J.E.: Modeling fractures and barriers as interfaces for flow in porous media. SIAM J. Sci. Comput. **26**, 1667–1691 (2005)
10. Miehe, C., Mauthe, S., Teichtmeister, S.: Minimization principles for the coupled problem of Darcy-Biot-type fluid transport in porous media linked to phase field modeling of fracture. J. Mech. Phys. Solids **82**, 186–217 (2015)
11. Mishra, S., Jaffré, J.: On the upstream mobility scheme for two-phase flow in porous media. Comput. Geosci. **14**, 105–124 (2010)
12. Peaceman, D.W.: Fundamentals of Numerical Reservoir Simulation. Elsevier, Amsterdam (1977)
13. The CGAL Project: CGAL user and reference manual. Version 4.14. https://doc.cgal.org/4.14/Manual/packages.html (2019)

Monotone Embedded Discrete Fracture Method for the Two-Phase Flow Model

Kirill D. Nikitin and Ruslan M. Yanbarisov

Abstract We propose an application of the new monotone embedded discrete fracture method (mEDFM) [13] to the two-phase flow model. The new method for modelling of flows in fractured media consists in coupling of the embedded discrete fracture method (EDFM) with the nonlinear monotone finite volume (FV) scheme with two-point flux approximation, which preserves non-negativity of the discrete solution. The resulting method combines effectiveness and simplicity of the standard EDFM approach with accuracy and physical relevance of the nonlinear FV schemes for non-orthogonal grids and anisotropic media. Numerical experiments show that the two-phase flow modelling with the mEDFM provides much more accurate solution compared to the conventional EDFM, and is in a good agreement with the discrete fracture method, which directly applies the nonlinear FV method to a grid with fractures explicitly represented by 3D cells.

Keywords Finite volume method · Nonlinear discretization scheme · Fracture modelling · Embedded discrete fracture model · Flows in porous media · Two-phase flows

MSC (2010) 76S05 · 76M12 · 76T99

K. D. Nikitin (✉) · R. M. Yanbarisov
Marchuk Institute of Numerical Mathematics, Russian Academy of Sciences, Gubkina st. 8, Moscow, Russia
e-mail: nikitin.kira@gmail.com

R. M. Yanbarisov
e-mail: ruslan.yanbarisov@gmail.com

R. Klöfkorn et al. (eds.), *Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples*, Springer Proceedings in Mathematics & Statistics 323,
https://doi.org/10.1007/978-3-030-43651-3_52

1 Introduction

A significant amount of world's hydrocarbon reserves lies in reservoirs with fractures of various length scales.

One of popular methods of accounting fractures is the embedded discrete fractures method (EDFM). The method was first proposed in [7] as a hierarchical approach to modelling fractures in porous media. Small fractures were accounted implicitly by their effective properties, while large fractures were considered explicitly. This method can be coupled with any approach such as the dual-porosity dual-permeability method and others. The idea of representing large-scale fractures by embedded grids independent of the reservoir grids was presented in [8]. The family of EDFM methods was further developed in [4, 5, 10].

In EDFM fractures are considered as surfaces with prescribed apertures, and the connecting term between fractures and surrounding rock matrix can be derived by dimensionality reduction [6].

The original EDFM was proposed for the structured grid and isotropic media, thus the conventional linear two-point flux approximation (TPFA) scheme was used for all discrete fluxes. However, it is well known that the linear TPFA lacks approximation on non-$\mathbb{K}$-orthogonal grids. One popular alternative to the linear TPFA is the linear multi-point flux approximation [1], which is second-order accurate, but may be non-monotone for the cases with anisotropic media, which often coexists with fractures.

In our previous work [13] we proposed the monotone embedded discrete fractures method (mEDFM) which couples the original EDFM approach with two advanced nonlinear schemes: the monotone two-point flux approximation (NTPFA) [3] and the compact multi-point flux approximation (NMPFA) satisfying the discrete maximum principle (DMP) [2, 9]. The importance of the monotone and DMP schemes for the multi-phase flow models was studied in [11].

In this paper we consider the application of the mEDFM to the two-phase flows in porous media and compare the results with the original EDFM and with the discrete fracture method, which assumes explicit representation of fractures by the computational grid and uses the similar nonlinear scheme for the flux discretization.

2 Two-Phase Flow Model

The basic equations for the two-phase flow in a domain $\Omega \subset \mathbb{R}^3$ are the following:

1. Mass conservation for each phase:

$$\frac{\partial \rho_\alpha \varphi S_\alpha}{\partial t} + \operatorname{div}\left(\rho_\alpha \mathbf{u}_\alpha\right) = q_\alpha, \quad \alpha = w, o. \tag{1}$$

2. Darcy's law:

$$\mathbf{u}_\alpha = -\lambda_\alpha \mathbb{K}\left(\nabla p_\alpha - \rho_\alpha g \nabla z\right), \quad \alpha = w, o. \tag{2}$$

3. Two fluids fill the voids:
$$S_w + S_o = 1. \tag{3}$$

4. Pressure difference between phases is given by the capillary pressure $p_c = p_c(S_w)$:
$$p_o - p_w = p_c. \tag{4}$$

Here $\mathbb{K}$ is the absolute permeability tensor, $\varphi(p)$ is the porosity, g is the gravity term, z is the depth. For the phase α we have denoted: the pressure p_α (unknown), the saturation S_α (unknown), the Darcy's velocity $\mathbf{u}_\alpha$ (unknown), the density $\rho_\alpha(p) = \rho_{\alpha,0}/B_\alpha(p)$, the formation volume factor $B_\alpha(p)$, the mobility $\lambda_\alpha(p, S) = k_{r\alpha}(S)/\mu_\alpha(p)$, the relative permeability $k_{r\alpha}(S)$, the viscosity $\mu_\alpha(p)$, and the source/sink well term q_α (e.g. the injector or producer wells).

For the boundaries we consider no-flow condition, and for the wells the simple Peaceman formula is used [14]. For a cell T with center $\mathbf{x}_T$ connected to the well we have:
$$q_\alpha(\mathbf{x}) = \frac{\rho_\alpha k_{r\alpha}}{\mu_\alpha} WI\Big(p_{bh} - p - \rho_\alpha g(z_{bh} - z)\Big)\delta(\mathbf{x} - \mathbf{x}_T), \tag{5}$$

where p_{bh} is the bottom hole pressure, z_{bh} is the depth of the bottom hole, WI is the well index, which does not depend on the properties of fluids, but depends on the properties of the media, δ is the Dirac delta function.

3 Embedded Discrete Fracture Method

For representation of the fractured reservoir we use two types of media: the matrix domain $\Omega^m \subset \mathbb{R}^3$ and the fractures domain $\Omega^f \subset \mathbb{R}^3$ represented by n_f virtual domains $\Omega^f = \bigcup\limits_{i=1}^{n_f} \Omega^{f,i}$.

Each fracture $\Omega^{f,i}$ is considered as the surface extruded on the fracture aperture $w_{f,i}$. We assume that the fractures permeability and porosity are significantly larger than that of the porous media.

Next we define mass balance equation (1) for each of the domains Ω^m, Ω^f [5]:

$$\frac{\partial \rho_\alpha \varphi^m S_\alpha^m}{\partial t} + \operatorname{div}\left(\rho_\alpha \mathbf{u}_\alpha^{mm}\right) + \operatorname{div}\left(\rho_\alpha \mathbf{u}_\alpha^{mf}\right) = q_\alpha^m, \quad \text{in } \Omega^m,\ \alpha = w, o, \tag{6}$$
$$\frac{\partial \rho_\alpha \varphi^f S_\alpha^f}{\partial t} + \operatorname{div}\left(\rho_\alpha \mathbf{u}_\alpha^{fm}\right) + \operatorname{div}\left(\rho_\alpha \mathbf{u}_\alpha^{ff}\right) = q_\alpha^f, \quad \text{in } \Omega^f,\ \alpha = w, o, \tag{7}$$

where $\mathbf{u}_\alpha^{mm}$ is the cell-to-cell Darcy's flux identical to (2) for pressure p^m and saturation S^m defined in the matrix, $\mathbf{u}_\alpha^{ff}$ is the similar flux for the fractures domain for

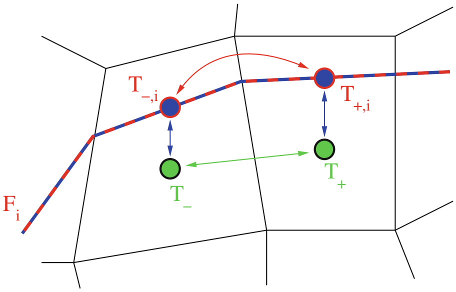


Fig. 1 Darcy fluxes for a fracture in porous media: cell-to-cell (green), cell-to-fracture (blue) and intra-fracture (red) exchanges

unknowns p^f and S^f defined in fractures, and $\mathbf{u}_\alpha^{mf} = -\mathbf{u}_\alpha^{fm}$ is the additional flow between the matrix and the fractures.

On the discrete level, for each grid cell T there is a set of the matrix unknowns $p^m_{\alpha,T}$, $S^m_{\alpha,T}$, and $n_{f,T}$ fracture unknowns p^f_{α,f_i}, S^f_{α,f_i}, where $n_{f,T}$ is the number of fractures F_i inside the cell.

The fully implicit scheme is used for the solution of the coupled equations. For the spatial discretization we use the finite volume method, however instead of one flux we need to approximate three types of fluxes $\mathbf{u}^*$ in Eqs. (6) and (7), which are schematically presented in Fig. 1. We use the following space discretizations for the fluxes (colors correspond to the ones in the figure):

- For the **cell-to-cell** flux between cells T_+ and T_- we use the nonlinear TPFA scheme [3] both for the pressure (including capillary) and the gravity terms:

$$\begin{aligned}\operatorname{div}\left(\rho_\alpha \mathbf{u}_\alpha^{mm}\right) \approx \operatorname{upw}&\Big[\rho_\alpha^{n+1}(p^m)\lambda_\alpha^{n+1}(S^m, p^m),\ \left(M_+(p^m)p^m_+ - M_-(p^m)p^m_-\right)\Big]\\ - \operatorname{upw}&\Big[\rho_\alpha^{n+1}(p^m)\lambda_\alpha^{n+1}(S^m, p^m),\ \left(M_+(z)\rho_+ g z_+ - M_-(z)\rho_- g z_-\right)\Big].\end{aligned}$$

- For the **fracture-to-cell** flux within cell T we use the conventional linear TPFA:

$$\begin{aligned}\operatorname{div}\left(\rho_\alpha \mathbf{u}_\alpha^{mf}\right) \approx \operatorname{upw}_{mf}&\Big[\rho_\alpha^{n+1}(p)\lambda_\alpha^{n+1}(S, p),\ M_T^{mf}\left(p^m - p^f\right)\Big]\\ - \operatorname{upw}_{mf}&\Big[\rho_\alpha^{n+1}(p)\lambda_\alpha^{n+1}(S, p),\ M_T^{mf}\left(\rho_m g z_m - \rho_f g z_f\right)\Big].\end{aligned}$$

- For the **intra-fracture** flux between virtual fracture cells $T_{-,i}$ and $T_{+,i}$ (intersection of the fracture F_i with cells T_- and T_+, respectively) we also use the linear TPFA:

$$\begin{aligned}\operatorname{div}\left(\rho_\alpha \mathbf{u}_\alpha^{ff}\right) \approx \operatorname{upw}&\Big[\rho_\alpha^{n+1}(p^f)\lambda_\alpha^{n+1}(S^f, p^f),\ M^{ff}\left(p^f_+ - p^f_-\right)\Big]\\ - \operatorname{upw}&\Big[\rho_\alpha^{n+1}(p^f)\lambda_\alpha^{n+1}(S^f, p^f),\ M^{ff}\left(\rho_+ g z_+ - \rho_- g z_-\right)\Big].\end{aligned}$$

Here, $M_\pm(p)$ are the coefficients of the nonlinear discretization scheme, M^* are the EDFM coefficients presented in [13], and ‘upw’ are the upwind functions:

$$\mathrm{upw}\big[f(C), v\big] = \begin{cases} f(C_+)v, & v \geq 0, \\ f(C_-)v, & v < 0, \end{cases} \quad \mathrm{upw}_{mf}\big[f(C), v\big] = \begin{cases} f(C_m)v, & v \geq 0, \\ f(C_f)v, & v < 0, \end{cases}$$

The resulting system of algebraic equations is nonlinear due to nonlinearity of the two-phase flow model, and the Newton method is used to solve it. Using the nonlinear flux discretization scheme does not introduce additional complexity for the nonlinear solver. However, in spite of being formally two-point, the nonlinear scheme produces a multi-point stencil for the Jacobian matrix, which results in more expensive linear system solution (on average, extra 25–100%) compared to the linear TPFA scheme. For more details about the Newton method and the construction of the Jacobian matrix for the nonlinear scheme we refer to [12].

4 Numerical Experiment for Two-Phase Flow

For the numerical experiment we simulate the two-phase flow for a standard five-spot problem with two wells in the opposite corners of a rectangular domain, and add two fractures as shown in Fig. 2.

The permeability tensor for the porous media is full anisotropic:

$$\mathbb{K}^m = R_z(-\alpha)\begin{pmatrix} k_1 & 0 & 0 \\ 0 & k_2 & 0 \\ 0 & 0 & k_3 \end{pmatrix} R_z(\alpha), \qquad R_z(\alpha) = \begin{pmatrix} \cos\alpha & \sin\alpha & 0 \\ -\sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

where $k_1 = 10^3$ [md], $k_2 = k_3 = 10^2$ [md], $\alpha = \frac{\pi}{4}$, and the porosity is $\phi^m = 0.15$.

The permeability tensor for the fractures is scalar $\mathbb{K}^f = k^f \mathbb{I}$, $k^f = 10^6$ [md], $w_f = 0.13$ [ft] and the porosity is $\phi^f = 0.15$.

Domain dimensions are: $[0, 100] \times [0, 100] \times [0, 10]$ ft. Tables for capillary pressure and relative permeabilities are similar to the two-phase flow experiments from [12]. For the wells we set the bottom hole pressures $p_{inj} = 4100$ [psi] and

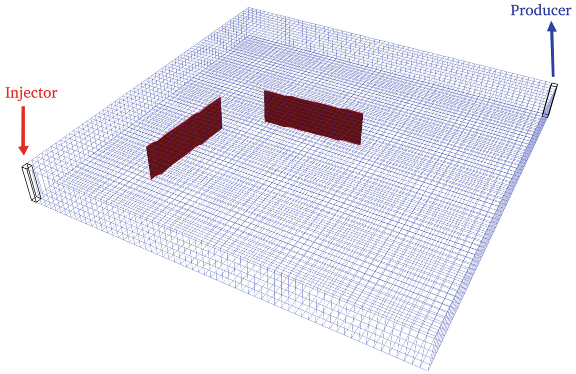


Fig. 2 Setup for the five-spot problem with two fractures

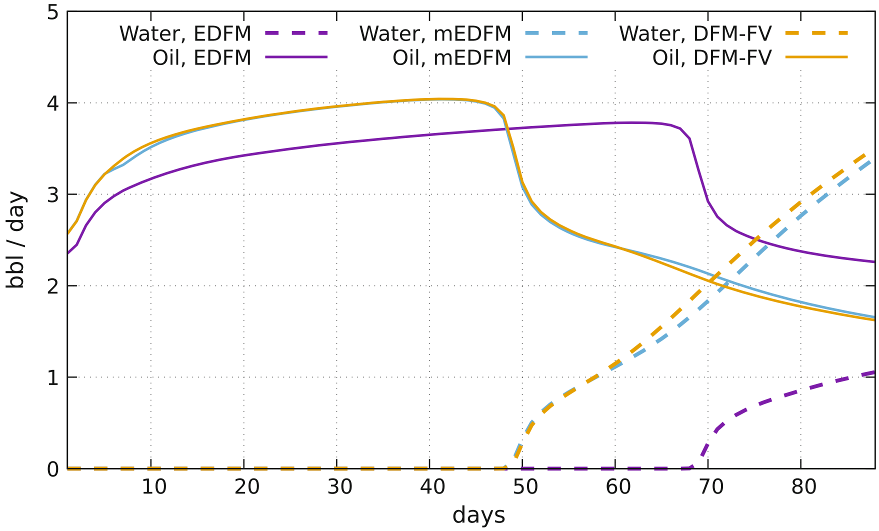


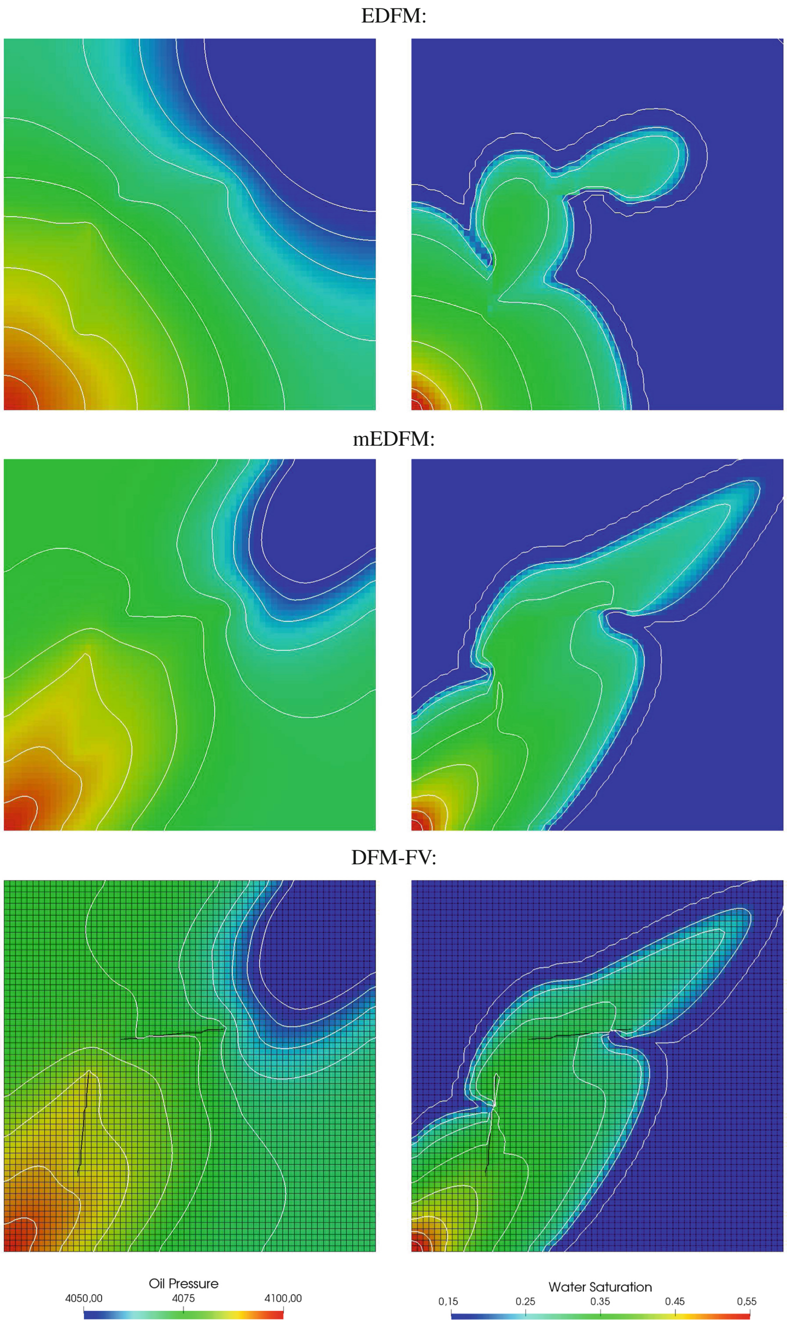
Fig. 3 Oil and water rates for EDFM, mEDFM (NTPFA) and DFM-FV (NTPFA) solutions

$p_{prod} = 3900\,[\text{psi}]$. The initial pressure is $p_0 = 4000\,[\text{psi}]$, and the initial saturation is $S_0 = 0.15$.

We simulate water injection for 90 days with time step $\Delta t = 1$ day and compare three solutions: (1) the EDFM solution with the linear TPFA discretization for all flux types, (2) the mEDFM solution with the NTPFA discretization, and (3) the discrete fracture method (DFM-FV) solution with the NTPFA scheme, which directly applies the original FV discretization for the mesh with cut-cells and a thin layer of 3D cells representing the fracture.

Water and oil rates for the producer well are shown in the Fig. 3. The mEDFM and the DFM-FV schemes produce very close results with similar rates and breakthrough times since the NTPFA scheme provides the approximation for non-$\mathbb{K}$-orthogonal grids. On the contrast, the original EDFM provides a different solution, with 40% larger breakthrough time.

Figure 4 shows the oil pressure and the water saturation fields at the time $T = 45$ days. One can see that the mEDFM (NTPFA) and the DFM-FV methods produce almost identical results, whereas the EDFM solution is noticeably different from them. It should be noted that the DFM-FV requires grid modification to take fractures into account explicitly, which may complicate the reservoir simulation. The mEDFM provides a viable alternative.


EDFM:

mEDFM:

DFM-FV:

Fig. 4 Oil pressure (left) and water saturation (right) fields for the two-phase flow, T = 45 days. Top: EDFM solution; middle: mEDFM (NTPFA) solution; bottom: DFM-FV (NTPFA) solution

5 Conclusion

We present the application of the new monotone embedded discrete fracture method (mEDFM) for the two-phase flows in fractured media. The method combines the EDFM approach with the monotone nonlinear two-point flux approximation.

Numerical experiments show that in anisotropic media the two-phase flow modelling with the mEDFM provides the accurate solution (in contrast to the conventional EDFM), and is in a good agreement with the discrete fracture method, which assumes explicit representation of the fractures by the grid.

Acknowledgements The development of the monotone embedded discrete fracture method was supported by Russian Foundation for Basic Research, project number 19-31-90110. The development of the two-phase flow numerical model and the mesh modification method for DFM-FV was supported by Moscow Center for Fundamental and Applied Mathematics (agreement with the Ministry of Education and Science of the Russian Federation No. 075-15-2019-1624).

References

1. Aavatsmark, I., Eigestad, G., Mallison, B., Nordbotten, J.: A compact multipoint flux approximation method with improved robustness. Num. Meth. Part. D. E. **24**(5), 1329–1360 (2008)
2. Chernyshenko, A., Vassilevski, Y.: A finite volume scheme with the discrete maximum principle for diffusion equations on polyhedral meshes. In: Finite Volumes for Complex Applications VII-Methods and Theoretical Aspects, pp. 197–205. Springer, Berlin (2014)
3. Danilov, A., Vassilevski, Y.: A monotone nonlinear finite volume method for diffusion equations on conformal polyhedral meshes. Russ. J. Numer. Anal. Math. Model. **24**(3), 207–227 (2009)
4. Hajibeygi, H., Karvounis, D., Jenny, P.: A hierarchical fracture model for the iterative multiscale finite volume method. J. Comput. Phys. **230**, 8729–8743 (2011)
5. Jiang, J., Younis, R.: An improved projection-based embedded discrete fracture model (pEDFM) for multiphase flow in fractured reservoirs. Adv. Water Resour. **109**, 267–289 (2017)
6. Kumar, K., List, F., Pop, I., Radu, F.: Formal upscaling and numerical validation of fractured flow models for richards equation. J. Comput. Phys. **407**, 109138 (2019)
7. Lee, S.H., Lough, M.F., Jensen, C.L.: Hierarchical modeling of flow in naturally fractured formations with multiple length scales. Water Resour. Res. **37**(3), 443–455 (2001)
8. Li, L., Lee, S.H.: Efficient field-scale simulation of black oil in a naturally fractured reservoir through discrete fracture networks and homogenized media. SPE Reserv. Eval. Eng. **11**, 750–758 (2008)
9. Lipnikov, K., Svyatskiy, D., Vassilevski, Y.: Minimal stencil finite volume scheme with the discrete maximum principle. Russ. J. Numer. Anal. Math. Modelling **27**(4), 369–385 (2012)
10. Moinfar, A., Varavei, A., Sepehrnoori, K., Johns, R.T.: Development of an efficient embedded discrete fracture model for 3D compositional reservoir simulation in fractured reservoirs. SPE J. **19**(2), 289–303 (2014)
11. Nikitin, K., Novikov, K., Vassilevski, Y.: Nonlinear finite volume method with discrete maximum principle for the two-phase flow model. Lobachevskii J. Math. **37**(5), 570–581 (2016)
12. Nikitin, K., Terekhov, K., Vassilevski, Y.: A monotone nonlinear finite volume method for diffusion equations and multiphase flows. Comput. Geosci. **18**(3–4), 311–324 (2014)
13. Nikitin, K., Yanbarisov, R.: Monotone embedded discrete fractures method for flows in porous media. J. Comput. Appl. Math. **364**, 112353 (2020)
14. Peaceman, D.W.: Interpretation of well-block pressures in numerical reservoir simulation. SPE J. **18**(3), 183–194 (1978)

A Robust VAG Scheme for a Two-Phase Flow Problem in Heterogeneous Porous Media

Konstantin Brenner, R. Masson, and E. H. Quenjel

Abstract A positive Vertex Approximate Gradient (VAG) scheme is proposed to discretize the total velocity formulation of two-phase Darcy flow problems in heterogeneous porous media. The discretization is based on the physical variables and allows for multiple rock types with highly contrasted petrophysical and hydrodynamical properties. The numerical experiment shows that, compared to the Phase Potential Upwind (PPU) version of VAG scheme, this new discretization is more robust and efficient in terms of nonlinear convergence.

Keywords Two-phase Darcy flow · Finite volume scheme · Heterogeneous capillary pressure · VAG scheme

MSC 35K65 · 65M08 · 65M22 · 6T99

1 Introduction

We apply the Vertex Approximate Gradient (VAG) discretization [4] to the two-phase Darcy flow problem written in the total velocity formulation [2]. The choice of the numerical scheme is motivated by the ability of the VAG discretization to properly capture saturation jumps across different rock types. An upwind approximation [7] of the capillary diffusion term with respect to the capillary VAG fluxes is proposed in order to avoid possible undershoots and overshoots of the saturation, that could result from the non monotonicity of the VAG capillary fluxes. The time integration is chosen

K. Brenner · R. Masson · E. H. Quenjel (✉)
Laboratoire J. A. Dieudonné, Team Coffee, Université Côte d'Azur, Inria, CNRS,
Parc Valrose, 06108 Nice cedex 02, France
e-mail: el-houssaine.quenj@univ-cotedazur.fr

K. Brenner
e-mail: konstantin.brenner@univ-cotedazur.fr

R. Masson
e-mail: roland.masson@univ-cotedazur.fr

R. Klöfkorn et al. (eds.), *Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples*, Springer Proceedings in Mathematics & Statistics 323,
https://doi.org/10.1007/978-3-030-43651-3_53

implicit to avoid severe time step restrictions in high velocity regions. For stability reasons, the pressure and saturation are fully coupled to account for the nonlinear transmission conditions at different rock type interfaces [1]. The resulting discretization can be viewed as an extension of the Hybrid Upwind (HU) transport scheme to the VAG discretization. The HU scheme was developed in the framework of the Two-Point Flux Approximation [6] in order to improve the convergence of the nonlinear solver. The authors used only one unknown per rock type interface, typically the saturation or the capillary pressure, whereas our discretization uses both pressure and saturation interface variables. This latter approach allows to accurately represent the capillary barriers. The positive approximation of the transport and capillary diffusion terms ensure a discrete maximum principle on the saturations. Following [3], our choice of the primary unknowns at interface nodes between different rock types is based on a generalization of variable switch techniques which allows to stabilize Newton's method. As shown in the numerical section, this new HU VAG discretization provides faster nonlinear convergence than the VAG discretization based on the Phase Potential Upwinding (PPU) [3].

2 Two-Phase Darcy Flow Model

Let $t_f > 0$ and Ω be a bounded domain of $\mathbb{R}^d$ $(d \geq 1)$ such that $\overline{\Omega} = \bigcup_{\mathrm{rt}\in\mathscr{RT}} \overline{\Omega}_{\mathrm{rt}}$, where $\mathscr{RT}$ is the set of rock types. The total velocity formulation of the two-phase Darcy flow model reads

$$\left\{\begin{array}{l} \phi(\mathbf{x})\partial_t s^{\mathrm{nw}} + \mathrm{div}\Big(f^{\mathrm{nw}}\mathbf{V}^T - D\Lambda(\mathbf{x})\big(\nabla p_c + (\rho_{\mathrm{w}} - \rho_{\mathrm{nw}})\mathbf{g}\big)\Big) = 0, \\ \mathrm{div}\mathbf{V}^T = 0, \\ \mathbf{V}^T = \mathbf{V}^{\mathrm{nw}} + \mathbf{V}^{\mathrm{w}} = -\sum\limits_{\alpha\in\{\mathrm{nw},\,\mathrm{w}\}} \eta^\alpha(\mathbf{x}, s^\alpha)\Lambda(\mathbf{x})(\nabla p^\alpha - \rho_\alpha \mathbf{g}), \\ p_c = p^{\mathrm{nw}} - p^{\mathrm{w}} \in \widetilde{P}_c(\mathbf{x}, s^{\mathrm{nw}}), \\ s^{\mathrm{nw}} + s^{\mathrm{w}} = 1, \end{array}\right. \tag{1}$$

with {nw, w} denoting the set of non-wetting and wetting phases. In (1), $\phi(\mathbf{x})$ denotes the porosity, s^α the phase saturation, $\Lambda(\mathbf{x})$ the permeability tensor, p^α the phase pressure, $\mathbf{V}^T$ the total velocity, and p_c the capillary pressure. The phase density ρ_α is assumed constant. The gravity acceleration vector is denoted by $\mathbf{g}$ and its norm by g. The phase mobility function $\eta^\alpha(\mathbf{x}, s^\alpha)$ is defined as the ratio of the relative permeability of the phase over its viscosity. Let $\widetilde{P}_c(\mathbf{x}, s)$ denote the monotone graph extension of the capillary pressure function (see [1]). It is assumed that $\widetilde{P}_c$ and η^α are spatially homogeneous in each subdomain Ω_{rt}, rt $\in \mathscr{RT}$; in addition we assume that the total mobility function $\eta(\mathbf{x}, s) = \eta^{\mathrm{nw}}(\mathbf{x}, s) + \eta^{\mathrm{w}}(\mathbf{x}, 1 - s)$ verifies $\eta(\mathbf{x}, s) \geq \eta_{\min} > 0$. We then denote by f^{nw} the non-wetting phase fractional flow function $f^{\mathrm{nw}}(\mathbf{x}, s) = \eta^{\mathrm{nw}}(\mathbf{x}, s)/\eta(\mathbf{x}, s)$, and by $D(\mathbf{x}, s)$ the capillary diffusion coefficient $D(\mathbf{x}, s) = \eta^{\mathrm{w}}(\mathbf{x}, 1 - s)f^{\mathrm{nw}}(\mathbf{x}, s)$. The system (1) is completed by some initial distribution of s^{nw} and the boundary conditions

$$\mathbf{V}^{\alpha} \cdot \mathbf{n} = 0 \quad \text{on } \Gamma^{N} \times (0, t_f), \quad p^{\alpha} = p^{\alpha}_{\text{Dir}} \quad \text{on } \Gamma^{\text{Dir}} \times (0, t_f) \quad \text{for } \alpha \in \{\text{nw}, \text{w}\},$$

where $\mathbf{n}$ is the outward normal to Γ^{N}, and $\partial\Omega = \Gamma^{N} \cup \Gamma^{\text{Dir}}$ with $\left|\Gamma^{\text{Dir}}\right| > 0$.

3 Positive VAG Discretization for Two-Phase Darcy Flows

3.1 VAG Mesh, Fluxes and Pore Volumes

The VAG discretization considers generalized polyhedral meshes of Ω [4]. Let us briefly recall some notations. Let $\mathcal{M}$ be the set of polyhedral cells of Ω. For each $k \in \mathcal{M}$ we denote the set of nodes of the cell k by $\mathcal{V}_k$ and we also denote by $\mathcal{V} = \bigcup_{k\in\mathcal{M}} \mathcal{V}_k$ the set of all vertices of the mesh, and by $\mathcal{M}_{\mathbf{s}}$ the subset of cells sharing the node $\mathbf{s} \in \mathcal{V}$. Note that the mesh is supposed to be conforming w.r.t. the partition of Ω in subdomains Ω_{rt}, $\text{rt} \in \mathcal{RT}$, and w.r.t. the partition $\{\Gamma^{N}, \Gamma^{\text{Dir}}\}$ of $\partial\Omega$. We then denote by $\mathcal{V}_{\text{Dir}}$ the set of nodes located at $\overline{\Gamma}^{\text{Dir}}$.

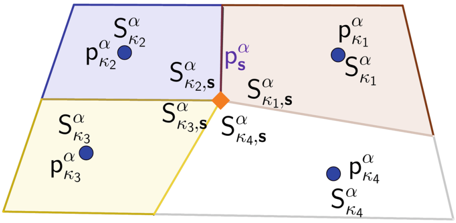
Let $X_{\mathcal{D}} = \{v_k \in \mathbb{R}, v_{\mathbf{s}} \in \mathbb{R}, \text{ for } k \in \mathcal{M}, \mathbf{s} \in \mathcal{V}\}$ be the vector space of degrees of freedom (d.o.f.). The VAG scheme is a control volume scheme in the sense that it results, for each d.o.f. not located at the Dirichlet boundary, in a volume balance equation. The two main ingredients are therefore the conservative fluxes and the pore volumes. For $u_{\mathcal{D}} \in X_{\mathcal{D}}$, the VAG fluxes $F_{k,\mathbf{s}}(u_{\mathcal{D}})$ connect the cell $k \in \mathcal{M}$ to its nodes $\mathbf{s} \in \mathcal{V}_k$. They can be expressed as $F_{k,\mathbf{s}}(u_{\mathcal{D}}) = \sum\limits_{\mathbf{s}'\in\mathcal{V}_k} \mathbb{T}_k^{\mathbf{s},\mathbf{s}'}(u_k - u_{\mathbf{s}'})$, with $(\mathbb{T}_k^{\mathbf{s},\mathbf{s}'})_{\mathbf{s},\mathbf{s}'\in\mathcal{V}_k}$ a symmetric positive definite matrix obtained from the $\mathbb{P}_1$ finite element subspace defined on a tetrahedral submesh of $\mathcal{M}$.

As described in [3], the portions $\phi_{k,\mathbf{s}}$ of each cell's pore volume $\int_k \phi(\mathbf{x})d\mathbf{x}$ is distributed to its nodes $\mathbf{s} \in \mathcal{V}_k \setminus \mathcal{V}_{\text{Dir}}$. We then define $\phi_k = \int_k \phi(\mathbf{x})d\mathbf{x} - \sum_{\mathbf{s}\in\mathcal{V}_k\setminus\mathcal{V}_{\text{Dir}}} \phi_{k,\mathbf{s}}$ as the remaining cell pore volume.

3.2 Choice of the Primary Unknowns

We recall that the mesh is conforming w.r.t. the rock type subdomains, therefore a single rock type rt_k is assigned to each cell $k \in \mathcal{M}$. We denote by $\chi_{\mathbf{s}} = \{\text{rt}_k,\ k \in \mathcal{M}_{\mathbf{s}}\}$ the set of rock types surrounding the node $\mathbf{s} \in \mathcal{V}$, and we set $\chi_k = \{\text{rt}_k\}$ for all $k \in \mathcal{M}$.

The choice of the primary variables follows the variable switching strategy introduced in [3]. We use the pressure of the non-wetting phase as the first primary variable for all d.o.f.; then for the cells and the nodal d.o.f. associated with a single rock type the second primary unknown is the saturation, while for the nodes $\mathbf{s}$ located at rock type interfaces we invoke the variable switching based on a parametrization of $\widetilde{P}_{c,\text{rt}}$, $\text{rt} \in \chi_{\mathbf{s}}$. For such nodes we construct a set of non-decreasing con-



$\mathsf{S}^{\alpha}_{\kappa 2}$
$\mathsf{p}^{\alpha}_{\kappa 2}$
$\mathsf{S}^{\alpha}_{\kappa 2,\mathbf{s}}$
$\mathsf{p}^{\alpha}_{\mathbf{s}}$
$\mathsf{p}^{\alpha}_{\kappa 1}$
$\mathsf{S}^{\alpha}_{\kappa 1}$
$\mathsf{S}^{\alpha}_{\kappa 1,\mathbf{s}}$
$\mathsf{S}^{\alpha}_{\kappa 3}$
$\mathsf{S}^{\alpha}_{\kappa 3,\mathbf{s}}$
$\mathsf{S}^{\alpha}_{\kappa 4,\mathbf{s}}$
$\mathsf{p}^{\alpha}_{\kappa 3}$
$\mathsf{p}^{\alpha}_{\kappa 4}$
$\mathsf{S}^{\alpha}_{\kappa 4}$

Fig. 1 Phase pressure and saturation discrete unknowns in the four cells sharing the vertex **s**. Each color represents a possible different rock type

tinuous functions $P_{c,\chi_{\mathbf{s}}}(\tau)$ and $\big(S^{\rm nw}_{\chi_{\mathbf{s}},{\rm rt}}(\tau)\big)_{{\rm rt}\in\chi_{\mathbf{s}}}$ defined for $\tau\in[0,1]$ and satisfying $P_{c,\chi_{\mathbf{s}}}(\tau)\in\widetilde{P}_{c,{\rm rt}}(S^{\rm nw}_{\chi_{\mathbf{s}},{\rm rt}}(\tau))$ for all ${\rm rt}\in\chi_{\mathbf{s}}$ and $\tau\in[0,1]$; in addition we require $P_{c,\chi_{\mathbf{s}}}(\tau)+\sum_{{\rm rt}\in\chi_{\mathbf{s}}}S^{\rm nw}_{\chi_{\mathbf{s}},{\rm rt}}(\tau)$ to be strictly increasing. We note that this parametrization allows to deal both with a vanishing capillary diffusion and the capillary barriers. The functions $\tau\mapsto P_{c,\chi_{\mathbf{s}}}(\tau)$, $\big(S^{\rm nw}_{\chi_{\mathbf{s}},{\rm rt}}(\tau)\big)_{{\rm rt}\in\chi_{\mathbf{s}}}$ have to be chosen carefully in order to improve the nonlinear solver and we refer to [3] for a detailed discussion. In order to unify the notations, for the cells or the nodes associated with a single rock type, we introduce a trivial parametrization defined by $S^{\rm nw}_{\chi,{\rm rt}}(\tau)=\tau$ for χ reduced to a single rock type.

Given the primary unknowns $p^{\rm nw}_{\mathscr{D}}=(p^{\rm nw}_{\nu})_{\nu\in\mathscr{M}\cup\mathscr{V}}$ and $\tau_{\mathscr{D}}=(\tau_{\nu})_{\nu\in\mathscr{M}\cup\mathscr{V}}$, we define

$$\left\{\begin{array}{lll}
p_{c,\mathscr{D}} & =(p_{c,\nu})_{\nu\in\mathscr{M}\cup\mathscr{V}}, & \text{with } p_{c,\nu}=P_{c,\chi_{\nu}}(\tau_{\nu}),\\
p^{\rm w}_{\mathscr{D}} & =(p^{\rm w}_{\nu})_{\nu\in\mathscr{M}\cup\mathscr{V}}, & \text{with } p^{\rm w}_{\nu}=p^{\rm nw}_{\nu}-p_{c,\nu},\\
\Phi^{\alpha}_{\mathscr{D}} & =p^{\alpha}_{\mathscr{D}}+\rho_{\alpha}gZ_{\mathscr{D}}, & \text{with } Z_{\mathscr{D}}=(z_{\nu})_{\nu\in\mathscr{M}\cup\mathscr{V}},\\
s^{\alpha}_{k} & =S^{\alpha}_{\chi_k,{\rm rt}_k}(\tau_k), & k\in\mathscr{M},\\
s^{\alpha}_{k,\mathbf{s}} & =S^{\alpha}_{\chi_{\mathbf{s}},{\rm rt}_k}(\tau_{\mathbf{s}}), & \mathbf{s}\in\mathscr{V},\ k\in\mathscr{M}_{\mathbf{s}}.
\end{array}\right.$$

To sum up, as exhibited in Fig. 1, the discrete phase pressure is single-valued for all d.o.f. while the saturation is single-valued for all cells (and single rock type nodes) and multi-valued at any node sharing multiple rock types.

3.3 Hybrid Upwinding (HU) VAG Scheme for the Diphasic Model

The gravity and capillary gradient fluxes are defined by

$$G_{k,\mathbf{s}}=(\rho_{\rm nw}-\rho_{\rm w})gF_{k,\mathbf{s}}\Big(Z_{\mathscr{D}}\Big),\quad C_{k,\mathbf{s}}=F_{k,\mathbf{s}}(p_{c,\mathscr{D}}).$$

Let us introduce the total velocity fluxes $V^T_{k,\mathbf{s}}=V^T_{k,\mathbf{s}}(p^{\rm nw}_{\mathscr{D}},\tau_{\mathscr{D}})$, for all $k\in\mathscr{M}$, $\mathbf{s}\in\mathscr{V}_k$, as well as the following discrete phase saturation functions at each d.o.f.

$$\gamma^{\alpha}_{k}(\tau)=S^{\alpha}_{\chi_k,{\rm rt}_k}(\tau),\quad k\in\mathscr{M},\quad \gamma^{\alpha}_{\mathbf{s}}(\tau)=\sum_{k\in\mathscr{M}_{\mathbf{s}}}\frac{\phi_{k,\mathbf{s}}}{\phi_{\mathbf{s}}}S^{\alpha}_{\chi_{\mathbf{s}},{\rm rt}_k}(\tau),\quad \mathbf{s}\in\mathscr{V}\setminus\mathscr{V}_{\rm Dir}.$$

with $\phi_{\mathbf{s}} = \sum_{k\in\mathcal{M}_{\mathbf{s}}} \phi_{k,\mathbf{s}}$. For $N \in \mathbb{N}^*$, we consider the time subdivision $t^0 = 0 < t^1 < \cdots < t^{n-1} < t^n \cdots < t^N = t_f$ of $[0, t_f]$. We denote the time steps by $\Delta t^n = t^n - t^{n-1}$ for all $n = 1, \cdots, N$. In the sequel, we omit the time superscript t^n in the flux terms. Then, for a given $\tau^0_{\mathcal{D}} \in [0, 1]^{\mathcal{M}\cup\mathcal{V}}$, the scheme consists in finding $(p^{\mathrm{nw},n}_{\mathcal{D}}, \tau^n_{\mathcal{D}})$, solutions of the following system of equations for $\alpha \in \{\mathrm{nw}, \mathrm{w}\}$:

$$\begin{cases} \dfrac{\phi_k}{\Delta t^n}\Big(\gamma^{\alpha}_k(\tau^n_k) - \gamma^{\alpha}_k(\tau^{n-1}_k)\Big) \\ \quad + \displaystyle\sum_{\mathbf{s}\in\mathcal{V}_k} f^{\mathrm{nw}}_{k,\mathbf{s}}\ V^T_{k,\mathbf{s}} + D^{\mathrm{cap}}_{k,\mathbf{s}} C_{k,\mathbf{s}} + D^{\mathrm{g}}_{k,\mathbf{s}} G_{k,\mathbf{s}} = 0, \quad k \in \mathcal{M}, \\ \dfrac{\phi_{\mathbf{s}}}{\Delta t^n}\Big(\gamma^{\alpha}_{\mathbf{s}}(\tau^n_{\mathbf{s}}) - \gamma^{\alpha}_{\mathbf{s}}(\tau^{n-1}_{\mathbf{s}})\Big) \\ \quad - \displaystyle\sum_{k\in\mathcal{M}_{\mathbf{s}}} f^{\mathrm{nw}}_{k,\mathbf{s}}\ V^T_{k,\mathbf{s}} + D^{\mathrm{cap}}_{k,\mathbf{s}}\ C_{k,\mathbf{s}} + D^{\mathrm{g}}_{k,\mathbf{s}}\ G_{k,\mathbf{s}} = 0, \ \mathbf{s} \in \mathcal{V}\setminus\mathcal{V}_{\mathrm{Dir}}, \\ p^{\mathrm{nw},n}_{\mathbf{s}} = p^{\mathrm{nw}}_{\mathbf{s},\mathrm{Dir}}, \quad \tau^n_{\mathbf{s}} = \tau_{\mathbf{s},\mathrm{Dir}}, \qquad\qquad \mathbf{s} \in \mathcal{V}_{\mathrm{Dir}}. \end{cases} \tag{2}$$

Summing the conservation equations over the phases $\alpha \in \{\mathrm{nw}, \mathrm{w}\}$ provides the discrete divergence-free property of the total velocity fluxes

$$\sum_{\mathbf{s}\in\mathcal{V}_k} V^T_{k,\mathbf{s}} = 0 \text{ for all } k \in \mathcal{M}, \quad \sum_{k\in\mathcal{M}_{\mathbf{s}}} V^T_{k,\mathbf{s}} = 0 \text{ for all } \mathbf{s} \in \mathcal{V}\setminus\mathcal{V}_{\mathrm{Dir}}. \tag{3}$$

Fractional flow term: We first specify two expressions of $V^T_{k,\mathbf{s}}$ using the upwind mobilities

$$V^T_{k,\mathbf{s}} = \sum_{\alpha\in\{\mathrm{nw},\,\mathrm{w}\}} \eta^{\alpha}_{\mathrm{rt}_k}(s^{\alpha}_k) F_{k,\mathbf{s}}(\Phi^{\alpha}_{\mathcal{D}})^+ - \eta^{\alpha}_{\mathrm{rt}_k}(s^{\alpha}_{k,\mathbf{s}}) F_{k,\mathbf{s}}(\Phi^{\alpha}_{\mathcal{D}})^-, \tag{4}$$

with $x^{\pm} = \max(\pm x, 0)$, or alternatively, using the cell mobilities

$$V^T_{k,\mathbf{s}} = \sum_{\alpha\in\{\mathrm{nw},\,\mathrm{w}\}} \eta^{\alpha}_{\mathrm{rt}_k}(s^{\alpha}_k) F_{k,\mathbf{s}}(\Phi^{\alpha}_{\mathcal{D}}). \tag{5}$$

This last choice is expected to be more stable than (4). Then, the fractional flow fluxes are defined by

$$f^{\mathrm{nw}}_{k,\mathbf{s}}\ V^T_{k,\mathbf{s}} = f^{\mathrm{nw}}_k(s^{\mathrm{nw}}_k)\ (V^T_{k,\mathbf{s}})^+ - f^{\mathrm{nw}}_k(s^{\mathrm{nw}}_{k,\mathbf{s}})\ (V^T_{k,\mathbf{s}})^-.$$

Capillary term: The capillary gradient flux is not monotone. To tackle this issue, we perform a positive correction as follows

$$D^{\mathrm{cap}}_{k,\mathbf{s}} C_{k,\mathbf{s}} = \frac{\eta^{\mathrm{nw}}_{\mathrm{rt}_k}(s^{\mathrm{nw}}_k)\eta^{\mathrm{w}}_{\mathrm{rt}_k}(s^{\mathrm{w}}_{k,\mathbf{s}})}{\eta_{\mathrm{rt}_k}(s^{\mathrm{nw},n-1}_k)} C^+_{k,\mathbf{s}} - \frac{\eta^{\mathrm{nw}}_{\mathrm{rt}_k}(s^{\mathrm{nw}}_{k,\mathbf{s}})\eta^{\mathrm{w}}_{\mathrm{rt}_k}(s^{\mathrm{w}}_k)}{\eta_{\mathrm{rt}_k}(s^{\mathrm{nw},n-1}_k)} C^-_{k,\mathbf{s}}. \tag{6}$$

Gravity term: The gravity contribution is defined similarly as in (6).

Remark 3.1 Note that the explicit approximation $\eta_{\mathrm{rt}_k}(s_k^{\mathrm{nw},n-1})$ is considered in (6). This choice improves the nonlinear convergence without involving any restriction on the time step.

We now state two properties of the scheme (2) and (3) that can be proved assuming that the capillary functions are bounded. We refer to [5] for the proof.

Proposition 3.1 *Let $\tau_{\mathscr{D}}^0 \in [0,1]^{\mathscr{M}\cup\mathscr{V}}$, then, every solution $(p_{\mathscr{D}}^{\mathrm{nw},n}, \tau_{\mathscr{D}}^n)$, to the system (2) and (3) ensures $\tau_{\mathscr{D}}^n \in [0,1]^{\mathscr{M}\cup\mathscr{V}}$, meaning that the saturations and the capillary pressure satisfy the physical bounds.*

Proposition 3.2 *Let $\tau_{\mathscr{D}}^0 \in [0,1]^{\mathscr{M}\cup\mathscr{V}}$ and let us assume that the total velocity fluxes are given. Then, the saturation equation (2) admits a solution $\tau_{\mathscr{D}}^n$.*

4 Numerical Results

We test the ability of the introduced HU VAG scheme to simulate oil migration in a fractured reservoir. It is compared to the PPU VAG version [3] which upwinds the phase mobility w.r.t the phase potential. The domain is $\Omega = (0, 100\,\mathrm{m})^3$ containing a network of planar fractures of aperture 1 cm (see Fig. 2). The matrix porosity is set to $\phi_m = 0.2$ and the fracture porosity to $\phi_f = 0.4$. The matrix permeability is isotropic and set to $\Lambda_m = 10^{-16}\,\mathrm{m}^2$ while the fracture tangential permeability is set to $\Lambda_f = 10^{-10}\,\mathrm{m}^2$. The oil density is $\rho_{\mathrm{nw}} = 700\,\mathrm{Kg/m}^3$ and the water density is $\rho_{\mathrm{w}} = 1000\,\mathrm{Kg/m}^3$. The oil viscosity is set to $\mu^{\mathrm{nw}} = 0.005\,\mathrm{Pa.s}$ and the water viscosity is $\mu^{\mathrm{w}} = 0.001\,\mathrm{Pa.s}$. The fracture and matrix relative permeabilities are given by $k_{r,f}^{\alpha}(s^{\alpha}) = (s^{\alpha})^{1.2}$ and $k_{r,m}^{\alpha}(s^{\alpha}) = (s^{\alpha})^2$, $\alpha = nw, w$. The capillary pressure is $P_{c,m}(s^{nw}) = -10^4\log(1-s^{nw})$ in the matrix and $P_{c,f}(s^{nw}) = -10^3\log(1-s^{nw})$ in the fracture network. The domain is meshed using a tetrahedral mesh with 47670 cells and 1670 fracture faces. The reservoir is initially saturated with water. Dirichlet

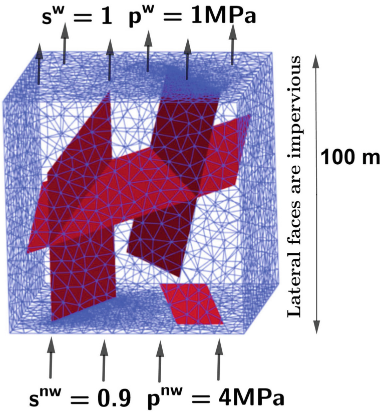

$s^w = 1$ $p^w = 1\mathrm{MPa}$
Lateral faces are impervious
100 m
$s^{nw} = 0.9$ $p^{nw} = 4\mathrm{MPa}$

Fig. 2 Test case configuration

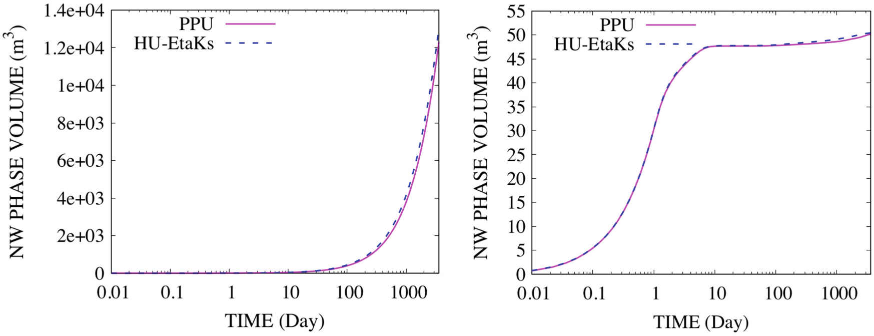


Fig. 3 Oil saturation volumes in the matrix (left) and in the fracture network (right) as a function of time obtained for the PPU and HU-EtaKs VAG schemes

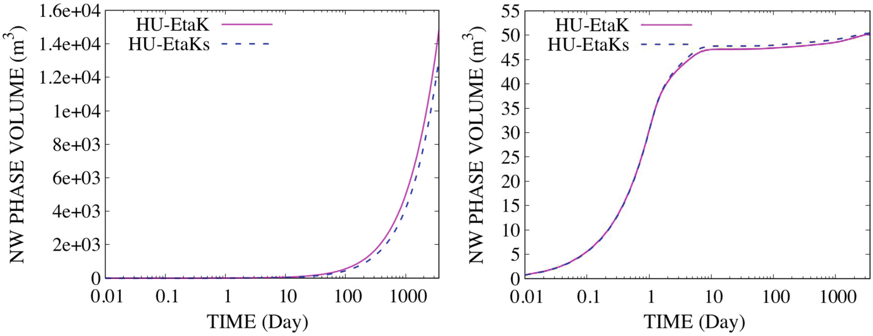


Fig. 4 Oil saturation volumes in the matrix (left) and in the fracture network (right) as a function of time obtained for the HU VAG schemes with EtaKs and EtaK

Table 1 Number of the time steps $N_{\Delta t}$, number of the time step chops N_{chop}, average number of Newton iterations per time step N_{Newton}, average number of GMRES iterations per Newton step N_{GMRes} and CPU time for the three VAG schemes

Scheme	$N_{\Delta t}$	N_{chop}	N_{Newton}	N_{GMRes}	CPU (s)
PPU VAG	106	8	6.7	14.7	507
HU-EtaKs VAG	82	0	3.9	13.4	205
HU-EtaK VAG	82	0	4.0	13.9	203

boundary conditions are imposed at the top boundary with a wetting phase pressure of 1 MPa and $s^w = 1$, while at the lower boundary, intersected by the fracture network, we impose $p^w = 4$ MPa and the capillary pressure resulting in the matrix saturation s^{nw} equal to 0.9. The lateral boundaries are assumed impervious and the final time is fixed to $t_f = 3600$ days. The time stepping is defined by $\Delta t^1 = \Delta t_{init}$ and for all $n \geq 1$ by $\Delta t^{n+1} = \max(\Delta t_{max}, 1.2\Delta t^n)$, in case of a successful time step Δt^n, and $\Delta t^{n+1} = 0.5\Delta t^n$, if Newton's method fails to converge in 25 iterations. We have used the values $\Delta t_{init} = 0.01$ and $\Delta t_{max} = 100$ days.

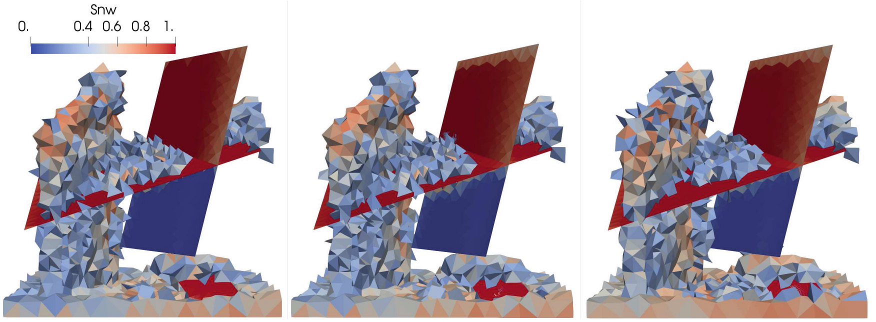
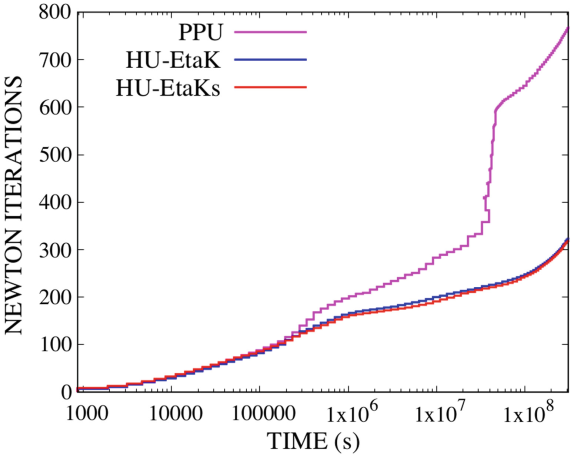


Fig. 5 Oil saturation in the matrix for $s^{nw} > 0.25$ and in the fractures at final time obtained for the PPU (left) and HU-EtaKs (middle) and HU-EtaK (right) VAG schemes

Fig. 6 Total number of Newton iterations as a function of time for the three VAG schemes



We display in Figs. 3, 4, 5 and 6 and Table 1 the results obtained for the PPU and HU VAG discretizations corresponding to the upwind mobility (4) labeled with HU-EtaKs and the centered one (5) labeled with HU-EtaK. Figure 3 shows a very good match between the PPU and HU-ETaKs schemes while Figs. 4 and 5 exhibit small differences in the matrix between both HU VAG schemes.

The increased robustness of the nonlinear convergence provided by the HU VAG schemes compared with the PPU version is clearly seen in this Table 1 as well as in Fig. 6.

In conclusion, the positive HU VAG scheme provides similar solutions than the PPU version and exhibits an additional robustness in terms of nonlinear convergence for the simulation of highly heterogeneous media. Let us refer to [5] for more test-cases including large fracture networks.

Acknowledgements This work was supported by the French National Research Agency under the CHARMS ANR project ANR-16-CE06-0009.

References

1. Andreianov, B., Brenner, K., Cancès, C.: Approximating the vanishing capillarity limit of two-phase flow in multi-dimensional heterogeneous porous medium. ZAMM - J. App. Math. Mecha./Zeit. für Ange. Math. und Mecha. **94** (7–8), 655–667 (2014)
2. Brenier, Y., Jaffré, J.: Upstream differencing for multiphase flow in reservoir simulation. SIAM J. Numer. Anal. **28**(31), 685–696 (1991)
3. Brenner, K., Groza, M., Jeannin, L., Masson, R., Pellerin, J.: Immiscible two-phase Darcy flow model accounting for vanishing and discontinuous capillary pressures: application to the flow in fractured porous media. Comput. Geosci. **21**(5–6), 1075–1094 (2017)
4. Brenner, K., Masson, R.: Convergence of a vertex centred discretization of two-phase Darcy flows on general meshes. Int. J. Finite Vol. **10**, 1–37 (2013)
5. Brenner, K., Masson, R., Quenjel, E.H.: Vertex Approximate Gradient Discretization preserving positivity for two-phase Darcy flows in heterogeneous porous media. J. Comput. Phy. (2020). https://doi.org/10.1016/j.jcp.2020.109357
6. Hamon, F.P., Mallison, B.T., Tchelepi, H.A.: Implicit hybrid upwinding for two-phase flow in heterogeneous porous media with buoyancy and capillarity. Comput. Methods Appl. Mech. Eng. **331**, 701–727 (2018)
7. Quenjel, E.H.: Enhanced positive vertex-centered finite volume scheme for anisotropic convection-diffusion equations. ESAIM Math. Model. Numer. Anal. **54**(2), 591–618 (2020)

Design of Coupled Finite Volume Schemes Minimizing the Grid Orientation Effect in Reservoir Simulation

Karine Laurent, Éric Flauraud, Christophe Preux, Quang Huy Tran, and Christophe Berthon

Abstract In this paper, we present and compare two nine-point finite volume schemes to reduce the so-called *grid orientation effect* (GOE) which occurs in the simulation of unstable two phase flow in porous media. The first scheme is a more classical nine-point scheme with one tuning parameter whereas the second one, more original, uses two parameters (one per direction). A numerical test problem testify to the improvement brought by the new scheme.

Keywords Grid orientation effect · Reservoir simulation · Finite volume schemes · Nine-point scheme

MSC (2010) 35Q35 · 65M08 · 76S05

1 Introduction

In oil reservoir simulation, engineers are often faced with a phenomenon called *grid orientation effect* (GOE). This unpleasant effect arises when coupled finite volume schemes are used on structured grids in order to simulate the thrust of a viscous fluid (oil) by a less viscous one (water), which is typical of an injection scenario for enhanced oil recovery. The GOE gives rise to a more or less marked distortion of the computed solution whereas, in particular, the exact solution is radial. As a consequence, the simulation of predicted production of a well also depends on the grid orientation and may not be accurate. Since the 1970s, a wide range of ideas have explored to reduce the GOE. The literature on this problem is so vast that we

K. Laurent · É. Flauraud (✉) · C. Preux · Q. H. Tran
IFP Energies nouvelles, 1 et 4 avenue de Bois-Préau, 92852 Rueil-Malmaison Cedex, France
e-mail: eric.flauraud@ifpen.fr

C. Berthon
Laboratoire Jean Leray, Université de Nantes, 2 rue de la Houssinière, BP 92208, 44322 Nantes Cedex 3, France
e-mail: christophe.berthon@univ-nantes.fr

R. Klöfkorn et al. (eds.), *Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples*, Springer Proceedings in Mathematics & Statistics 323,
https://doi.org/10.1007/978-3-030-43651-3_54

cannot claim to provide an exhaustive review. One of the precursors is Yanosik & McCracken [6], who developed a *nine-point* (9P) scheme obtained by superimposing two five-point (5P) schemes associated with two square grids rotated by $\pi/4$ relative to each other. By involving diagonal neighbors into the stencil, the resulting scheme significantly reduces the GOE over square meshes and met an instant success. Two generalizations of this 9P scheme to rectangular meshes were then proposed in [4] and in [1]. The difference between these two versions lies in the weighting heuristic for the diagonal cells. Since then, the 9P philosophy has been extended to other porous two-phase models, for example to account for dispersion [5]. The objectionable aspect of these works is that the error analysis—whenever available—is only concerned with the pressure, while the quantity of interest is the saturation. In [2], R. Eymard et al. designed another 9P scheme and propose a weighting parameter for the saturation equation discretization over square meshes. This methodology is more satisfactory from the theoretical standpoint. However, since the basic idea is to request that the diffusion matrix of the equivalent equation be invariant by a $\pi/4$-rotation, the extension to rectangular meshes does not seem obvious.

In this paper, we present and compare two 9P schemes on a two-phase flow problem. The first one, defined in Sect. 3.1 and called 9P1s, has a scalar tuning parameter θ that allows several "historical" schemes such as [1, 4, 6] to appear as special cases of a unified framework. The second one, defined in Sect. 3.2 and called 9P2s, has two scalar tuning parameters (θ_x, θ_y), a novelty that we introduce in order to further reduce the GOE. Finally, in Sect. 4 numerical results on radial problem confirm the ability of the 9P2s scheme to further reduce the GOE on rectangular meshes compared to the 9P1s scheme.

2 The Two-Phase Flow Model

Let $\Omega \subset \mathbb{R}^2$ be a bounded open connected domain with a regular boundary. The two-phase flow is characterized by the common pressure $p(\boldsymbol{x}, t) > 0$ and the water saturation $s(\boldsymbol{x}, t) \in [0, 1]$, where $\boldsymbol{x} = (x, y) \in \Omega$ and $t \geq 0$ are respectively space and time variables. These quantities of interest solve

$$\boldsymbol{u} = -\kappa\lambda(s)\nabla p, \tag{1a}$$

$$\operatorname{div}(\boldsymbol{u}) = q, \tag{1b}$$

$$\phi\partial_t s + \operatorname{div}(f(s)\boldsymbol{u}) = q_w, \tag{1c}$$

where the total velocity $\boldsymbol{u}(\boldsymbol{x}, t)$ is given by the Darcy's law (1a), and

$$\lambda(s) = \frac{\kappa_{r,w}(s)}{\mu_w} + \frac{\kappa_{r,o}(1-s)}{\mu_o} \tag{2}$$

is the total mobility. From now on, Eq. (1b) is referred to as the pressure equation, since it gives $-\mathrm{div}(\kappa\lambda(s)\nabla p) = q$ when combined with (1a). The symbol κ stands for the permeability tensor, restricted here to be a scalar. The water relative permeability $\kappa_{r,w}(s)$ is an increasing function of s, while the oil relative permeability $\kappa_{r,o}(1-s)$ is a decreasing function of s. Moreover, the two scalars $\mu_w > 0$ and $\mu_o > 0$ denote the water and oil viscosities. The quantity $\phi(\boldsymbol{x}) \in [0, 1]$ represents the (known) porosity of the medium. Here, without loss of generality, we impose $\phi \equiv 1$.

The water fractional flow $f(s)$ in (1c) is defined as

$$f(s) = \frac{\kappa_{r,w}(s)/\mu_w}{\kappa_{r,w}(s)/\mu_w + \kappa_{r,o}(1-s)/\mu_o}, \tag{3}$$

where we have set $\kappa_{r,w}(s) = \kappa^\sharp_{r,w}\,\kappa^*_{r,w}(s)$ and $\kappa_{r,o}(1-s) = \kappa^\sharp_{r,o}\,\kappa^*_{r,o}(1-s)$. The normalized relative permeabilities $\kappa^*_{r,w}(s)$ and $\kappa^*_{r,o}(1-s)$ are assumed to be in [0, 1], while $\kappa^\sharp_{r,w}$ and $\kappa^\sharp_{r,o}$ are given dimensionless constants. The water fractional flow f is a smooth positive and non-decreasing function of s, i.e., $f \geq 0$ and $f' \geq 0$ for $s \in [0, 1]$. It can be put under the reduced form

$$f(s) = \frac{M\kappa^*_{r,w}(s)}{M\kappa^*_{r,w}(s) + \kappa^*_{r,o}(1-s)}, \qquad \text{where} \qquad M = \frac{\mu_o\kappa^\sharp_{r,w}}{\mu_w\kappa^\sharp_{r,o}} \tag{4}$$

is the mobility ratio between the displacing water and the displaced oil. M measures, in some sense, the stiffness of the problem. Indeed, as soon as M is larger than some critical threshold, the system (1) turns out to be unstable and thus amplifies the numerical errors. In such a context, the errors due to the GOE may become prevailing. In the right-hand sides of (1), q and q_w are source terms expressing the produced or injected total and water flow in the domain.

3 Nine-Point Finite Volume Methods

Usually, system (1) is discretized in time using the IMPES technique where the pressure p is solved implicitly in a first step and the saturation s is solved explicitly (at least for the convection part) in a second step. Adopting a semi discrete formulation in space, the IMPES scheme reads

$$\boldsymbol{u}^{n+1} = -\kappa\lambda(s^n)\nabla p^{n+1}, \tag{5a}$$

$$\mathrm{div}(\boldsymbol{u}^{n+1}) = q^{n+1}, \tag{5b}$$

$$\Delta t^{-1}(s^{n+1} - s^n) + \mathrm{div}(f(s^n)\boldsymbol{u}^{n+1}) = q_w^{n+1}, \tag{5c}$$

where the time-step $\Delta t > 0$ must be restricted by a CFL-like condition. Regarding the discretization in space of the two divergence operators in (5), there are two finite volume schemes, one for the pressure equation (5b) and another one for the saturation equation (5c). In this section, we describe two discretizations in space, namely: (i) in Sect. 3.1, the 9P1s scheme which makes use of one scalar parameter; (ii) in Sect. 3.2, the 9P2s scheme which makes use of two scalar parameters.

The domain Ω is divided into uniform rectangular cells $K_{i,j} = (x_{i-1/2}, x_{i+1/2}) \times (y_{j-1/2}, y_{j+1/2})$ of side lengths $(x_{i+1/2} - x_{i-1/2}, y_{j+1/2} - y_{j-1/2}) = (\Delta x, \Delta y) \in (\mathbb{R}_*^+)^2$. We denote by $\boldsymbol{x}_{i,j} = (x_i, y_j)$ the center of the cell $K_{i,j}$.

3.1 A Nine-Point Scheme with One Parameter θ

The approximation of the Eq. (5b) with a nine-point finite volume scheme gives the discrete flux balance

$$\begin{aligned} &F^{\theta}_{i+1/2,j} - F^{\theta}_{i-1/2,j} + F^{\theta}_{i,j+1/2} - F^{\theta}_{i,j-1/2} \\ &\quad + F^{\theta\nearrow}_{i+1/2,j+1/2} - F^{\theta\nearrow}_{i-1/2,j-1/2} + F^{\theta\nwarrow}_{i-1/2,j+1/2} - F^{\theta\nwarrow}_{i+1/2,j-1/2} = \Delta x \Delta y\, q_{i,j}, \end{aligned} \tag{6}$$

in the cell $K_{i,j}$, where the numerical fluxes are defined by

$$F^{\theta}_{i+1/2,j} = \kappa\widetilde{\lambda}(s^n_{i,j}, s^n_{i+1,j})[z - 2\theta(z + z^{-1})](p^{n+1}_{i,j} - p^{n+1}_{i+1,j}), \tag{7a}$$

$$F^{\theta}_{i,j+1/2} = \kappa\widetilde{\lambda}(s^n_{i,j}, s^n_{i,j+1})[z^{-1} - 2\theta(z + z^{-1})](p^{n+1}_{i,j} - p^{n+1}_{i,j+1}), \tag{7b}$$

$$F^{\theta\nearrow}_{i+1/2,j+1/2} = \kappa\widetilde{\lambda}(s^n_{i,j}, s^n_{i+1,j+1})\theta(z + z^{-1}) \ (p^{n+1}_{i,j} - p^{n+1}_{i+1,j+1}), \tag{7c}$$

$$F^{\theta\nwarrow}_{i-1/2,j+1/2} = \kappa\widetilde{\lambda}(s^n_{i,j}, s^n_{i-1,j+1})\theta(z + z^{-1}) \ (p^{n+1}_{i,j} - p^{n+1}_{i-1,j+1}). \tag{7d}$$

The total mobilities are approximated between two cells using a harmonic average $\widetilde{\lambda}(s_L, s_R) = 2\lambda(s_L)\lambda(s_R)/[\lambda(s_L) + \lambda(s_R)]$. z is the ratio between the mesh sizes $z = \Delta y/\Delta x$. The selected orientation of the eight numerical fluxes (7) is displayed in Fig. 1. The arrows $\nearrow$ and $\nwarrow$ indicate the direction in which the flux takes a positive value. Finally, $q_{i,j}$ is an approximation of the source term q in $K_{i,j}$.

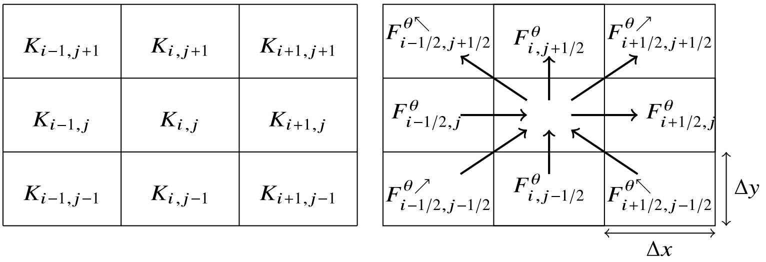


Fig. 1 Nine-point stencil (left) and orientation of numerical fluxes (right)

Once the pressure field is computed, the saturation equation (5c) can be discretized with a scheme having a similar nine-point and eight-flux structure. Rearranged as a discrete balance, the update of saturation takes the form

$$\begin{aligned}\Delta x\,\Delta y\,\Delta t^{-1}(s_{i,j}^{n+1} - s_{i,j}^{n}) + G_{i+1/2,j}^{\theta} - G_{i-1/2,j}^{\theta} + G_{i,j+1/2}^{\theta} - G_{i,j-1/2}^{\theta}\\ + G_{i+1/2,j+1/2}^{\theta\nearrow} - G_{i-1/2,j-1/2}^{\theta\nearrow} + G_{i-1/2,j+1/2}^{\theta\nwarrow} - G_{i+1/2,j-1/2}^{\theta\nwarrow} = \Delta x\,\Delta y\, q_{w;i,j},\end{aligned} \tag{8}$$

where the fluxes are upwinded as

$$G_{i+1/2,j}^{\theta} = f(s_{i,j}^{n})[F_{i+1/2,j}^{\theta}]^{+} \qquad + f(s_{i+1,j}^{n})[F_{i+1/2,j}^{\theta}]^{-}, \tag{9a}$$

$$G_{i+1/2,j+1/2}^{\theta\nearrow} = f(s_{i,j}^{n})[F_{i+1/2,j+1/2}^{\theta\nearrow}]^{+} + f(s_{i+1,j+1}^{n})[F_{i+1/2,j+1/2}^{\theta\nearrow}]^{-}, \tag{9b}$$

where $[F]^{+} = \max(F, 0)$ and $[F]^{-} = \min(F, 0)$ are respectively the positive and negative parts of F. The upwinding for $G_{i,j+1/2}^{\theta}$ and $G_{i-1/2,j+1/2}^{\theta\nwarrow}$ are similar. The term $q_{w;i,j}$ expresses an approximation of the source term.

The crucial point of this scheme is the choice of the parameter θ which determines the accuracy of the method. Some values have been proposed in the literature [1, 4, 6] and in [3] we propose a rigorous approach to deduce an optimal value of θ which decreases as much as possible the anisotropy of the numerical saturation error when the exact solution is radial. This parameter is given by

$$\theta^{\flat} = \frac{1}{4}\left(\frac{\Delta x + \Delta y}{\sqrt{\Delta x^2 + \Delta y^2}} - 1\right). \tag{10}$$

Note that, for a square mesh ($\Delta x = \Delta y = h$), the optimal value degenerates to

$$\theta^{\flat} = \frac{\sqrt{2} - 1}{4} \approx 0.103553, \tag{11}$$

which coincides with the parameter recommended in [2].

3.2 A Nine-Point Scheme with Two Parameters θ_x and θ_y

In this second and new scheme, we introduce two parameters instead of one. After all, since we have two privileged directions x, y, two grid-steps Δx, Δy, it seems natural to have θ_x, θ_y in the definition of the scheme. However, the introduction of these two parameters forces us to modify the definition of the fluxes in the balance equation (6) and in particular the diagonal fluxes. Then the numerical fluxes (7) are now defined as linear combinations of the standard two-point flux approximations used in the 5P scheme. In other words,

$$
\begin{aligned}
F^{\theta}_{i+1/2,j} &= (1-4\theta_x)F_{i+1/2,j}, \qquad F^{\theta}_{i-1/2,j} = (1-4\theta_x)F_{i-1/2,j}, && \text{(12a)}\\
F^{\theta}_{i,j+1/2} &= (1-4\theta_y)F_{i,j+1/2}, \qquad F^{\theta}_{i,j-1/2} = (1-4\theta_y)F_{i,j-1/2}, && \text{(12b)}\\
F^{\theta\nearrow}_{i+1/2,j+1/2} &= \theta_y F_{i,j+1/2} + \theta_x F_{i+1/2,j+1} + \theta_x F_{i+1/2,j} + \theta_y F_{i+1,j+1/2}, && \text{(12c)}\\
F^{\theta\nearrow}_{i-1/2,j-1/2} &= \theta_y F_{i-1,j-1/2} + \theta_x F_{i-1/2,j} + \theta_x F_{i-1/2,j-1} + \theta_y F_{i,j-1/2}, && \text{(12d)}\\
F^{\theta\nwarrow}_{i-1/2,j+1/2} &= \theta_y F_{i,j+1/2} - \theta_x F_{i-1/2,j+1} - \theta_x F_{i-1/2,j} + \theta_y F_{i-1,j+1/2}, && \text{(12e)}\\
F^{\theta\nwarrow}_{i+1/2,j-1/2} &= \theta_y F_{i+1,j-1/2} - \theta_x F_{i+1/2,j} - \theta_x F_{i+1/2,j-1} + \theta_y F_{i,j-1/2}, && \text{(12f)}
\end{aligned}
$$

where θ is now defined as the vector $(\theta_x,\ \theta_y)$ and

$$
\begin{aligned}
F_{i+1/2,j} &= \kappa\widetilde{\lambda}(s^n_{i,j}, s^n_{i+1,j})\, z\, (p^{n+1}_{i,j} - p^{n+1}_{i+1,j}),\\
F_{i,j+1/2} &= \kappa\widetilde{\lambda}(s^n_{i,j}, s^n_{i,j+1})\, z^{-1}\, (p^{n+1}_{i,j} - p^{n+1}_{i,j+1}).
\end{aligned}
$$

Once the pressure field is computed, the saturations are deduced from the Eq. (8) with the upwinded fluxes (9) in which (7) are replaced by (12). Once again in [3], an analysis of the saturation error is carried out in order to define two optimal parameters θ_x and θ_y in the sense that they minimize the anisotropy of the error when the solution is radial. The optimal values of these parameters are given by

$$
\theta^\flat_x(z,\omega) = \frac{\sqrt{1+\omega^2}(z\omega^2+1) - (1+z\omega^3)}{8z\omega}, \qquad \theta^\flat_y(z,\omega) = \frac{z\theta^\flat_x(z,\omega)}{\omega}, \tag{13}
$$

where

$$
\omega(z) = \begin{cases} 7z/2 & \text{if } 0 \le z \le 2/7,\\ 1 & \text{if } 2/7 \le z \le 7/2,\\ 2z/7 & \text{otherwise.} \end{cases}
$$

For a square mesh ($\Delta x = \Delta y$), we have $z = \omega = 1$ and recover $\theta^\flat_x = \theta^\flat_y = \theta^\flat = \frac{\sqrt{2}-1}{4}$.

4 Numerical Results

The numerical test presented in this chapter is inspired by [2]. The problem models a water injector well placed at the center of a homogeneous domain $\Omega = [-0.5, 0.5]^2$ initially saturated with oil. Consider the system

$$
\begin{aligned}
\boldsymbol{u} &= -\lambda(s)\nabla p, && \text{(14a)}\\
\partial_t s + \operatorname{div}(f(s)\boldsymbol{u}) &= \delta_0, && \text{(14b)}\\
\operatorname{div}(\boldsymbol{u}) &= \delta_0, && \text{(14c)}
\end{aligned}
$$

in $\Omega \times [0, T]$, $T = 0.05$, with the initial data $s(\boldsymbol{x}, t = 0) = 0$ in Ω. In (14), $q = q_w = \delta_0$ are Dirac sources expressing water injection at $\boldsymbol{x} = \mathbf{0}$. The absolute permeability has been assigned the constant value $\kappa = 1$, while the relative permeabilities are

$$\kappa_{r,w}(s) = s^2 \qquad \text{and} \qquad \kappa_{r,o}(1-s) = (1-s)^2. \tag{15}$$

As a consequence, the water fractional flux is

$$f(s) = \frac{Ms^2}{Ms^2 + (1-s)^2}, \qquad \text{with} \qquad M = \frac{\mu_o}{\mu_w}. \tag{16}$$

Setting $\mu_o = 200$ and $\mu_w = 1$ results in $M = 200$, which is a highly unfavorable mobility ratio. The system (14) is completed with the following boundary condition

$$-\lambda(s)\nabla p \cdot \boldsymbol{n} = \frac{1}{2\pi r}\boldsymbol{e}_r \cdot \boldsymbol{n}, \tag{17}$$

where $\boldsymbol{n}$ denotes the unit outward normal vector of $\partial\Omega$. System (14) with the boundary condition (17) has an analytical radial solution (see [2, 3]).

In Figs. 2 and 3, we plot the isovalues of the saturation and the profiles of the saturation along the x-axis, y-axis and the $diagonal$-axis for the 5P scheme, the 9P1s scheme and the 9P2s scheme. The simulations are run on two uniform grids: a 201 × 201 square mesh (Fig. 2) and a 201 × 601 rectangular mesh (Fig. 3). We note that for the square mesh (Fig. 2), the two 9P schemes suppress the GOE which on observes with the 5P scheme. Besides, the saturation profiles obtained with the 9P1s and 9P2s schemes are similar and close to the analytical solution which is not surprising since both schemes use the same optimal value for $\theta^\flat$ (11). However, for the rectangular mesh, the numerical solution with the 9P1s scheme is more diffused in the x-direction and becomes oval while it remains more radial with the 9P2s scheme. Indeed, the saturation profiles obtained with the 9P2s scheme are closed to the analytical solution in the three directions. Thus, this simple test shows the improvement brought by the new scheme to better mitigate the GOE on rectangular meshes.

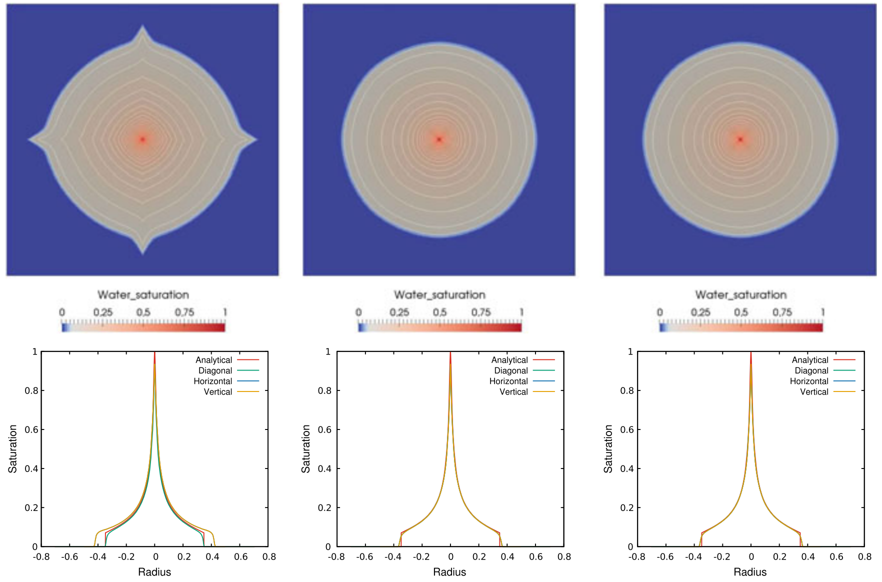


Fig. 2 Saturation contours (top) and saturation profiles (bottom) on square mesh ($\Delta x = \Delta y$) for the 5P scheme (left), 9P1s scheme (middle) and 9P2s scheme (right)

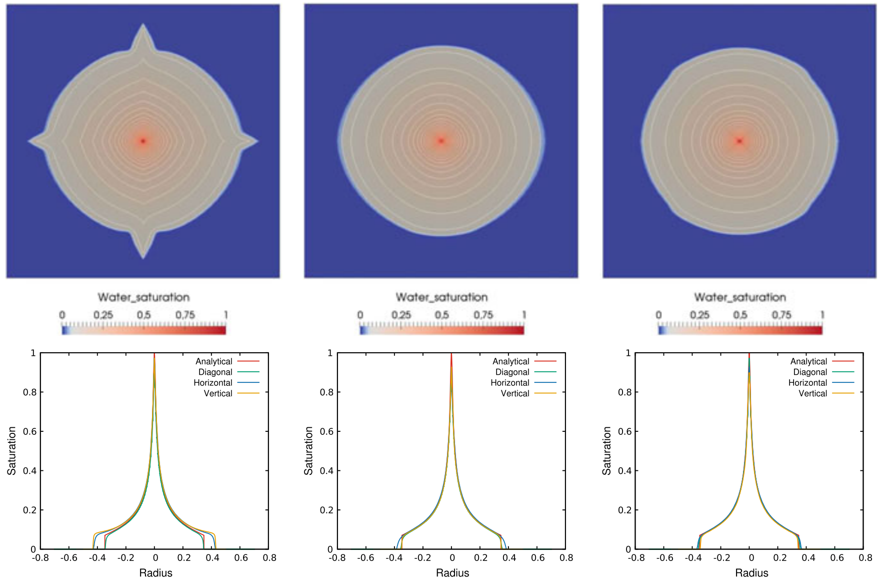


Fig. 3 Saturation contours (top) and saturation profiles (bottom) on rectangular mesh ($\Delta x = 3\Delta y$) for the 5P scheme (left), 9P1s scheme (middle) and 9P2s scheme (right)

References

1. Coats, K.H., Modine, A.D.: A consistent method for calculating transmissibilities in nine-point difference equations. In: SPE Reservoir Simulation Symposium, San Francisco, California, 15–18 Nov 1983
2. Eymard, R., Guichard, C., Masson, R.: Grid orientation effect in coupled finite volume schemes. IMA J. Numer. Anal. **33**, 582–608 (2013)
3. Laurent, K., Flauraud, E., Preux, C., Tran, Q.H., Berthon, C.: Design of coupled finite volume schemes minimizing the grid orientation effect in reservoir simulation (2019). https://hal.archives-ouvertes.fr/hal-02387696
4. Shah, P.C.: A nine-point finite difference operator for reduction of the grid orientation effect. SPE, 171–174 (1983)
5. Shiralkar, Gautam S., Stephenson, Robert E.: A general formulation for simulating physical dispersion and a new nine-point scheme. SPE Reserv. Eng. **6**, 115–120 (1991)
6. Yanosik, J.L., McCracken, T.A.: A nine-point, finite-difference reservoir simulator for realistic prediction of adverse mobility ratio displacements. SPE J. **19**, 253–262 (1979)

A Comparison of Consistent Discretizations for Elliptic Problems on Polyhedral Grids

Øystein S. Klemetsdal, Olav Møyner, Xavier Raynaud, and Knut-Andreas Lie

Abstract In this work, we review a set of consistent discretizations for second-order elliptic equations, and compare and contrast them with respect to accuracy, monotonicity, and factors affecting their computational cost (degrees of freedom, sparsity, and condition numbers). Our comparisons include the linear and nonlinear TPFA method, multipoint flux-approximation (MPFA-O), mimetic methods, and virtual element methods. We focus on incompressible flow and study the effects of deformed cell geometries and anisotropic permeability.

Keywords Reservoir simulation · Elliptic discretizations · Finite volumes · TPFA · NTPFA · MPFA · MFD · VEM

MSC (2010) 76S05 · 35L65 · 65M08 · 35J15

1 Introduction

Models of petroleum reservoirs with complex geology tend to have grids with general hexahedral or polyhedral cell geometries and tensor permeabilities. The standard two-point flux-approximation (TPFA) method is only consistent for K-orthogonal grids in which the principal directions of the permeability tensor align with vectors joining cell and face centroids.[1] Simulation models are often the result of upscaling [9], which tends to generate nonzero off-diagonal permeabilities, and as a rule, simulation grids will not be K-orthogonal, at least in some parts of the reservoir. The TPFA method is then not consistent and convergent, and will introduce grid-orientation effects that adversely affect the accuracy. Much research has therefore been devoted to develop consistent methods on non-K-orthogonal grids.

[1] Other choices of primary pressure points are also possible, e.g., circumcenter for triangular grids.

Ø. S. Klemetsdal (✉) · O. Møyner · X. Raynaud · K.-A. Lie
SINTEF Digital, Oslo, Norway
e-mail: oystein.klemetsdal@sintef.no

R. Klöfkorn et al. (eds.), *Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples*, Springer Proceedings in Mathematics & Statistics 323,
https://doi.org/10.1007/978-3-030-43651-3_55

The multipoint flux-approximation (MPFA) scheme [1] accounts for transversal pressure variations by introducing auxiliary pressure points at the cell interfaces, which are coupled inside local interaction regions that together form a dual grid. MPFA methods retain the same low number of unknowns as TPFA, but have a larger stencil and can be somewhat cumbersome to implement for complex grids.

Mimetic methods [5] also introduce auxiliary pressure points to ensure consistency, which are kept as primary unknowns. An inherent free stabilization parameter gives a variety of specific schemes that reduce to other known discretizations on simple grids [15]. The main drawbacks of mimetic methods are that they use a mixed-hybrid formulation and involve significantly more unknowns than cell-centered methods. Mimetic methods have later been developed into virtual element methods (VEM) [2, 3], which constitute a uniform and flexible framework for higher-order discretizations on general polyhedral cells. MPFA, mimetic, and VEM are only conditionally monotone and may introduce nonphysical pressure oscillations. The nonlinear two-point scheme (NTPFA) [17, 19, 20] uses pressure-dependent transmissibilities to define a consistent *and monotone* method, but requires the solution of a nonlinear system of equations.

In this work, we compare the performance of these methods applied to the type of grid models encountered in real reservoir simulation using the open-source MRST software [14]. Our test cases involve deformed cell geometries and anisotropic permeabilities. The paper can therefore be seen as an update of [15] and [12]. Further comparisons can be found in, e.g., [8].

2 Consistent Discretizations on Polyhedral Grids

For simplicity, we consider incompressible single-phase flow,

$$\nabla \cdot \mathbf{v} = q, \quad \mathbf{v} = -\mathbf{K}\nabla p, \quad \mathbf{x} \in \Omega \subset \mathbb{R}^d. \tag{1}$$

Discretized by a mesh consisting of n_c polygonal or polyhedral cells Ω_i with constant permeability $\mathbf{K}_i$ on each, the control-volume formulation of (1) reads

$$\int\limits_{\partial\Omega_i} \mathbf{v} \cdot \mathbf{n}\, \mathrm{d}s = \int_{\Omega_i} q\, \mathrm{d}\mathbf{x} = q_i. \tag{2}$$

Methods differ in the way they approximate the flux across intercell faces. Consider two neighboring cells as in Fig. 1, with common interface Γ_{ij}. The normal vector $\mathbf{n}_{i,j}$ points from Ω_i to Ω_j, and similarly, $\mathbf{n}_{j,i} = -\mathbf{n}_{i,j}$. For the flux $v_{i,j}$ across Γ_{ij} in the direction of $\mathbf{n}_{i,j}$, local conservation of mass requires $v_{i,j} = -v_{j,i}$.

Discrete conservation of mass is a natural requirement, and also necessary to avoid nonphysical solutions in multiphase simulations; consistency is needed for a correct solution, typically used together with coercivity to prove convergence [5, 20]; whereas monotonicity is desirable to produce physically meaningful solutions with

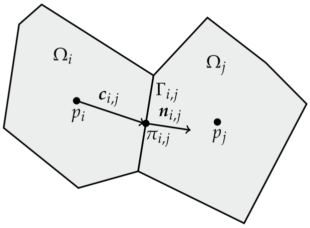


Fig. 1 Two neighboring cells and geometric quantities used to discretize the flux $v_{i,j}$

properties inherent to elliptic problems [10]. For linear discretizations, a sufficient condition for monotonicity is that the discretization produces a so-called M-matrix. Lack of coercivity may nonetheless lead to convergence breakdown, even for consistent methods [6]. We present a set of consistent discretization methods for (1), and discuss some of these properties; see Fig. 3 for a schematic comparison.

2.1 Two-Point Flux-Approximation

With an auxiliary pressure point $\pi_{i,j}$ at the centroid of Γ_{ij}, we can use a one-sided finite-difference to approximate the pressure gradient in Darcy's law,

$$v_{i,j} = \int\limits_{\Gamma_{i,j}} \mathbf{v} \cdot \mathbf{n}_{i,j} \, \mathrm{d}s \approx |\Gamma_{i,j}| \frac{\mathbf{c}_{i,j}^T \mathbf{K}_i \mathbf{n}_{i,j}}{|\mathbf{c}_{i,j}|^2} (p_i - \pi_{i,j}) = T_{i,j}(p_i - \pi_{i,j}). \tag{3}$$

Here, $\mathbf{K}_i$ is the constant value of $\mathbf{K}$ on Ω_i, and $T_{i,j}$ is referred to as the one-sided transmissibility. Imposing flux continuity across interfaces, $v_{ij} = v_{i,j} = -v_{j,i}$, and continuity of face pressures, $\pi_{ij} = \pi_{i,j} = \pi_{j,i}$, gives the system

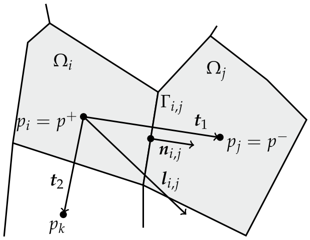
$$\sum\nolimits_{j=1}^{n_c} T_{ij}(p_i - p_j) = q_i, \qquad T_{ij} = \left(T_{i,j}^{-1} + T_{j,i}^{-1}\right)^{-1}, \qquad i = 1, \dots, n_c, \tag{4}$$

where T_{ij} is the transmissibility. If these cells do not share an interface, the transmissibility T_{ij} is zero. This yields an M-matrix, which guarantees that the method is monotone. However, the TPFA method is only consistent for K-orthogonal grids, for which a sufficient condition is that $\mathbf{K}_i\mathbf{n}_{i,j}$ is parallel to $\mathbf{c}_{i,j}$ for all cells.

2.2 Multipoint Flux Approximation

For a consistent method, one must account for pressure gradients parallel to cell faces. MFPA-O constructs an interaction region around each grid node and defines linear basis functions for pressure inside, with pressure continuity at face centroids and flux continuity across face patches. Continuity and mass conservation gives a

Fig. 2 The NTPFA vector $\mathbf{l}_{i,j} = \mathbf{K}_i\mathbf{n}_{i,j}$



consistent method, with unknown cell pressures and face pressures along the *outer* boundary. This gives a denser linear system than for TPFA. However, the method is only monotone under specific conditions, and we can not expect it to be monotone for very skewed grid cells and/or severely anisotropic permeabilities [10]. See, e.g., [1, 7] for more details of MPFA schemes.

2.3 Nonlinear Two-Point Flux Approximation

The NTPFA method [17, 19, 21] also uses additional points to estimate fluxes,

$$v_{i,j} = T_{i,j}(\mathbf{p})p_i - T_{j,i}(\mathbf{p})p_j.$$

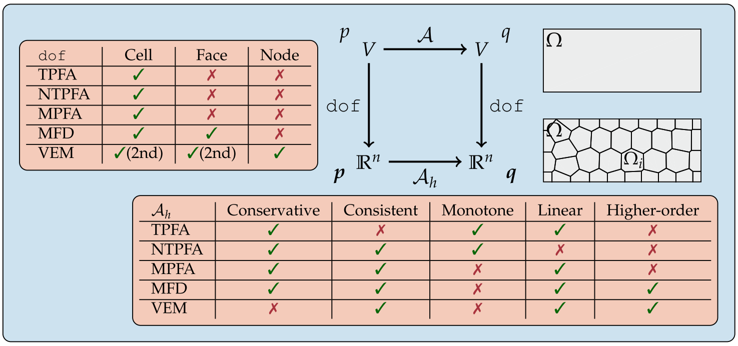
The transmissibilities $T_{i,j}$ are positive functions that depend on one or more pressure values, giving a nonlinear method. To derive such a scheme, we consider the vector $\mathbf{l}_{i,j} = \mathbf{K}_i\mathbf{n}_{i,j}$, (see Fig. 2). Whereas TPFA approximates $\mathbf{l}_{i,j}$ using only vector components normal to the interface, NTPFA uses a decomposition onto a basis of d vectors in d spatial dimensions. This is used to obtain consistent discretizations of $v_{i,j}$ and $v_{j,i}$, with v_{ij} taken as a convex combination of these. The result is a consistent and monotone two-point flux approximation, where the transmissibilities depend on pressure values not included in the two-point stencil.

2.4 Mimetic Finite Differences

TPFA and MPFA-O can be seen as special cases of a wider family of mass-conservative schemes written in so-called *hybrid formulation*

$$\mathbf{v}_i = \mathbf{T}_i(\mathbf{e}_i p_i - \pi_i), \quad \text{in } \Omega_i.$$

Here, $\mathbf{v}_i$ is the vector of fluxes across the n_f cell faces, $\mathbf{e}_i = (1, \ldots, 1)^T \in \mathbb{R}^{n_f}$, π_i is the vector of face pressures, and $\mathbf{T}_i$ is a matrix of one-sided transmissibilities. Discrete mass conservation and flux continuity is imposed through separate



dof	Cell	Face	Node
TPFA	✓	✗	✗
NTPFA	✓	✗	✗
MPFA	✓	✗	✗
MFD	✓	✓	✗
VEM	✓(2nd)	✓(2nd)	✓

$$\begin{tikzcd} V \arrow[r, "\mathcal{A}"] \arrow[d, "\texttt{dof}"'] & V \arrow[d, "\texttt{dof}"] \\ \mathbb{R}^n \arrow[r, "\mathcal{A}_h"'] & \mathbb{R}^n \end{tikzcd}$$

$\mathcal{A}_h$	Conservative	Consistent	Monotone	Linear	Higher-order
TPFA	✓	✗	✓	✓	✗
NTPFA	✓	✓	✓	✗	✗
MPFA	✓	✓	✗	✓	✗
MFD	✓	✓	✗	✓	✓
VEM	✗	✓	✗	✓	✓

Fig. 3 Schematic overview of key properties of the methods compared in this paper

equations, see e.g., [15] for details. This formulation can be interpreted as a first-order mimetic finite difference method, where different choices of the inner product matrices $\mathbf{M}_i = \mathbf{T}_i^{-1}$ lead to different special cases (e.g., TPFA or MPFA-O, see [15, 16]).

2.5 The Virtual Element Method

In their present formulation, neither of the methods mentioned so far are easily extended to higher order. By using moments of the solution as degrees of freedom, it is possible to obtain a unified, higher-order framework for general polyhedral grids called the virtual element method (VEM) [2, 3]. Herein, we use this method as an example of a finite element-type discretization for polyhedral grids. This formulation is not locally conservative, and the result must be postprocessed in order to be used in transport simulations. Alternatively, it is possible to use a mixed formulation [4]. We will denote first- and second-order VEM by VEM1 and VEM2, respectively.

3 Numerical Experiments

All discretizations are implemented in MRST [14]. Full codes for the following two examples are available online,[2] and [11] gives a more elaborate description.

[2] https://bitbucket.org/strene/compare-elliptic/

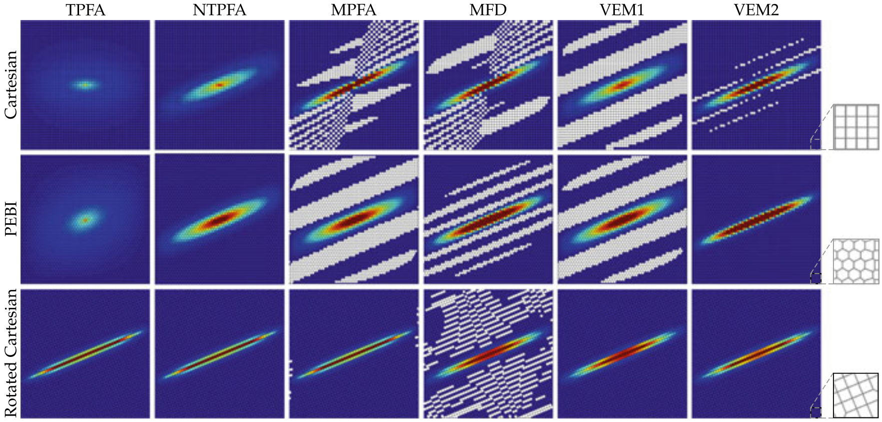


Fig. 4 Pressure solutions for the monotonicity test; white cells indicate negative pressure

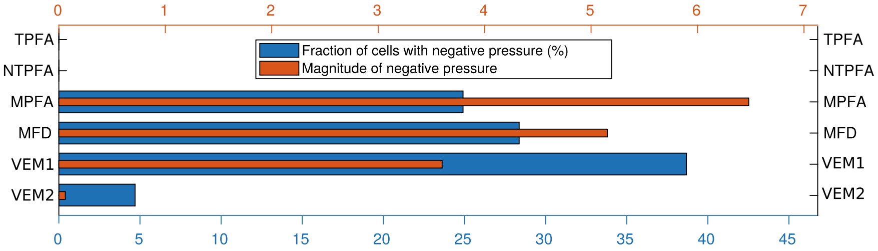


Fig. 5 Fraction and magnitude of negative pressure values for the solutions in the monotonicity test on the Cartesian mesh. Magnitude of negative pressure = $100 \sum_{p_i<0} |p_i| \Big/ \sum_i |p_i|$

3.1 Monotonicity

The fundamental elliptic maximum principle of (1) implies that if there is a single source within the domain, the pressure will decrease monotonically towards the boundary. To assess deviations from monotonicity, we consider anisotropic permeability $K_x/K_y = 500$, rotated by an angle $\pi/8$, and three different meshes: 51 × 51 Cartesian, honeycombed PEBI, and a rotated Cartesian mesh aligned with the principal axes of $\mathbf{K}$. We place a point source at the origin, and impose zero pressure boundary conditions. Figure 4 reports approximate solutions. All consistent methods, except NTPFA, give oscillations along the minor principal axis of $\mathbf{K}$. Figure 5 reports fraction and magnitude of negative pressures values.

NTPFA and TPFA are monotone by construction and have no cells with negative pressure. VEM1 has the highest fraction of negative pressures, but the magnitude is lower than for MFD and MPFA. MPFA has the highest magnitude of negative

Table 1 Key characteristics of the discrete systems for the monotonicity example: number of primary unknowns (dof), number of nonzero entries in discretization matrix (nnz), average number of nonzero entries per unknown (ratio = nnz/dof), and condition number (cond)

	Points for unknowns	Calculation	dof	nnz	Ratio	cond
TPFA	Cells	$51 \cdot 51$	2601	12801	4.92	1.45e+03
NTPFA	Cells	$51 \cdot 51$	2601	17208	6.62	2.83e+03
MPFA	Cells + outer faces	$51^2 + 2 \cdot 4 \cdot 51$	3009	23209	7.71	1.69e+03
MFD	Faces	$2 \cdot 51 \cdot 52 - 4 \cdot 51$	5100	35096	6.88	7.01e+03
VEM1	Vertices	$52 \cdot 52$	2704	22704	8.40	5.08e+04
VEM2	Cells + faces + vertices	$52^2 + 51^2 + 2 \cdot 51 \cdot 52$	10609	162817	15.35	1.07e+06

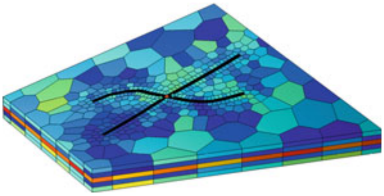


Fig. 6 Near-well case: **K** is log-normal, $K_x/K_y = 3$, rotated $\pi/6$ in the xy-plane. Fractures in black, topmost well-cell in red at the fracture intersection

pressures. VEM2 yields far better results in terms of physically meaningful pressure fields, even though the method is not guaranteed to be monotone.

Table 1 reports characteristics of the linear systems on the Cartesian mesh. MPFA has almost twice as many nonzero entries per unknown as TPFA, but similar condition number. NTPFA has a sparsity pattern similar to MPFA. MFD is less dense than for MFPA, but has three times higher condition number. VEM1 has fewer unknowns than MPFA, denser stencil, and condition number $\mathcal{O}(10)$ larger than the other. VEM2's stencil is more than three times denser than TPFA, with $\mathcal{O}(10^3)$ larger condition number. Results are similar on the rotated mesh. On the PEBI mesh, MFD and VEM are significantly denser, in particular MFD and VEM2, because the PEBI mesh has 1.5 as many faces as the Cartesian.

3.2 Near-Well Simulation

Grid blocks in real field models usually represent upscaled volumes containing significant permeability variation. We consider a near-well region with a vertical well, modelled as a source injecting 1 PV over 0.1 yrs. Two fractures intersect the well, modelled as volumetric objects with a much higher permeability (Fig. 6). Constant pressure is imposed on the vertical sides, with no-flow top/bottom. All the consistent schemes predict similar outflow through the four vertical sides, with VEM1 deviating most (7%) from the other four. TPFA differs with as much as 20%, which is reason for serious concern, if used for upscaling.

Table 2 Key characteristics of the discrete systems for the near-well example

	dof	nnz	Ratio	cond
TPFA	2465	19809	8.04	1.11e+04
NTPFA	2465	33608	13.63	3.26e+05
MPFA	8507	98579	11.59	6.74e+04
MFD	9658	130438	13.51	2.94e+09
VEM1	5274	170618	32.35	2.22e+11
VEM2	30173	2495409	82.70	1.44e+12

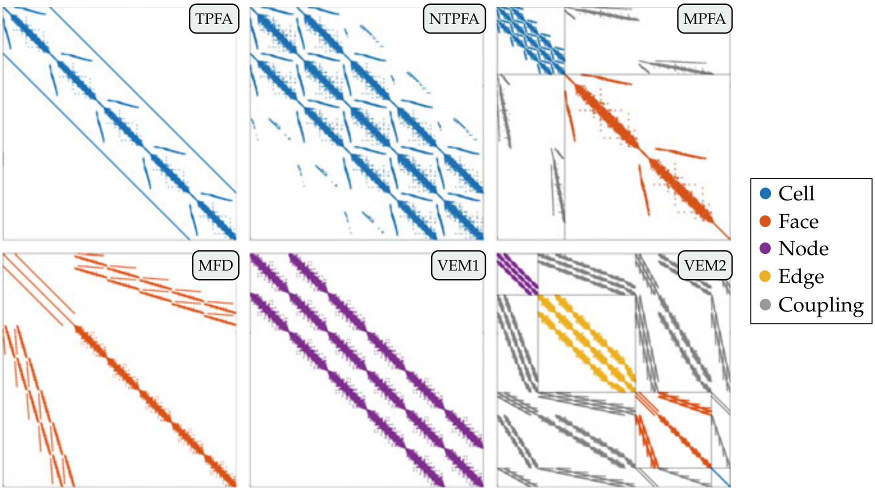


Fig. 7 Sparsity patterns from the near-well example, with different colors for each type of dof

Table 2 confirms that differences in algebraic complexity are accentuated compared to the 2D cases. VEM is *very* dense, with VEM2 having a ratio of 82.7. All methods have significantly higher condition numbers, with MFD and VEM being more ill-conditioned than the other methods. Figure 7 reports sparsity patterns. Since TPFA is not consistent on this grid, the converged NTPFA discretization is similar to that of MPFA instead of TPFA. VEM2 has a face-pressure block equal to MFD, and a node-pressure block equal to that of VEM1.

4 Closing Remarks

The novelty herein is that we compare a large set of discretizations on the same problem, with access to complete source codes. Our experiments here and in [12] do not clearly point to one preferred method that is significantly better than the others.

TPFA is inconsistent and has grid orientation effects, but is monotone and gives sparse matrices with low condition numbers. Consistent methods are convergent and reduce grid orientation effects, but have monotonicity issues and give denser and more ill-conditioned linear systems, particularly for VEM. NTPFA is monotone but requires the solution of a nonlinear system and is, in our experience, significantly less robust than e.g., mimetic methods. Our best advice is to compute representative flow solutions with more than one consistent scheme and use the results to estimate the level of error that may arise because of anisotropic permeability and skew and irregular cell geometries. For multiphase simulations, one should assess the quality of the resulting flow fields using e.g., flow diagnostics [14]: sweep, drainage, and well-pair regions, well-allocation factors, time-of-flight, and residence time distributions. In addition, one should also investigate the number and size of the connected components in the computed flux fields, as these will affect convergence behavior of nonlinear solvers used in each time step of a multiphase simulation [13, 18].

Acknowledgements Klemetsdal, Lie, and Raynaud were supported by the Research Council of Norway (244361). Møyner is funded by VISTA, a basic research programme funded by Equinor and conducted in close collaboration with The Norwegian Academy of Science and Letters.

References

1. Aavatsmark, I.: An introduction to multipoint flux approximations for quadrilateral grids. Comput. Geosci. **6**(3–4), 405–432 (2002)
2. Ahmad, B., Alsaedi, A., Brezzi, F., Marini, L.D., Russo, A.: Equivalent projectors for virtual element methods. Comput. Math. Appl. **66**(3), 376–391 (2013)
3. Beirão da Veiga, L., et al.: Basic principles of virtual element methods. Math. Model. Methods Appl. Sci. **23**(01), 199–214 (2013)
4. Brezzi, F., Falk, R.S., Donatella Marini, L.: Basic principles of mixed virtual element methods. ESAIM Math. Model. Numer. Anal. **48**(4), 1227–1240 (2014)
5. Brezzi, F., Lipnikov, K., Shashkov, M.: Convergence of the mimetic finite difference method for diffusion problems on polyhedral meshes. SIAM J. Numer. Anal. **43**(5), 1872–1896 (2005)
6. Droniou, J.: Finite volume schemes for diffusion equations: introduction to and review of modern methods. Math. Mod. Methods Appl. Sci. **24**(08), 1575–1619 (2014)
7. Edwards, M.G., Rogers, C.F.: A flux continuous scheme for the full tensor pressure equation. In: ECMOR IV-4th European Conference on the Mathematics of Oil Recovery (1994)
8. Eymard, R., et al.: 3D benchmark on discretization schemes for anisotropic diffusion problems on general grids. In: Fořt, J., et al. (eds.) Finite Volumes for Complex Applications VI Problems & Perspectives, pp. 895–930. Springer, Berlin, Heidelberg (2011)
9. Farmer, C.: Upscaling: a review. Int. J. Numer. Methods Fluids **40**(1–2), 63–78 (2002)
10. Keilegavlen, E., Aavatsmark, I.: Monotonicity for MPFA methods on triangular grids. Comput. Geosci. **15**(1), 3–16 (2011)
11. Klemetsdal, Ø.S.: Efficient solvers for field-scale simulation of flow and transport in porous media. Ph.D. thesis, Norwegian University of Science and Technology (2019)
12. Klemetsdal, Ø.S., et al.: Unstructured gridding and consistent discretizations for reservoirs with faults and complex wells. In: SPE Reservoir Simulation Conference (2017)
13. Klemetsdal, Ø.S., et al.: Efficient reordered nonlinear Gauss-Seidel solvers with higher order for black-oil models. Comput. Geosci. (2019)

14. Lie, K.A.: An Introduction to Reservoir Simulation Using MATLAB/GNU Octave. Cambridge University Press (2019)
15. Lie, K.A., et al.: Open-source MATLAB implementation of consistent discretisations on complex grids. Comput. Geosci. **16**(2), 297–322 (2012)
16. Lipnikov, K., Shashkov, M., Yotov, I.: Local flux mimetic finite difference methods. Numer. Math. **112**(1), 115–152 (2009)
17. Lipnikov, K., et al.: Monotone finite volume schemes for diffusion equations on unstructured triangular and shape-regular polygonal meshes. J. Comput. Phys. **227**(1), 492–512 (2007)
18. Natvig, J.R., et al.: An efficient discontinuous Galerkin method for advective transport in porous media. Adv. Water Resour. **30**(12), 2424–2438 (2007)
19. Nikitin, K., Terekhov, K., Vassilevski, Y.: A monotone nonlinear finite volume method for diffusion equations and multiphase flows. Comput. Geosci. **18**(3–4), 311–324 (2014)
20. Schneider, M., et al.: Convergence of nonlinear finite volume schemes for heterogeneous anisotropic diffusion on general meshes. J. Comput. Phys. **351**, 80–107 (2017)
21. Schneider, M., et al.: Monotone nonlinear finite-volume method for challenging grids. Comput. Geosci. **22**(2), 565–586 (2018)

Global Implicit Solver for Multiphase Multicomponent Flow in Porous Media with Multiple Gas Phases and General Reactions

Markus M. Knodel, Serge Kräutle, and Peter Knabner

Abstract Multiphase multicomponent flow processes in porous media have to be considered to study the efficiency of mineral trapping mechanisms for climate killing gas storage in deep layers. Robust predictions ask for the solution of large nonlinear coupled systems of diffusion-advection-reaction (partial differential) equations containing equilibrium reactions. In that we elaborate the fully globally implicit Kräutle-Knabner PDE reduction method (cf. a former paper Kräutle and Knabner in Water Resour Res 43(3):W03429 [8]) for the case of multiple gas phases, we solve the arising Finite Element discretized/Finite Volume stabilized equations by means of a semismooth nested Newton solver. We present preliminary simulation results for the case of mutual injection of CO_2, CH_4 and H_2S into deep layers and investigate the arising mineral trapping scenario. Our methods are applicable also to other fields such as nuclear waste storage or oil recovery.

Keywords PDE reduction method · Nested Newton · Equilibrium reactions · CO_2

MSC (2010) 86-08 · 35A25 · 35Q68 · 76V05 · 80A32 · 92E20 · 76T10 · 35K57

1 Introduction

To study the efficiency of trapping mechanisms (mineral trapping) for CO_2 storage in deep layers, multiphase multicomponent flow processes in porous media [1, 2] have to be considered. The precise prediction of gas and liquid flow asks for the

M. M. Knodel (✉) · S. Kräutle · P. Knabner
Chair of Applied Mathematics 1, Universität Erlangen-Nürnberg,
Cauerstr. 11, 91058 Erlangen, Germany
e-mail: markus.knodel@math.fau.de

S. Kräutle
e-mail: kraeutle@math.fau.de

P. Knabner
e-mail: knabner@math.fau.de

R. Klöfkorn et al. (eds.), *Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples*, Springer Proceedings in Mathematics & Statistics 323,
https://doi.org/10.1007/978-3-030-43651-3_56

solution of large nonlinear coupled systems of diffusion-advection-reaction partial differential equations (PDEs), algebraic equations (AEs) and ordinary differential equations (ODEs). The choice of a suitable formulation of the equations is important for efficient numerical solution. We apply the fully globally implicit Kräutle-Knabner PDE reduction method ("KKPRM", published in former papers [5, 8]) which enables to eliminate the equilibrium reactions based upon specific variable transformations. Separating the resulting remaining PDE/ODE/AE system into a global and a local system, we apply a semismooth nested Newton solver method [7] which enables fast and efficient computation of the dynamics of the system by means of the application of parallel solvers to the Finite Element discretized/Finite Volume stabilized [3] PDE system. Our computations of the behavior of the concentrations of the different species of the multiphase multicomponent flow are highly resolved in space and time. We extend the mineral trapping scenario of the predecessor of this study, the "Brunner/Knabner-paper" (BKP) [4] to the case of an arbitrary number of species in gaseous phase. Namely, we present first results for the case of the injection of various gas species. Whereas our results are so far preliminary, our techniques allow for predictions in very complex scenarios and are applicable also for similar cases such as nuclear waste storage and oil recovery. As a side effect, we are on the way to compute the present Sin benchmark [9].

2 Mathematical Model and Global Implicit Solver

If we denote $\alpha = g, \ell, s$ as the gaseous, liquid and solid phase, the PDEs describing multicomponent multiphase flow in porous media (with porosity ϕ) with general reactions, i.e. the reactive transport, read for I_α concentrations $\mathbf{c}_\alpha = (\mathbf{c}_\alpha^1, \ldots, \mathbf{c}_\alpha^{I_\alpha})$

$$\partial_t(\phi s_\alpha \mathbf{c}_\alpha^i) + \nabla \cdot (\mathbf{q}_\alpha \mathbf{c}_\alpha^i + \mathbf{j}_\alpha^i) = \mathbf{f}_\alpha^i, \qquad i = 1, 2, \ldots, I_\alpha. \tag{1}$$

For gas and liquid phase, we have saturation s_α, Darcy velocity $\mathbf{q}_\alpha$, diffusive flux $\mathbf{j}_\alpha$. For all states (g, ℓ, s), the right hand side $\mathbf{f}_\alpha$ describes kinetic and equilibrium reactions. We define the transport operator $\mathscr{L}_\alpha$ (with mole fractions $\chi_\alpha = \mathbf{c}_\alpha / \rho_\alpha^{mol}$)

$$\mathscr{L}_\alpha \mathbf{c}_\alpha = \nabla \cdot \left(-\rho_\alpha^{mol} \mathbf{D}_\alpha \nabla \boldsymbol{\chi}_\alpha + \mathbf{q}_\alpha \mathbf{c}_\alpha\right) \tag{2}$$

for gaseous and liquid phase which also comprises the diffusion-dispersion tensor $\mathbf{D}_\alpha$. For detailed formulae, parameter and further variable definitions, we refer to the BKP [4]. The complete system of equations does not consist only of the PDEs (1), but in addition of ODEs and AEs. These ODEs and AEs result from the equations of state (EOS) for the gaseous and liquid phase, the equilibrium reactions, and relations between different components. The phenomena described by means of AEs are:

- The equilibrium reactions (which are vector-valued, also the coefficients):
  - We have exchange $g \leftrightarrow \ell$ (gas–liquid). Henry's law is a realization:

$$\boldsymbol{\Phi}_{eq}^{ex} := \mathbf{p}_g - \mathbf{H}\,\boldsymbol{\chi}_\ell \overset{!}{=} \mathbf{0} \qquad (\text{Henry constant } \mathbf{H}, \text{ partial gas pressure } \mathbf{p}_g). \tag{3}$$

  (**H**: diagonal matrix.) For CO_2 exchange, other laws such as the experimentally based spline description derived by Spycher-Pruess [10] should be used.
  - The law of mass action (LaMA) is applied for all reactions in and between liquid and solid phase, with vector of equilibrium constants **K** and stoichiometric matrix **S**: Aquatic reactions between different aqueous species $\ell \leftrightarrow \ell$ $\boldsymbol{\Phi}_{mob}$, sorption reactions between aqueous species and nonmineral solids: $\ell \leftrightarrow s^{\cancel{min}}$: $\boldsymbol{\Phi}_{sorp}$, and mineral reactions between aqueous species and mineral solids $\ell \leftrightarrow s^{min}$ $\boldsymbol{\Phi}_{min}$, hence (details: BKP, note: $\ln(\mathbf{c}_\ell) = (\ln(\mathbf{c}_\ell^1), \ldots, \ln(\mathbf{c}_\ell^{I_\ell})))$:

$$\begin{aligned} \boldsymbol{\Phi}_{eq}^{\ell s} &= (\boldsymbol{\Phi}_{mob}, \boldsymbol{\Phi}_{sorp}, \boldsymbol{\Phi}_{min}) \\ \text{e.g.} \qquad \boldsymbol{\Phi}_{mob} &:= (\mathbf{S}^T)_\ell^{mob} \ln(\mathbf{c}_\ell) - \ln(\mathbf{K}^{mob}) \overset{!}{=} \mathbf{0}\,. \end{aligned} \tag{4}$$

- The EOS for gas: The most simple version is the ideal gas law, Peng-Robinson is more involved already. Experimentally based splines such as the EOS of Duan-Moeller (strictly speaking: only pure CO_2 injection) are the most realistic ones.
- For the EOS of the liquid, in "simple" cases such as the Sin benchmark [9], constant molar mass density of the liquid phase can be used. However, more involved experimentally based spline descriptions exist, namely Garcias law if we consider only CO_2 exchange. (All our EOS are scalar-valued.)

We may write the final equation system in compact form[1]

$$\begin{aligned} \partial_t(\phi s_g\, \mathbf{c}_g) + \mathscr{L}_g\, \mathbf{c}_g &= \phi s_\ell \mathbf{S}_g^{ex}\, \mathbf{R}_{eq} && (5) \\ \partial_t(\phi s_\ell\, \mathbf{c}_\ell) + \mathscr{L}_\ell\, \mathbf{c}_\ell &= \phi s_\ell\, \mathbf{S}_\ell^{kin}\, \mathbf{R}_{kin} + \phi s_\ell \mathbf{S}_\ell^{eq}\, \mathbf{R}_{eq} && (6) \\ \partial_t(\phi s_\ell\, \mathbf{c}_s) &= \phi s_\ell\, \mathbf{S}_s^{kin}\, \mathbf{R}_{kin} + \phi s_\ell \mathbf{S}_s^{eq}\, \mathbf{R}_{eq} && (7) \\ \boldsymbol{\Phi}_{eq}^{ex}(\mathbf{c}_g, \mathbf{c}_\ell) &= \mathbf{0} && (8) \\ \boldsymbol{\Phi}_{eq}^{\ell s}(\mathbf{c}_\ell, \mathbf{c}_s) &= \mathbf{0} && (9) \\ \rho_\alpha^{mol} - f_\alpha(\mathbf{c}_\alpha) &= 0 \qquad (\alpha = g, \ell \text{ for the EOS}). && (10) \end{aligned}$$

The aim of the KKPRM [5, 8] is to remove the equilibrium reactions and thus reduce the number of PDEs to enhance efficiency and accuracy. We construct orthogonal matrices (basis of complete space, $\mathbf{S}_\alpha^\star = \mathbf{S}_\alpha\, \mathbf{A}_\alpha$: linear independent part of $\mathbf{S}_\alpha$, the linear dependent parts are shifted to the $\mathbf{A}_\alpha = ((\mathbf{S}_\alpha^*)^T(\mathbf{S}_\alpha^*))(\mathbf{S}_\alpha^*)^T(\mathbf{S}_\alpha))$ [6]:

$$(\mathbf{S}_\alpha^\perp)^T\, \mathbf{S}_\alpha^* = \mathbf{0} \qquad (\mathbf{B}_\alpha^\perp)^T\, \mathbf{B}_\alpha = \mathbf{0} \qquad (\text{standard choice: } \mathbf{B}_\alpha = \mathbf{S}_\alpha^*). \tag{11}$$

[1] Besides the relations for molar and mass densities and (also capillary) pressures in g and ℓ phase.

Multiplying (5)–(7) with $((\mathbf{S}_\alpha^\perp)^T\,\mathbf{B}_\alpha^\perp)^{-1}(\mathbf{S}_\alpha^\perp)^T$ resp. $(\mathbf{B}_\alpha^T\,\mathbf{S}_\alpha^*)^{-1}\,\mathbf{B}_\alpha^T$ and transforming the concentrations to reaction invariant η and reaction participating $\boldsymbol{\xi}$ components

$$\boldsymbol{\eta}_\alpha = ((\mathbf{S}_\alpha^\perp)^T\,\mathbf{B}_\alpha^\perp)^{-1}(\mathbf{S}_\alpha^\perp)^T\,\mathbf{c}_\alpha \qquad\qquad \boldsymbol{\xi}_\alpha = (\mathbf{B}_\alpha^T\,\mathbf{S}_\alpha^*)^{-1}\,\mathbf{B}_\alpha^T\,\mathbf{c}_\alpha \tag{12}$$

transfers the PDEs (several matrix combinations collapse) to the structure (excerpt):

$$\begin{array}{rcll}
\partial_t(\phi s_g \boldsymbol{\xi}_g^{ex}) + \mathscr{L}_g \boldsymbol{\xi}_g^{ex} &=& \phi s_\ell\,\mathbf{R}_{eq}^{ex} & \quad \rceil^{-1} \\
\partial_t(\phi s_\ell \boldsymbol{\eta}_\ell) + \mathscr{L}_\ell \boldsymbol{\eta}_\ell &=& \mathbf{0} & \\
\partial_t\left(\phi s_\ell \boldsymbol{\xi}_\ell^{ex}\right) + \mathscr{L}_\ell \boldsymbol{\xi}_\ell^{ex} &=& \phi s_\ell\,\mathbf{R}_{eq}^{ex} + \phi s_\ell\,\mathbf{A}_\ell^{ex}\,\mathbf{R}^{kin}\,. & \quad \leftarrow\!\rfloor^{+}
\end{array} \tag{13}$$

The directly afore indicated subtractions in (13) are the core of the KKPRM. The *subtractionprocedureleadstoastronglyreducednumberofPDEs*. The sub-system (13) reduces to (the remaining equations for the unknowns $\mathbf{R}_{eq}^{ex}$ can be dropped):

$$\partial_t(\phi s_\ell \boldsymbol{\eta}_\ell) + \mathscr{L}_\ell \boldsymbol{\eta}_\ell = \mathbf{0} \tag{14}$$

$$\partial_t\left(\phi s_\ell \boldsymbol{\xi}_\ell^{ex}\right) - \partial_t(\phi s_g \boldsymbol{\xi}_g^{ex}) + \mathscr{L}_\ell \boldsymbol{\xi}_\ell^{ex} - \mathscr{L}_g \boldsymbol{\xi}_g^{ex} = \phi s_\ell\,\mathbf{A}_\ell^{ex}\,\mathbf{R}^{kin}\,. \tag{15}$$

Widening the BKP algorithm, we split the remaining equation system and the variables into local and global parts with the *global system* and the *global variables*

$$\begin{array}{rcl}
\rho_\ell^{mol} - f_\ell(\mathbf{c}_\ell) &=& 0 \\
\boldsymbol{\Phi}_{eq}^{ex}(\mathbf{c}_g, \mathbf{c}_\ell) &=& \mathbf{0} \\
\partial_t(\phi s_g\, \boldsymbol{\eta}_g) + \mathscr{L}_g\, \boldsymbol{\eta}_g &=& \mathbf{0} \\
\partial_t(\phi s_\ell\, \boldsymbol{\eta}_s) \quad &=& \mathbf{0} \\
\partial_t(\phi s_\ell\, \boldsymbol{\eta}_\ell) + \mathscr{L}_\ell\, \boldsymbol{\eta}_\ell &=& \mathbf{0} \\
\partial_t\left(\phi s_\ell\, \boldsymbol{\xi}_\ell^{ex} + \phi s_g \hat{\boldsymbol{\xi}}_g^{ex}\right) && \\
+\mathscr{L}_\ell\, \boldsymbol{\xi}_\ell^{ex} + \mathscr{L}_g \hat{\boldsymbol{\xi}}_g^{ex} &=& \phi s_\ell\,\mathbf{A}_\ell^{ex}\,\mathbf{R}^{kin} \\
\partial_t\left(\phi s_\ell \bar{\boldsymbol{\xi}}_B^{sorp}\right) + \mathscr{L}_\ell\, \boldsymbol{\xi}_\ell^{sorp} &=& \phi s_\ell\,\mathbf{A}_{\ell-s}^{sorp}\,\mathbf{R}^{kin} \\
\partial_t\left(\phi s_\ell \bar{\boldsymbol{\xi}}_B^{min}\right) + \mathscr{L}_\ell\, \boldsymbol{\xi}_\ell^{min} &=& \phi s_\ell\,\mathbf{A}_\ell^{min}\,\mathbf{R}^{kin} \\
\partial_t\left(\phi s_\ell\, \boldsymbol{\xi}_\ell^{kin}\right) + \mathscr{L}_\ell\, \boldsymbol{\xi}_\ell^{kin} &=& \phi s_\ell\,\mathbf{A}_\ell^{kin}\,\mathbf{R}^{kin} \\
\bar{\boldsymbol{\xi}}_B^{sorp} - \boldsymbol{\xi}_\ell^{sorp} + \boldsymbol{\xi}_s^{sorp} &=& 0 \\
\bar{\boldsymbol{\xi}}_B^{min} - \boldsymbol{\xi}_\ell^{min} + \boldsymbol{\xi}_s^{min} &=& 0.
\end{array}
\qquad
\boldsymbol{\Xi}_{glob} = \begin{pmatrix}
\tilde{p}_c \\
\tilde{\mathbf{p}}_g^{partial} \\
\boldsymbol{\eta}_g \\
\boldsymbol{\eta}_s \\
\boldsymbol{\eta}_\ell \\
\boldsymbol{\xi}_\ell^{ex} \\
\boldsymbol{\xi}_\ell^{sorp} \\
\boldsymbol{\xi}_\ell^{min} \\
\boldsymbol{\xi}_\ell^{kin} \\
\bar{\boldsymbol{\xi}}^{sorp} \\
\bar{\boldsymbol{\xi}}^{min}
\end{pmatrix} \tag{16}$$

The global system consists out of the PDEs, the liquid EOS, and the gas exchange equilibrium reactions (Henry's law and/or Spycher-Pruess). New (not part of BKP) equations are indicated in red, and we use $\hat{\boldsymbol{\xi}}_{ex}^{g} = -\boldsymbol{\xi}_{ex}^{g}$ for comparison purposes with the BKP. The local system contains the law ϕ^{cap} for the capillary pressure p_c, various relations between the variables, and the LaMA based equilibrium reactions for liquid-solid interactions. The *local system* and *local variables* read

$$
\begin{aligned}
\phi^{cap}(s_g, p_c) &= 0 \\
\tilde{p}_g^{total} - \sum_{i=0}^{I_g} \tilde{p}_g^i &= 0 \\
\rho_g^{mol} - f_g(\mathbf{c}_g) &= 0 \\
\mathbf{p}_g^{partial}\, \rho_g^{mol} - \mathbf{c}_g\, p_g^{total} &= \mathbf{0} \\
\tilde{p}_c - \tilde{p}_g^{total} + p_\ell &= 0 \\
\boldsymbol{\Phi}^{mob}(\mathbf{c}_\ell) &= \mathbf{0} \\
\boldsymbol{\Phi}^{sorp}(\mathbf{c}_\ell, \mathbf{c}_s^{nm}) &= \mathbf{0} \\
\boldsymbol{\Phi}^{min}(\mathbf{c}_\ell, \mathbf{c}_s^{min}) &= \mathbf{0} \\
\partial_t \left(\phi s_\ell\, \boldsymbol{\xi}_s^{kin}\right) &= \phi s_\ell\, \mathbf{A}_s^{kin}\, \mathbf{R}^{kin} \\
\rho_{g,\ell}^{mass} - \sum_{i=1}^{I_{g,\ell}} M^i c_{g,\ell}^i &= 0 \\
\rho_\ell^{mol} - \sum_{i=1}^{I_\ell} c_\ell^i &= 0.
\end{aligned}
\qquad
\boldsymbol{\Xi}_{loc} = \begin{pmatrix} s_g \\ \tilde{p}_g^{total} \\ \rho_g^{mol} \\ \hat{\boldsymbol{\xi}}_g^{ex} \\ p_\ell \\ \boldsymbol{\xi}_\ell^{mob} \\ \boldsymbol{\xi}_s^{sorp} \\ \boldsymbol{\xi}_s^{min} \\ \boldsymbol{\xi}_s^{kin} \\ \rho_{g,\ell}^{mass} \\ \rho_\ell^{mol} \end{pmatrix}
\tag{17}
$$

We apply a semismooth nested Newton solver [7] to solve the coupled system of local and global equations and variables (cf. former papers [4, 5, 8]). The Newton iterator for the local problem (at each grid point) is nested in the Newton iterator for the global problem. The resolution function $\frac{\partial\, \boldsymbol{\Xi}^{loc}}{\partial\, \boldsymbol{\Xi}^{glob}}$ is indispensable and relates global and local variables based upon the relation of global and local system. Finally, we note the retransformation to get the entire physical concentrations

$$
\mathbf{c}_g = -\mathbf{S}_g^{ex}\, \hat{\boldsymbol{\xi}}_g^{ex} + \mathbf{B}_g^{\perp}\, \boldsymbol{\eta}_g \tag{18}
$$

$$
\begin{aligned}
\mathbf{c}_\ell = {} & \mathbf{S}_\ell^{ex}\, \boldsymbol{\xi}_\ell^{ex} + \mathbf{S}_\ell^{mob}\, \boldsymbol{\xi}_\ell^{mob} + \mathbf{S}_\ell^{sorp}(\bar{\boldsymbol{\xi}}_B^{sorp} + \boldsymbol{\xi}_s^{sorp}) \\
& + \mathbf{S}_\ell^{min}(\bar{\boldsymbol{\xi}}_B^{min} + \boldsymbol{\xi}_s^{min}) + (\mathbf{S}^*)_\ell^{kin}\, \boldsymbol{\xi}_\ell^{kin} + \mathbf{B}_\ell^{\perp}\, \boldsymbol{\eta}_\ell
\end{aligned}
\tag{19}
$$

$$
\mathbf{c}_s = \begin{pmatrix} \mathbf{c}^{\xcancel{min}} \\ \mathbf{c}_s^{min} \end{pmatrix} = \begin{pmatrix} \mathbf{S}_s^{sorp}\, \boldsymbol{\xi}_s^{sorp} + (\mathbf{S}^*)_s^{kin}\, \boldsymbol{\xi}_s^{kin} + \mathbf{B}_s^{\perp}\, \boldsymbol{\eta}_s \\ \boldsymbol{\xi}_s^{min} \end{pmatrix} \tag{20}
$$

where $\mathbf{c}_s^{\xcancel{min}}$ indicates the nonmineral solids and $\mathbf{c}_s^{min}$ the mineral ones.

Caring for code-reusabilty, we implemented the afore explained algorithm into our parallel M++ [11] based RICHY++ framework. The discretization is performed by means of adaptive implicit Euler in time and Finite Elements in space where we use Finite Volume stabilization for the advective parts [3]. To solve the equation system arising from the global Newton, we use BiCGStab as LinearSolver with a SuperLU preconditioner. The simulation presented in this study was performed at a 600 × 100 m sized 2D computational domain constructed by means of rectangles with size 20 × 20m (squares). The number of Degrees of Freedom (DoFs) reads:

	DoFs	
Level	global	local
0	2,790	2,976
1	10,065	10,736
2	38,115	40,656
3	148,215	158,096

3 Simulations/Results

We display preliminary results for the relevant case of mineral trapping for the injection of various gas species into deep layers.[2] Our case mixes elements of the BKP [4] with elements from the Sin benchmark [9]. We take the chemistry of BKP, but we add two additional gases. Most values which were used in the BKP [4] are still used. For the new parameters, we use in major part those given by the Sin benchmark description [9] and in part, we use heuristic values for parameters.

$$CO_2^g \longleftrightarrow CO_2^\ell \tag{21}$$

$$CH_4^g \longleftrightarrow CH_4^\ell \tag{22}$$

$$H_2S^g \longleftrightarrow H_2S^\ell \tag{23}$$

$$CO_2^\ell + H_2O \longleftrightarrow HCO_3^- + H^+ \tag{24}$$

$$H_2S^\ell \longleftrightarrow H^+ + HS^- \tag{25}$$

$$\text{Calcite} + H^+ \longleftrightarrow Ca^{2+} + HCO_3^- \tag{26}$$

$$\text{MinA} + 3H^+ \longleftrightarrow Me^{3+} + SiO_2 \tag{27}$$

$$\text{MinB} + 2H^+ \longleftrightarrow Me^{3+} + HCO_3^-. \tag{28}$$

The blue lines in Eqs. (21)–(28) arise from the BKP case, the red ones from the Sin benchmark. All 3 gases are injected into the porous domain (note that the Sin benchmark does not contain gas injection). We use Spycher-Pruess for CO_2 exchange and Henry's law for CH_4 and H_2S exchange. Even though we use in part still heuristic parameters and still use the spline-based Garcias law EOS for the fluid which only

[2] Note that the terminus "gas injection" in our context always means that the gas is injected in purely dissolute form, but switches into equilibrium state within each time step.

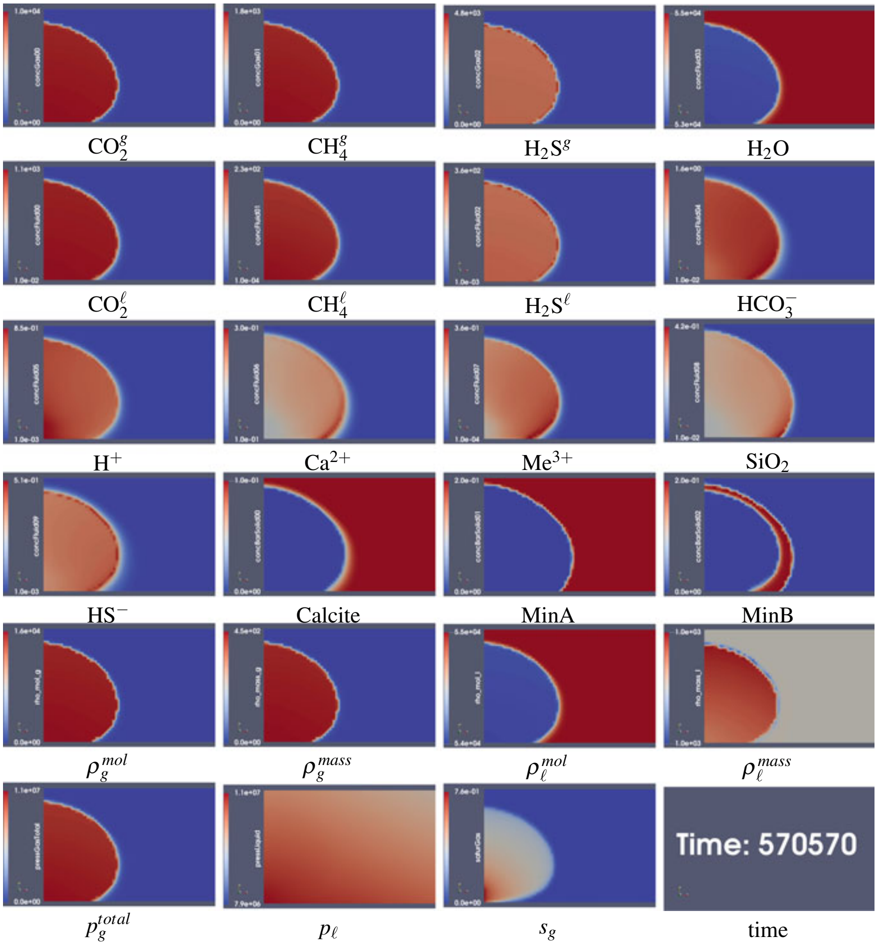


Fig. 1 "Merging" of the BKP with elements from the Sin benchmark: three gas injection (CO_2, CH_4, H_2S) into deep layers and mineral trapping. Spatial refinement level 3 (only a part of 600 × 100 m domain visible). Boundary conditions: Neumann flux $x = 0$ m, $z \leq 12.5$ m; Dirichlet (initial values) $x = 600$ m; else no flux. Screenshot: $t = 5.7 \cdot 10^5$ s, corresponding to about nine days

considers the CO_2 exchange (instead of the even simpler constant liquid density fluid EOS as applied by the Sin benchmark), and we also still apply the spline based Duan-Moeller gas EOS rather than ideal gas law or Peng-Robinson, our simulations are already close to the real case of gas injection into deep layers. Figure 1 displays a screenshot of the corresponding simulation.

4 Summary, Conclusions and Outlook

We are working on multiphase multicomponent flow in porous media with general equilibrium and kinetic reactions. To evaluate the highly complex equations, we apply the globally implicit solver for multiphase multicomponent flow based on the refined KKPRM [8]. We presented the extension of the BKP [4] from one to various gas species which get injected into deep layers leading to mineral trapping. We presented first results with various gases, gas liquid exchange, and mineral trapping. Such computations are at the high end of present technology and therefore, our present example already demonstrates the efficiency of our approach. We intend to apply our approach also to the Sin Benchmark [9], which does not contain gas injection, but more complex chemistry. To this end, we plan extended studies of our code concerning also numerical grid convergence and weak and strong scaling.

Even though so far in part we still use heuristic parameters, our approach likely is one of the very first successful application of globally implicit solvers for the case of the injection and reactive transport of multiple gas phases into deep layers of porous media inducing mineral trapping and we are approaching highly realistic scenarios.

Acknowledgements M. M. K. thanks Etienne Ahusborde, Université de Pau et des Pays de l'Adour, France, and Florian Frank, Universität Erlangen, for very stimulating discussions on the subject.

References

1. Ahusborde, E., Amaziane, B., El Ossmani, M.: Improvement of numerical approximation of coupled multiphase multicomponent flow with reactive geochemical transport in porous media. Oil Gas Sci. Technol. Revue d'IFP Energies nouvelles **73**, 73 (2018)
2. Becker, B., Guo, B., Bandilla, K., Celia, M.A., Flemisch, B., Helmig, R.: An adaptive multiphysics model coupling vertical equilibrium and full multidimensions for multiphase flow in porous media. Water Resour. Res. **54**(7), 4347–4360 (2018)
3. Brunner, F., Frank, F., Knabner, P.: FV upwind stabilization of FE discretizations for advection–diffusion problems. In: J. Fuhrmann, M. Ohlberger, C. Rohde (eds.) Finite Volumes for Complex Applications VII-Methods and Theoretical Aspects, pp. 177–185. Springer International Publishing (2014)
4. Brunner, F., Knabner, P.: A global implicit solver for miscible reactive multiphase multicomponent flow in porous media. Comput. Geosci. **23**(1), 127–148 (2019)
5. Hoffmann, J., Kräutle, S., Knabner, P.: A parallel global-implicit 2-D solver for reactive transport problems in porous media based on a reduction scheme and its application to the MoMaS benchmark problem. Comput. Geosci. **14**(3), 421–433 (2010)
6. Hoffmann, J., Kräutle, S., Knabner, P.: A general reduction scheme for reactive transport in porous media. Comput. Geosci. **16**(4), 1081–1099 (2012)
7. Kräutle, S.: The semismooth newton method for multicomponent reactive transport with minerals. Adv. Water Resour. **34**(1), 137–151 (2011)
8. Kräutle, S., Knabner, P.: A reduction scheme for coupled multicomponent transport-reaction problems in porous media: generalization to problems with heterogeneous equilibrium reactions. Water Resour. Res. **43**(3), W03429 (2007)

9. Sin, I., Lagneau, V., Windt, L.D., Corvisier, J.: 2D simulation of natural gas reservoir by two-phase multicomponent reactive flow and transport—description of a benchmarking exercise. Math. Comput. Simul. **137**, 431–447 (2017)
10. Spycher, N., Pruess, K.: CO_2–H_2O mixtures in the geological sequestration of CO_2. ii. partitioning in chloride brines at 12100 °C and up to 600 bar. Geochimica et Cosmochimica Acta **69**(13), 3309 – 3320 (2005)
11. Wieners, C.: A geometric data structure for parallel finite elements and the application to multigrid methods with block smoothing. Comput. Visual. Sci. **13**(4), 161–175 (2010)

Partitioned Coupling Schemes for Free-Flow and Porous-Media Applications with Sharp Interfaces

Alexander Jaust, Kilian Weishaupt, Miriam Mehl, and Bernd Flemisch

Abstract We investigate a partitioned coupling scheme applied to a system of free flow over a porous medium. The coupling scheme follows a partitioned approach which means that the flow fields in the two domains are solved separately and information is exchanged over the sharp interface that separates the free-flow and the porous-medium domain. Technically, the coupling is realized via the open-source library preCICE, employing a pure black-box approach such that different solver frameworks can be used with highly specialized solvers in each of the flow domains. We investigate the partitioned coupling approach numerically by comparing it to a monolithic coupling scheme with respect to convergence and accuracy. This is the first time a partitioned black-box coupling is used for coupling free flow and porous-media flow. The coupling approach is numerically validated and different partitioned coupling approaches are compared with each other.

Keywords Partitioned coupling · Finite-volume methods · Darcy flow · Stokes equation · preCICE · DuMu^x

MSC (2010) 76S05 · 76M12 · 76D07 · 65N08

A. Jaust (✉) · M. Mehl
Institute for Parallel and Distributed Systems, University of Stuttgart,
Universitätsstraße 38, 70569 Stuttgart, Germany
e-mail: alexander.jaust@ipvs.uni-stuttgart.de

M. Mehl
e-mail: miriam.mehl@ipvs.uni-stuttgart.de

K. Weishaupt · B. Flemisch
Institute for Modelling Hydraulic and Environmental Systems,
University of Stuttgart, Pfaffenwaldring 61, 70569 Stuttgart, Germany
e-mail: kilian.weishaupt@iws.uni-stuttgart.de

B. Flemisch
e-mail: bernd.flemisch@iws.uni-stuttgart.de

R. Klöfkorn et al. (eds.), *Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples*, Springer Proceedings in Mathematics & Statistics 323,
https://doi.org/10.1007/978-3-030-43651-3_57

1 Introduction

Many real-world applications involve flow in different media such as the flow of water over a river bed or the flow of air over fabric for drying. These cases represent coupled multiphysics systems with free flow in one part and flow through a porous medium (soil or fabric) in the other part, potentially with multiple fluid phases.

We want to simulate these kinds of problems numerically. Solving the equations monolithically usually requires direct linear solvers or the development of specialized preconditioners due to the poor condition number of the system matrix [1]. However, this can be very restrictive due to time and memory requirements and prevents the reuse of solvers that are already highly specialized for one of the flow regimes. Thus, partitioned methods that solve the flows in both domains based on domain decomposition ideas are a popular alternative. Partitioned approaches for porous-media applications have been extensively employed and analyzed for finite element discretizations, see for example [3, 6, 7] and the references therein.

In this work, we investigate a partitioned coupling procedure that is not limited to a particular type of numerical discretization for the flow problems, but uses a "black-box" approach. The flow problems in each of the domains are solved separately. Data are exchanged between the two domains and post-processed in a suitable way until the underlying fixed-point problem is solved accurately enough. This procedure is based on ideas that are especially popular for fluid-structure interaction [5]. The respective interface quasi-Newton methods have not gotten much attention for porous-media applications yet.

For integrating the partitioned coupling into our simulation framework DuMu$^{\mathrm{x}}$ [8, 10], we use the open-source software library preCICE [2]. preCICE allows for an easy integration of the coupling procedure into existing codes such that these codes can be easily reused. An overview over the capabilities of preCICE can be found in [2]. This is the first time a partitioned black-box coupling is used for coupling free flow and porous-media flow. The coupling approach is validated and different partitioned coupling approaches are compared with each other.

The structure of this paper is as follows: First, we shortly introduce the governing equations and the coupling conditions on the domain interface. Afterwards, we briefly describe the spatial discretization methods used for each subdomain which is followed by the presentation of the coupling procedure and numerics. We present numerical results for a simple stationary test case and end with a conclusion and outlook.

2 Problem Description

We solve a flow problem on the domain Ω with suitable boundary conditions on the domain boundary $\Gamma = \partial\Omega$. This domain can be split into two subdomains Ω_{ff} and Ω_{pm} with $\Omega = \Omega_{\mathrm{ff}} \cup \Omega_{\mathrm{pm}}$, see Fig. 1. In domain Ω_{ff}, free flow is considered while

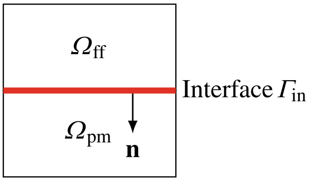


Fig. 1 An example of the domain Ω split into the subdomains with free flow Ω_{ff} and porous-media flow Ω_{pm}. The sharp interface Γ_{in} separating the subdomains is highlighted in red

domain Ω_{pm} corresponds to a porous medium. The two subdomains are separated by a sharp interface $\Gamma_{\text{in}} = \Omega_{\text{ff}} \cap \Omega_{\text{pm}}$. Suitable coupling conditions have to be employed at the interface Γ_{in}.

We are solving for velocities $\mathbf{u}$ and the pressure p in Ω_{ff} while p is the only primary variable in Ω_{pm}. The problems are stationary and incompressible. The influence of gravity can be neglected. Problem dependent parameters are the viscosity μ and the permeability tensor $\mathbf{K}$. We consider an isotropic and homogeneous porous medium. The normal vector $\mathbf{n}$ on the interface Γ_{in} points from the free-flow domain into the porous-medium domain and $\mathbf{t}$ is a unit tangential vector at the interface. Subscripts pm (porous medium) and ff (free flow) are added where necessary.

2.1 Governing Equations

We solve the incompressible Stokes equations,

$$\nabla \cdot \mathbf{u}_{\text{ff}} = 0, \tag{1a}$$

$$(\mathbf{u}_{\text{ff}} \cdot \nabla)\mathbf{u}_{\text{ff}} + \frac{\mu}{\varrho} \Delta \mathbf{u}_{\text{ff}} = -\frac{1}{\varrho} \nabla p_{\text{ff}}, \tag{1b}$$

in the free-flow domain Ω_{ff} for modeling mass (1a) and momentum conservation (1b). ϱ is the fluid density.

In the porous-medium domain Ω_{pm}, we solve

$$\nabla \cdot \mathbf{u}_{\text{pm}} = 0, \tag{2a}$$

$$-\frac{\mathbf{K}}{\mu} \nabla p_{\text{pm}} = \mathbf{u}_{\text{pm}}, \quad \text{with } \mathbf{K} = \begin{pmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{pmatrix} = \begin{pmatrix} K & 0 \\ 0 & K \end{pmatrix}, \tag{2b}$$

which describes again mass conservation (2a) and momentum conservation modeled by Darcy's law (2b). Note that (2a) and (2b) can be summarized to be $-\nabla \cdot \frac{\mathbf{K}}{\mu} \nabla p_{\text{pm}} = 0$, i.e., $\mathbf{u}_{\text{pm}}$ is actually not a primary, but an auxiliary variable.

2.2 Coupling Conditions

In order to ensure consistency, we need to enforce coupling conditions on the interface Γ_{in}. We do so by applying the following conditions:

$$\mathbf{u}_{\text{ff}} \cdot \mathbf{n} = \mathbf{u}_{\text{pm}} \cdot \mathbf{n}, \tag{3a}$$

$$\mathbf{n} \cdot \left(\varrho \mathbf{u}_{\text{ff}} {\mathbf{u}_{\text{ff}}}^T - \mu \left(\nabla \mathbf{u}_{\text{ff}} + (\nabla \mathbf{u}_{\text{ff}})^T\right)\right) \cdot \mathbf{n} + p_{\text{ff}} = p_{\text{pm}}, \tag{3b}$$

$$\left(-\frac{\sqrt{K_{11}}}{\alpha_{\text{BJS}}}(\nabla \mathbf{u}_{\text{ff}} + (\nabla \mathbf{u}_{\text{ff}})^T)\mathbf{n} - \mathbf{u}_{\text{ff}}\right) \cdot \mathbf{t} = 0. \tag{3c}$$

The equations describe the conservation of mass (3a) and the balance of normal forces (3b). The last condition is the Beavers–Joseph–Saffman slip condition [11] which contains the problem-dependent parameter $\alpha_{\text{BJS}} > 0$. Note that the last condition is in fact a boundary condition for the free-flow domain.

3 Solvers and Partitioned Setup

We use standard finite volume methods for the spatial discretization of the flow problems. In the free-flow domain, the inherently stable staggered grid approach [9] is used while in the porous-medium domain, a standard cell-centered finite volume method is used. We solve the subproblems in both domains using DuMu$^{\text{x}}$ [8, 10], an open-source toolbox for flow and transport in porous media.

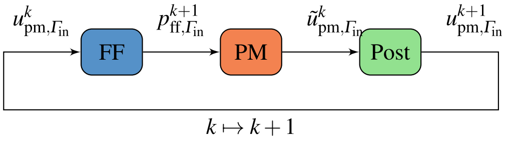
For the partitioned coupling scheme, we solve the flow problems on the two subdomains via two different solvers. The solver referred to as FF solves the free-flow problem on Ω_{ff} and the second solver referred to as PM solves the porous-medium problem on Ω_{pm}. We iterate between the two solvers until convergence of the interface values. More details about this procedure can be found in [2, 5] and the references therein.

Let k be the iteration index. Given the normal velocity $u^k_{\text{pm},\Gamma_{\text{in}}}$ at the interface Γ_{in} and the boundary condition (3c), the free-flow solver computes a new flow state that leads to a new pressure $p^{k+1}_{\text{ff},\Gamma_{\text{in}}}$ on the interface. In the same way, the porous-medium solver computes a new flow state leading to a new normal velocity $u^{k+1}_{\text{pm},\Gamma_{\text{in}}}$ on the interface based on the pressure $p^{k+1}_{\text{ff},\Gamma_{\text{in}}}$ specified on the interface with this setting, the interface coupling conditions can be interpreted as a fixed-point problem

$$\text{PM}(\text{FF}(u_{\text{pm},\Gamma_{\text{in}}})) = u_{\text{pm},\Gamma_{\text{in}}} \quad \Leftrightarrow \quad R(u_{\text{pm},\Gamma_{\text{in}}}) := \text{PM}(\text{FF}(u_{\text{pm},\Gamma_{\text{in}}})) - u_{\text{pm},\Gamma_{\text{in}}} = 0 \tag{4}$$

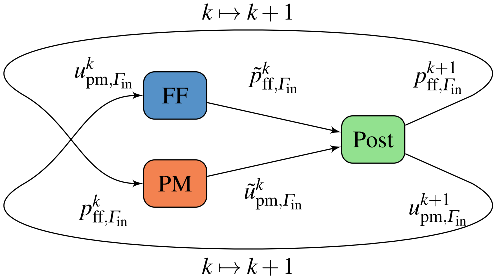
that should recover the monolithic solution when the residual R is zero.

We use (i) serial-implicit couplings, see Fig. 2, and (ii) a parallel-implicit coupling, see Fig. 3. In the serial coupling the solvers compute a new solution sequentially while for the parallel coupling the two solvers compute a new solution concurrently. Both


$u^k_{\mathrm{pm},\Gamma_\mathrm{in}}$
FF
$p^{k+1}_{\mathrm{ff},\Gamma_\mathrm{in}}$
PM
$\tilde{u}^k_{\mathrm{pm},\Gamma_\mathrm{in}}$
Post
$u^{k+1}_{\mathrm{pm},\Gamma_\mathrm{in}}$
$k \mapsto k+1$

$$u^{k+1}_{\mathrm{pm},\Gamma_\mathrm{in}} = \mathrm{Post}(\mathrm{PM}(\mathrm{FF}(u^k_{\mathrm{pm},\Gamma_\mathrm{in}})))$$

Fig. 2 Serial-implicit coupling: The free-flow solver (FF) computes a new flow state given the normal velocity $u^k_{\mathrm{pm},\Gamma_\mathrm{in}}$. The pressure $p^{k+1}_{\mathrm{ff},\Gamma_\mathrm{in}}$ is used in the coupling condition of the porous-medium solver (PM) computing a new flow state. The post-processing scheme (Post) gets the velocity $\tilde{u}^k_{\mathrm{pm},\Gamma_\mathrm{in}}$ and computes a new velocity $u^{k+1}_{\mathrm{pm},\Gamma_\mathrm{in}}$ based on the current value and a history of previous input values


$k \mapsto k+1$
$u^k_{\mathrm{pm},\Gamma_\mathrm{in}}$
FF
$\tilde{p}^k_{\mathrm{ff},\Gamma_\mathrm{in}}$
$p^{k+1}_{\mathrm{ff},\Gamma_\mathrm{in}}$
Post
PM
$p^k_{\mathrm{ff},\Gamma_\mathrm{in}}$
$\tilde{u}^k_{\mathrm{pm},\Gamma_\mathrm{in}}$
$u^{k+1}_{\mathrm{pm},\Gamma_\mathrm{in}}$
$k \mapsto k+1$

$$\begin{pmatrix} u^{k+1}_{\mathrm{pm},\Gamma_\mathrm{in}} \\ p^{k+1}_{\mathrm{ff},\Gamma_\mathrm{in}} \end{pmatrix} = \mathrm{Post}\begin{pmatrix} \mathrm{FF}(u^k_{\mathrm{pm},\Gamma_\mathrm{in}}) \\ \mathrm{PM}(p^k_{\mathrm{ff},\Gamma_\mathrm{in}}) \end{pmatrix}$$

Fig. 3 Parallel-implcit coupling: The free-flow solver (FF) and the porous-medium solver (PM) compute a new pressure $\tilde{p}^k_{\mathrm{ff},\Gamma_\mathrm{in}}$ and a new velocity $\tilde{u}^k_{\mathrm{pm},\Gamma_\mathrm{in}}$ based on previous values $u^k_{\mathrm{pm},\Gamma_\mathrm{in}}$ and $p^k_{\mathrm{ff},\Gamma_\mathrm{in}}$. The resulting values $\tilde{p}^k_{\mathrm{ff},\Gamma_\mathrm{in}}$ and $\tilde{u}^k_{\mathrm{pm},\Gamma_\mathrm{in}}$ are used by the post-processing scheme (Post) in order to obtain $p^{k+1}_{\mathrm{ff},\Gamma_\mathrm{in}}$ and $u^{k+1}_{\mathrm{pm},\Gamma_\mathrm{in}}$ based on the current values and a history of previous input values

coupling types correspond to a fixed-point iteration accelerated by a so-called post-processing Post as shown in Figs. 2 and 3. We use the relative convergence measure

$$\|p^{k+1}_{\mathrm{ff},\Gamma_\mathrm{in}} - p^k_{\mathrm{ff},\Gamma_\mathrm{in}}\|_2 < \varepsilon \cdot \|p^{k+1}_{\mathrm{ff},\Gamma_\mathrm{in}}\|_2 \ \text{ and } \ \|\tilde{u}^k_{\mathrm{pm},\Gamma_\mathrm{in}} - u^k_{\mathrm{pm},\Gamma_\mathrm{in}}\|_2 < \varepsilon \cdot \|\tilde{u}^k_{\mathrm{pm},\Gamma_\mathrm{in}}\|_2, \quad (5)$$

in the serial-implicit case and

$$\|\tilde{p}^k_{\mathrm{ff},\Gamma_\mathrm{in}} - p^k_{\mathrm{ff},\Gamma_\mathrm{in}}\|_2 < \varepsilon \cdot \|\tilde{p}^k_{\mathrm{ff},\Gamma_\mathrm{in}}\|_2 \ \text{ and } \ \|\tilde{u}^k_{\mathrm{pm},\Gamma_\mathrm{in}} - u^k_{\mathrm{pm},\Gamma_\mathrm{in}}\|_2 < \varepsilon \cdot \|\tilde{u}^k_{\mathrm{pm},\Gamma_\mathrm{in}}\|_2, \quad (6)$$

for the parallel-implicit case. The tolerance $\varepsilon > 0$ is user-defined.

In this work, we choose the inverse least squares interface quasi-Newton method as post-processing method. The method approximates the inverse of the Jacobian of the interface problem (4) based on the given input/output data pairs and carries out a norm minimization in order to compute an improved solution for the interface values. We refer to the overview paper of Degroote [5] for more information about this method while implementation details are given in the corresponding preCICE publication [2].

To validate and analyze the partitioned setup, we compare it to a monolithic coupling approach that is part of DuMu$^{\text{x}}$ [10]. The monolithic coupling solves the flow problem on the whole domain Ω in one single system. The resulting linear system of equations is solved with a direct solver since it is too ill-conditioned to be solved iteratively. This comes at the cost of limited flexibility, as one cannot easily reuse existing solvers, together with larger memory requirements and worse complexity compared to the partitioned coupling.

4 Numerical Results

We simulate a pressure-driven horizontal free flow over a porous medium. The free-flow domain $\Omega_{\text{ff}} = [0, 1] \times [1, 2]$ and the porous-medium domain $\Omega_{\text{pm}} = [0, 1] \times [0, 1]$ feature the same spatial extent. We use equidistant grids with $N = 40$ grid cells in each spatial dimension and the same number of grid cells in the free-flow grid and the porous-medium grid. This way, we avoid possible errors introduced due to data mapping between non-matching grids. The superscript part and mono corresponds to a solution computed using the partitioned or the monolithic coupling. No-flow/no-slip conditions hold at the upper boundary of Ω_{ff} while fixed-pressure conditions are set at the left and right boundaries. All sides of Ω_{pm} (except the coupling interface) are closed with no-flow conditions.

The pressure difference driving the flow in Ω_{ff} is set to $\Delta p = 10^{-9}\text{Pa}$. The permeability is set to $K = 10^{-6}\text{m}^2$ and we set the Beavers–Joseph–Saffman parameter to $\alpha_{\text{BJS}} = 1.0$. The density and the viscosity of the fluid are $\rho = 10^3\text{kg/m}^3$ and $\mu = 10^{-3}\text{Pa} \cdot \text{s}$. The initial velocities and initial pressure are zero in both domains. The flows in the free-flow and the porous-medium domain are solved with DuMu$^{\text{x}}$. The partitioned coupling is realized via preCICE with $\varepsilon = 10^{-8}$ while the monolithic coupling uses the coupling capabilities of DuMu$^{\text{x}}$. The system of equations is solved using Newton's method and the linear system of equations is solved using the direct solver provided by UMFPACK [4].

In Fig. 4 we present the convergence behavior for the serial-implicit coupling (*case a*, Fig. 2), the serial-implicit coupling with interchanged order of the solvers (*case b*, similar to Fig. 2, but with $p^{k+1}_{\text{ff},\Gamma_{\text{in}}} = \text{Post}(\text{FF}(\text{PM}(p^{k}_{\text{ff},\Gamma_{\text{in}}})))$) and the parallel-implicit coupling (Fig. 3, *case c*). All employed coupling approaches converge and reach the defined tolerance within at most 7 iterations. The coupling in *case a* needs one iteration less than *case b*. This is due to the initialization of the solvers with zero initial conditions which means that the solver PM does not compute any flow in the first iteration for *case b*. This could be improved by providing better initial data on the coupling interface. However, *case b* shows a monotonous convergence behavior and also converges to much smaller residuals for the pressure. The coupling in *case c* needs the most coupling iterations, but is still able to reach the defined tolerance. This behavior is also expected since the parallel-implicit coupling is more "loose" than the serial couplings, but allows the concurrent execution of both solvers.

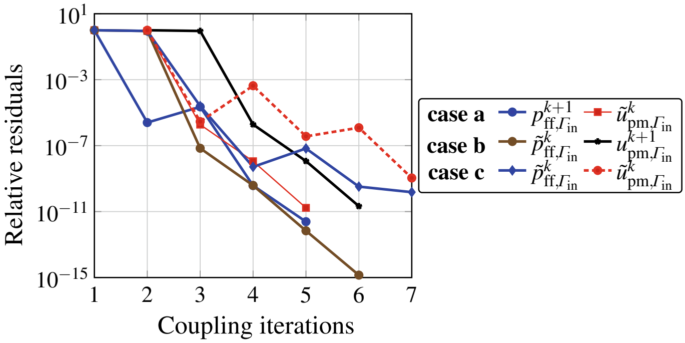


Fig. 4 We plot the relative residuals computed according to (5) (*case a* and *b*) and (6) (*case c*) for the user-given tolerance $\varepsilon = 10^{-8}$. The residuals are plotted for a serial-implicit coupling as in Fig. 2 (*case a*), serial-implicit coupling with interchanged order of solvers (*case b*) and parallel-implicit coupling (*case c*) as described in Fig. 3

Table 1 Relative differences between the monolithic and partitioned solution for $\varepsilon = 10^{-8}$ and the three different couplings employed. The partitioned solutions coincide for at least 8 significant digits due to ε. The differences have been computed using single precision floating point numbers and thus the errors coincide for all cases

case	$\Vert\mathbf{u}_\mathrm{ff}^\mathrm{mono} - \mathbf{u}_\mathrm{ff}^\mathrm{part}\Vert_\infty^\mathrm{rel}$	$\Vert\mathbf{u}_\mathrm{pm}^\mathrm{mono} - \mathbf{u}_\mathrm{pm}^\mathrm{part}\Vert_\infty^\mathrm{rel}$	$\Vert p_\mathrm{ff}^\mathrm{mono} - p_\mathrm{ff}^\mathrm{part}\Vert_\infty^\mathrm{rel}$	$\Vert p_\mathrm{pm}^\mathrm{mono} - p_\mathrm{pm}^\mathrm{part}\Vert_\infty^\mathrm{rel}$
a/b/c	3.906576E-03	1.011851E-06	3.971279E-05	2.092258E-06

In Table 1 we present the relative differences between the solutions obtained with the partitioned and the monolithic coupling. All three partitioned couplings types give very similar results such that the errors are identical relative to the monolithic solution which is not unexpected as $\varepsilon = 10^{-8}$. The errors are reasonably small, but it stands out that the errors in the free-flow domain Ω_ff are bigger than in the porous-medium domain Ω_pm. The coupling conditions are possibly handled in a slightly different way in the partitioned approach than in the monolithic one.

So far, the monolithic approach has been faster than the partitioned approach, but the implementation of the partitioned approach within DuMu$^\mathrm{x}$ has not been optimized yet. In particular, the same direct solver is employed for the two subproblems. Assuming roughly the same number of unknowns $n/2$ in the porous medium and the free flow-domain and that the computational effort in 2d for a typical band matrix is n^2, using only direct solvers requires $2(n/2)^2 = n^2/2$ per iteration and yields a computational benefit only if two or less iterations are sufficient. Therefore we do not report any CPU times for the studied cases.

5 Conclusion and Outlook

In this work, we have investigated a partitioned black-box coupling scheme for free-flow and porous-media applications where the two flow regimes are separated by a sharp interface. Iterative partitioned coupling schemes with an inverse least squares interface quasi-Newton method as post-processing have been employed.

All investigated couplings types converge to the desired tolerance with different iteration counts. The serial-implicit coupling solving the free-flow problem needs the least (5) iterations while the parallel-implicit coupling needs the most (7) iterations for our test case. The serial-implicit coupling solving the porous-medium problem first shows the best convergence behavior.

When comparing the solution obtained from the partitioned coupling scheme with a monolithic coupling, we obtain reasonably small errors that coincide for the studied couplings. The deviations in the free-flow domain are larger than in the porous-medium domain which might be caused by a slightly different handling of the coupling conditions in the implementation and should be further investigated.

Future work will focus on understanding what introduces the errors in the free-flow domain. We will investigate further coupling procedures and other post-processing methods, parallel simulation at higher resolution and higher Reynolds numbers as well as time-dependent flows. Moreover, we are interested in studying the influence of the execution order of solvers and other coupling conditions. We plan to utilize the black-box nature of the coupling by using different solvers in the different domains which would also include unstructured and non-matching grids. This will allow to increase the efficiency of the partitioned approach.

Acknowledgements This work has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—Project Number 327154368—SFB 1313.

References

1. Benzi, M., Golub, G.H., Liesen, J.: Numerical solution of saddle point problems. Acta Numer. **14**, 1–137 (2005). https://doi.org/10.1017/s0962492904000212
2. Bungartz, H.J., Lindner, F., Gatzhammer, B., Mehl, M., et al.: preCICE - a fully parallel library for multi-physics surface coupling. Advances in Fluid-Structure Interaction. Comput. Fluids **141**, 250–258 (2016). https://doi.org/10.1016/j.compfluid.2016.04.003
3. Caiazzo, A., John, V., Wilbrandt, U.: On classical iterative subdomain methods for the Stokes-Darcy problem. Comput. Geosci. **18**(5), 711–728 (2014). https://doi.org/10.1007/s10596-014-9418-y
4. Davis, T.A.: Algorithm 832: UMFPACK V4.3—an Unsymmetric-Pattern Multifrontal Method. ACM Trans. Math. Softw. **30**(2), 196–199 (2004). https://doi.org/10.1145/992200.992206
5. Degroote, J.: Partitioned simulation of fluid-structure interaction. Arch. Comput. Methods Eng. **20**(3), 185–238 (2013). https://doi.org/10.1007/s11831-013-9085-5
6. Discacciati, M., Gerardo-Giorda, L.: Optimized Schwarz methods for the Stokes-Darcy coupling. IMA J. Numer. Anal. **38**(4), 1959–1983 (2017). https://doi.org/10.1093/imanum/drx054

7. Discacciati, M., Quarteroni, A.: Navier-Stokes/Darcy coupling: modeling, analysis, and numerical approximation. Rev. Mat. Complut. **22**(2), 315–426 (2009). https://doi.org/10.5209/rev_REMA.2009.v22.n2.16263
8. Flemisch, B., Darcis, M., Erbertseder, K., Faigle, B., et al.: DuMu^x: DUNE for multi-{phase, component, scale, physics,…} flow and transport in porous media. Adv. Water Resour. **34**(9), 1102–1112 (2011). https://doi.org/10.1016/j.advwatres.2011.03.007
9. Harlow, F.H., Welch, J.E.: Numerical calculation of time-dependent viscous incompressible flow of fluid with free surface. Phys. Fluids **8**(12), 2182–2189 (1965). https://doi.org/10.1063/1.1761178
10. Koch, T., Gläser, D., Weishaupt, K., Ackermann, S., et al.: DuMux 3—an open-source simulator for solving flow and transport problems in porous media with a focus on model coupling. Comput Math Appl. (2020). https://doi.org/10.1016/j.camwa.2020.02.012
11. Saffman, P.G.: On the boundary condition at the surface of a porous medium. Stud. Appl. Math. **50**(2), 93–101 (1971). https://doi.org/10.1002/sapm197150293

Challenges in Drift-Diffusion Semiconductor Simulations

Patricio Farrell and Dirk Peschka

Abstract We study and compare different discretizations of the van Roosbroeck system for charge transport in bulk semiconductor devices that can handle nonlinear diffusion. Three common challenges corrupting the precision of numerical solutions will be discussed: boundary layers, discontinuities in the doping profile, and corner singularities in L-shaped domains. The most problematic of these challenges are boundary layers in the quasi-Fermi potentials near ohmic contacts, which can have a drastic impact on the convergence order.

Keywords Finite volume method · Finite element method · Nonlinear diffusion · Scharfetter-gummel scheme · Semiconductors · Van roosbroeck system · Convergence order · Diffusion enhancement

MSC (2010) 35Q99 · 82D37 · 65M08 · 65M06 · 65M60

1 Introduction

The present paper aims at comparing different discretization philosophies for semiconductor problems. We study three major challenges for recent finite element and finite volume schemes which are designed to deal with nonlinear diffusion in a thermodynamic consistent way and are based on quasi-Fermi potentials as primary variables. In particular, we study the error and convergence rate of the numerical solutions in the presence of: boundary layers, discontinuous doping profile and corner singularities.

P. Farrell (✉) · D. Peschka
Weierstrass Institute (WIAS), Mohrenstr. 39, 10117 Berlin, Germany
e-mail: patricio.farrell@wias-berlin.de

D. Peschka
e-mail: dirk.peschka@wias-berlin.de

R. Klöfkorn et al. (eds.), *Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples*, Springer Proceedings in Mathematics & Statistics 323,
https://doi.org/10.1007/978-3-030-43651-3_58

2 Modelling Semiconductors with Ohmic Contacts

2.1 Stationary van Roosbroeck System

The van Roosbroeck system is a drift-diffusion model, which describes the recombination and transport of charge carriers driven by diffusion and by electric fields within a semiconductor device. It consists of three nonlinear, coupled partial differential equations for the electrostatic potential $\psi : \Omega \to \mathbb{R}$ as well as the non-negative electron and hole densities $n : \Omega \to \mathbb{R}^+$ and $p : \Omega \to \mathbb{R}^+$, namely a Poisson equation and two continuity equations. We consider a homogeneous material and some domain $\Omega \subseteq \mathbb{R}^d$ for $d \in \{1, 2, 3\}$ in an isothermal setting. Then the stationary van Roosbroeck system is given by the system of elliptic partial differential equations

$$-\nabla \cdot (\varepsilon_0 \varepsilon_r \nabla \psi) = q \left(C + p(\psi, \varphi_p) - n(\psi, \varphi_n) \right), \tag{1a}$$

$$\nabla \cdot \mathbf{j}_n = +qR, \tag{1b}$$

$$\nabla \cdot \mathbf{j}_p = -qR, \tag{1c}$$

where q denotes the elementary charge, ε_0 is the vacuum permittivity and ε_r is the relative permittivity of the material. The recombination rate R and the charge-carrier currents $\mathbf{j}_n, \mathbf{j}_p$ depend on the solution n, p, ψ and vanish in thermal equilibrium. The given doping concentration $C : \Omega \to \mathbb{R}$ (intentionally introduced impurities) varies spatially and can have discontinuities. The equations of state are given by

$$n(\psi, \varphi_n) = N_c \mathscr{F} \left(\frac{q(\psi - \varphi_n) - E_c}{k_B T} \right), \tag{2a}$$

$$p(\psi, \varphi_p) = N_v \mathscr{F} \left(\frac{q(\varphi_p - \psi) + E_v}{k_B T} \right), \tag{2b}$$

where the statistical distribution function $\mathscr{F}$ relates the electron and hole densities n, p to the quasi-Fermi potentials φ_n, φ_p. Working with quasi-Fermi potentials has all the advantages mentioned in the introduction, in particular from a modeling and computational point of view. Furthermore, we set the recombination rate to zero as it plays a minor role for most of our considerations.

The effective density of states for electrons in the conduction band N_c and holes in the valence band N_v as well as the corresponding band-edge energies E_c, E_v and the band gap $E_g = E_c - E_v$ are material parameters and assumed to be spatially constant in this paper. Temperature and the Boltzmann constant are denoted with T and k_B. The three most important reference cases for the statistical distribution functions are the Boltzmann, Blakemore and Fermi-Dirac function. For each distribution function, the corresponding current densities in (1b) and (1c) are

$$\mathbf{j}_n = -q\mu_n n \nabla\varphi_n = -q\mu_n n\nabla\psi + qD_n\nabla n, \tag{3a}$$

$$\mathbf{j}_p = -q\mu_p p \nabla\varphi_p = -q\mu_p p\nabla\psi - qD_p\nabla p. \tag{3b}$$

Using the thermal voltage $U_T = \frac{k_B T}{q}$, the diffusion coefficients D_n, D_p are linked to the carrier mobilities μ_n, μ_p via generalized Einstein relations

$$\frac{D_n}{\mu_n} = U_T\, g\big(\eta_n\big), \qquad \frac{D_p}{\mu_p} = U_T\, g\big(\eta_p\big), \qquad g(\eta) = \frac{\mathscr{F}(\eta)}{\mathscr{F}'(\eta)}, \tag{4}$$

where g is the diffusion enhancement as motivated in [8].

The system (1) is supplied with mixed Dirichlet-Neumann boundary conditions.

3 Discretization of the van Roosbroeck System Using Potentials

In the following we are going to explain standard discretization methods to solve the van Roosbroeck system.

3.1 Finite Element Method

Assume $\Omega \subset \mathbb{R}^2$ is a polygonal domain and let $\mathscr{T}_h$ be an admissible decomposition of Ω into N_{tria} triangles and N_{vert} vertices, such that $\Omega = \bigcup_{t=1}^{N_{\text{tria}}} \tau_t$ for $\tau_t \in \mathscr{T}_h$. Similar as in [2], we solve the stationary van Roosbroeck system (1) using a standard P_1 finite element method. We seek the electrostatic potential and the quasi-Fermi potentials $\mathbf{u}^h = (\psi^h, \varphi_n^h, \varphi_p^h) \in V^h$, such that the van Roosbroeck system can be written in the weak form as

$$0 = \int_\Omega \Big(\varepsilon_0\varepsilon_r \nabla\psi^h \cdot \nabla v_i - q\big(C + p(\psi^h, \varphi_p^h) - n(\psi^h, \varphi_n^h)\big)v_i\Big)\,\mathrm{d}x, \tag{5a}$$

$$0 = \int_\Omega \Big(q\mu_n n(\psi^h, \varphi_n^h)\nabla\varphi_n^h \cdot \nabla v_j - qR\left(n(\psi^h, \varphi_n^h), p(\psi^h, \varphi_p^h)\right) v_j\Big)\,\mathrm{d}x, \tag{5b}$$

$$0 = \int_\Omega \Big(q\mu_p p(\psi^h, \varphi_p^h)\nabla\varphi_p^h \cdot \nabla v_k + qR\left(n(\psi^h, \varphi_n^h), p(\psi^h, \varphi_p^h)\right) v_k\Big)\,\mathrm{d}x, \tag{5c}$$

for all suitable test functions $\mathbf{v}^h = (v_i, v_j, v_k) \in V^h$, where $V^h \cong \mathbb{R}^{N_{\text{vert}}\times 3}$ is the $3N_{\text{vert}}$ dimensional space of vectorial continuous functions which are piecewise linear on each triangle τ_t.

3.2 Finite Volume Method

In this section, we present a Voronoï finite volume technique [4, 6, 7]. Similar as for finite elements, we start by partitioning the domain Ω into non-intersecting, convex polyhedral control volumes ω_k such that $\Omega = \bigcup_{k=1}^{N_{\text{vert}}} \omega_k$. We associate with each control volume ω_k a node $\mathbf{x}_k \in \omega_k$. For every boundary intersecting control volume, we demand that this node lies on the boundary $\mathbf{x}_k \in \partial\Omega \cap \omega_k$. Assuming the partition is admissible [3], i.e. for two adjacent control volumes ω_k and ω_l, the edge $\overline{\mathbf{x}_k\mathbf{x}_l}$ of length h_{kl} is orthogonal to $\partial\omega_k \cap \partial\omega_l$, the normal vectors to $\partial\omega_k$ can be calculated by $\boldsymbol{\nu}_{kl} = (\mathbf{x}_l - \mathbf{x}_k)/\|\mathbf{x}_l - \mathbf{x}_k\|$. We note that the variables $(\psi, \varphi_n, \varphi_p)$ are of interest only at the nodes, not at the edges.

For each control volume ω_k, the finite volume discretization is given by the three equations:

$$\sum_{\omega_l \in \mathcal{N}(\omega_k)} |\partial\omega_k \cap \partial\omega_l| j_{\psi;k,l} = q|\omega_k| \left(C_k + p(\psi_k, \varphi_{p;k}) - p(\psi_k, \varphi_{n;k})\right), \tag{6a}$$

$$\sum_{\omega_l \in \mathcal{N}(\omega_k)} |\partial\omega_k \cap \partial\omega_l| j_{n;k,l} = +q|\omega_k| R(\psi_k, \varphi_{n;k}, \varphi_{p;k}), \tag{6b}$$

$$\sum_{\omega_l \in \mathcal{N}(\omega_k)} |\partial\omega_k \cap \partial\omega_l| j_{p;k,l} = -q|\omega_k| R(\psi_k, \varphi_{n;k}, \varphi_{p;k}). \tag{6c}$$

We denote with $\mathcal{N}(\omega_k)$ the set of all control volumes neighboring ω_k. In 2D, the measure $|\partial\omega_k \cap \partial\omega_l|$ corresponds to the length of the boundary line segment and in 3D to the area of the intersection of the boundary surfaces.

The unknowns $\psi_k, \varphi_{n;k}, \varphi_{p;k}$ correspond to the electrostatic potential as well as the quasi-Fermi potentials for electrons and holes evaluated at node $\mathbf{x}_k$. To approximate the fluxes in (6) using general $\mathcal{F}$, ideas from [1] are useful to derive a finite volume scheme for convection-diffusion problems in a *thermodynamically consistent* way by averaging the nonlinear diffusion term appropriately.

4 Numerical Examples

In this section, we are going to present numerical solutions of the van Roosbroeck system via FE and the Scharfetter-Gummel FV discretization. We focus on two challenges, which have an impact on the convergence rate of solutions: the size of a boundary layer and the regularity of the doping. Since in this section we are mostly concerned with numerical solutions, we will drop the superindex h. If necessary, we replace it with the acronym of the corresponding discretization method. Also we remind the reader that we solve the van Roosbroeck system without recombination, *i.e.*, $R \equiv 0$. Throughout this section, we use the Blakemore distribution function.

4.1 Resolution of Boundary Layer

In Fig. 1 the densities n, p and the doping C are shown for the two cases $\kappa = 5 \cdot 10^2$ and $\kappa = 5 \cdot 10^5$ at $V_{\text{ext}} = 3\,\text{V}$. Note that in both cases, the hole density p has a boundary layer at $x = 0$ and the electron density n has a boundary layer at $x = 0.3\,\mu\text{m}$. This boundary layer, however, is on the length scale of λ_D and therefore nicely resolved by the mesh. On the level of the plot, the difference in solutions corresponding to the two alternative doping profiles is not visible. In the left panel of Fig. 2 we show the potentials $(\psi, \varphi_n, \varphi_p)$ for $V_{\text{ext}} = 3\,\text{V}$. While the electrostatic potential in both cases is a rather smooth function (blue line), the quasi-Fermi potentials have a boundary layer of size ℓ_J (green and red line) that can not be resolved on any of the uniform meshes. This logarithmic boundary layer is predicted by our analysis in [5]. As one can see in Fig. 2 (middle and right panel), the solution effectively jumps within the last interval before the ohmic contact.

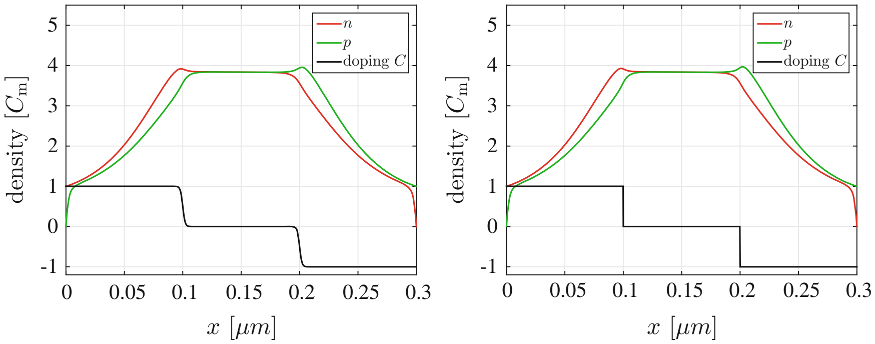


Fig. 1 1D electron and hole densities n, p and doping C at bias $V_{\text{ext}} = 3\,\text{V}$ shown (left) with $\kappa = 500$ and (right) with $\kappa = 5 \cdot 10^5$, the former yielding a smooth doping profile and the latter practically a discontinuous one

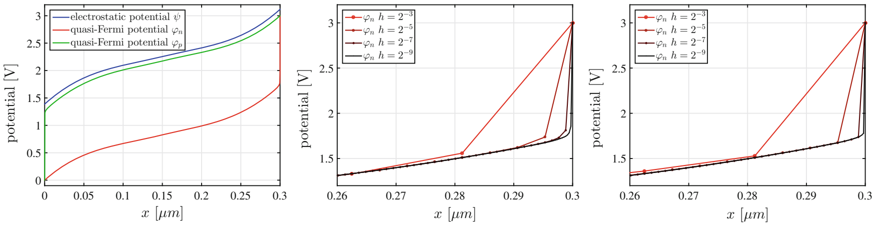


Fig. 2 1D quasi-Fermi potentials of electrons and holes φ_n, φ_p and electrostatic potential ψ (left) with bias $V_{\text{ext}} = 3\,\text{V}$ as well as boundary layers in the electron quasi-Fermi potential φ_n near $x = 0.3\,\mu\text{m}$ for different mesh resolutions h (middle) for finite element and (right) Scharfetter-Gummel type finite volume discretization for $\kappa = 500$

4.2 Regularity of the Doping

Next, we discuss the influence of the smoothness of the doping on the convergence order for the different discretization methods. Whenever we compare a coarse discrete solution (of size $2h$) to a finer one (of size h), we restrict the finer solution to the coarser mesh. Then we can subtract $\mathbf{u}^h$ from $\mathbf{u}^{2h}$ and slightly abusing the notation write $\|\mathbf{u}^h - \mathbf{u}^{2h}\|$ for the corresponding norm. Provided that the doping is sufficiently smooth and the carrier densities converge sufficiently fast, then the FV discretization of the Poisson equation is second order accurate, see the convergence for n, p, ψ in the top right panel of Fig. 3. When the doping is discontinuous ($\kappa = 5 \cdot 10^5$), the bottom row in Fig. 3 shows that also the convergence order of the FV electrostatic potential becomes linear, which is plausible by standard FE error estimates. From Fig. 3 it appears that while the error in the FE method is dominated by the quasi-Fermi potentials, the error in the FV method is dominated by the lack of regularity in the doping.

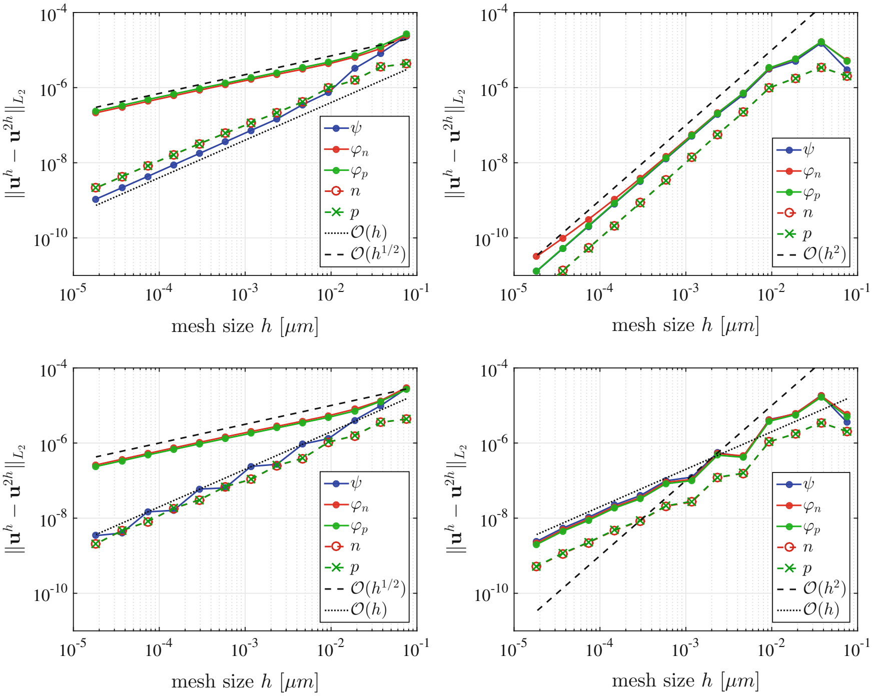


Fig. 3 L_2 convergence rates in 1D for solution (left) of the FE discretization and (right) of the FV discretization with $\kappa = 500$ in the top row and for $\kappa = 5 \cdot 10^5$ in the bottom row at $V_{\text{ext}} = 3\,\text{V}$

4.3 Corner Singularities and Boundary Adapted Meshes

Semiconductor devices may often be angular-shaped. However, in particular L-shaped domains pose numerical difficulties which we would like to study for the FE and FV methods. We consider a two-dimensional L-shaped domain

$$\Omega = [0, 2L]^2 \setminus [0, L]^2 \subset \mathbb{R}^2, \tag{7}$$

and impose ohmic contacts at the boundaries $(x, 0)$ and $(0, y)$ for $L \le x, y \le 2L$. All other boundaries are supplied with homogeneous Neumann boundary conditions. The p-i-n doping concentration $C\colon \Omega \to \mathbb{R}$ is given by

$$C(\mathbf{x}) = \begin{cases} +C_0 & 0 \le x \le L/2, \\ -C_0 & 0 \le y \le L/2, \\ +2C_0(L-x)/L & L/2 < x \le L, \\ -2C_0(L-y)/L & L/2 < y \le L, \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

with $\mathbf{x} = (x, y)$ and as before $L = 10^{-7}m = 0.1\,\mu\text{m}$ and $C_0 = C_\text{m} = 10^{23}\,\text{m}^{-3}$. With this choice we ensure that the convergence order does not suffer from the regularity of the doping. However, constructing a non-convex domain with a corner angle $\vartheta = \theta\pi$ and $\theta = 3/2$ imposes a corner singularity of the form $\psi(\mathbf{x}) \sim r^{1/\theta}$ as $r \to 0$ for $r = \sqrt{(x - x_0)^2 + (y - y_0)^2}$ at $x_0 = y_0 = 0.1\,\mu\text{m}$.

The upper left panel of Fig. 4 shows the convergence of the electron quasi-Fermi potentials at $V_\text{ext} = 3$ V, where the FE and FV are compared on a sequence of uniform and a sequence of boundary adapted meshes. As in 1D, the FV method converges quadratically. Furthermore, for the FV discretization the error seems not to be influenced very much by the boundary adapted meshes. In contrast, the FE method again has a lower convergence order and local adaptivity improves the L_2 error of the solution by about one order of magnitude.

The lower panels of Fig. 4 show the solutions at $V_\text{ext} = 0.2$ V, where the boundary layer is moderate and solutions are closer to thermal equilibrium. Hence, the lower left panel shows the general tendency to have lower errors. However, the convergence is slower with an order between $\mathcal{O}(h)$ and $\mathcal{O}(h^{4/3})$, indicating a stronger influence of the corner singularity. This effect is even more pronounced in the lower right panel, in which for all the used methods the convergence of the electrostatic potential nicely follows the $\mathcal{O}(h^{4/3})$ order predicted by the error analysis of the corner singularity.

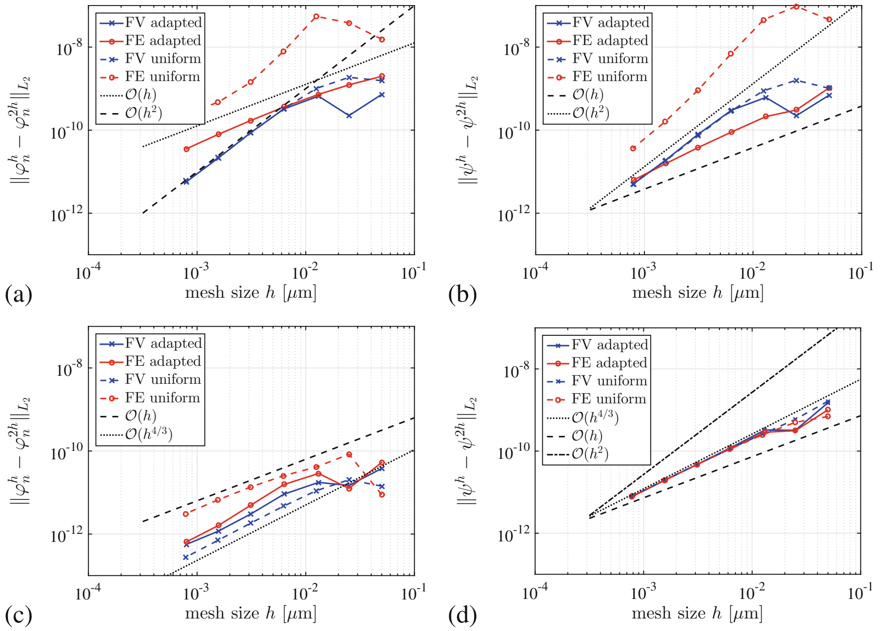


Fig. 4 Convergence of solutions on different meshes as a function of relative triangle size $h = 2^{-\ell}$ for **a** electron quasi-Fermi potential φ_n at $V_{\text{ext}} = 3\,\text{V}$, **b** electrostatic potential ψ at $V_{\text{ext}} = 3\,\text{V}$, **c** electron quasi-Fermi potential φ_n at $V_{\text{ext}} = 0.2\,\text{V}$, **d** electrostatic potential ψ at $V_{\text{ext}} = 0.2\,\text{V}$

5 Conclusion

Summarizing, in 2D both FE and FV discretizations deliver reasonable results. While the finite volume scheme often shows better convergence rates, the finite element method can be drastically improved by using meshes which are finer near ohmic contacts. We clearly observe that depending on the potential and the selected bias, the error is dominated by the boundary layer or the corner singularity. While the FV method generally handles the boundary layer well, the FE method in 2D introduces extra oscillations in the boundary layer, see [5] for details.

References

1. Bessemoulin-Chatard, M.: A finite volume scheme for convection-diffusion equations with nonlinear diffusion derived from the Scharfetter-Gummel scheme. Numer. Math. **121**(4), 637–670 (2012). https://doi.org/10.1007/s00211-012-0448-x
2. der Maur, MA., Povolotskyi, M., Sacconi, F., Pecchia, A., Romano, G., Penazzi, G., Di Carlo, A.: TiberCAD: towards multiscale simulation of optoelectronic devices. Opt. Quantum Electron. **40**(14-15), 1077–1083 (2008)

3. Eymard, R., Gallouët, T., Herbin, R.: Finite volume methods. In: Solution of Equation in $\mathbb{R}^n$ (Part 3), Techniques of Scientific Computing (Part 3), Handbook of Numerical Analysis, vol. 7, pp. 713 – 1018. Elsevier (2000)
4. Farrell, P., Rotundo, N., Doan, D.H., Kantner, M., Fuhrmann, J., Koprucki, T.: Mathematical methods: drift-diffusion models. In: J. Piprek (ed.) Handbook of Optoelectronic Device Modeling and Simulation, chap. 50, pp. 733–772. Taylor & Francis (2017)
5. Farrell, P., Peschka, D.: Nonlinear diffusion, boundary layers and nonsmoothness: analysis of challenges in drift-diffusion semiconductor simulations. Comput. Math. Appl. (2019). https://doi.org/10.1016/j.camwa.2019.06.007
6. Gajewski, H.: Analysis und Numerik von Ladungstransport in Halbleitern. WIAS Report (6) (1993). ISSN 0942-9077
7. Gärtner, K.: Existence of bounded discrete steady-state solutions of the Van Roosbroeck system on boundary conforming Delaunay grids. SIAM J. Sci. Comput. **31**(2), 1347–1362 (2009). https://doi.org/10.1137/070710950
8. van Mensfoort, S.L.M., Coehoorn, R.: Effect of Gaussian disorder on the voltage dependence of the current density in sandwich-type devices based on organic semiconductors. Phys. Rev. B **78**(8) (2008). https://doi.org/10.1103/physrevb.78.085207

Unipolar Drift-Diffusion Simulation of S-Shaped Current-Voltage Relations for Organic Semiconductor Devices

Jürgen Fuhrmann, Duy Hai Doan, Annegret Glitzky, Matthias Liero, and Grigor Nika

Abstract We discretize a unipolar electrothermal drift-diffusion model for organic semiconductor devices with Gauss–Fermi statistics and charge carrier mobilities having positive temperature feedback. We apply temperature dependent Ohmic contact boundary conditions for the electrostatic potential and use a finite volume based generalized Scharfetter-Gummel scheme. Applying path-following techniques we demonstrate that the model exhibits S-shaped current-voltage curves with regions of negative differential resistance, only recently observed experimentally.

Keywords Non-isothermal drift-diffusion · Organic semiconductors · Finite volumes · Generalized Scharfetter-Gummel scheme · Path following

MSC (2010) 65M08 · 35J92 · 80M12 · 80A20

1 Introduction

The temperature activated hopping transport of charge carriers in organic semiconductors results in a strong interplay between electric current and heat flow. It gives rize to interesting phenomena like S-shaped Current-Voltage (CV) relations with regions

J. Fuhrmann (✉) · A. Glitzky · M. Liero · G. Nika
Weierstrass Institute, Mohrenstraße 39, 10117 Berlin, Germany
e-mail: fuhrmann@wias-berlin.de

A. Glitzky
e-mail: glitzky@wias-berlin.de

M. Liero
e-mail: liero@wias-berlin.de

G. Nika
e-mail: nika@wias-berlin.de

D. Hai Doan
m4sim GmbH, Seydelstr. 31, 10117 Berlin, Germany
e-mail: duyhai.doan@m4sim.de

R. Klöfkorn et al. (eds.), *Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples*, Springer Proceedings in Mathematics & Statistics 323,
https://doi.org/10.1007/978-3-030-43651-3_59

of negative differential resistance in Organic Light Emitting Diodes (OLEDs) [12] or leads to inhomogeneous luminance in large-area OLEDs. Moreover, electrothermal effects influence the performance of transistors [10].

As demonstrated in [12], p-Laplace thermistor models that describe the total current and heat flow in a device, are able to capture the positive temperature feedback in OLEDs. Especially, they can reproduce experimentally observed S-shaped CV-relations and inhomogeneous current density and temperature distributions in large-area OLEDs. But, details such as separate electron and hole current flow, generation-recombination and related heat productions, as well as energy barriers at material interfaces cannot be included.

In this paper, we present a numerical approximation of an electrothermal drift-diffusion model for organic semiconductor devices and study its ability to reproduce S-shaped CV-relations. For simplicity, for this proof of concept we use vertically layered device structures.

2 Electrothermal Drift-Diffusion Description of Organic Semiconductor Devices

We restrict our considerations to the unipolar (n-doped) case, for the full model see [3]. Then the electrothermal behavior is described in a drift-diffusion setting by PDEs for the electrostatic potential ψ, the electrochemical potential φ_n, and the temperature T. In the device domain Ω we consider the stationary coupled system

$$\begin{aligned} -\nabla\cdot(\varepsilon\nabla\psi) &= q(C-n),\\ -\nabla\cdot j_n = 0, \quad j_n &= -qn\mu_n\nabla\varphi_n,\\ -\nabla\cdot(\lambda\nabla T) &= qn\mu_n|\nabla\varphi_n|^2 =: H. \end{aligned} \tag{1}$$

This system results from the coupling of a generalized, unipolar van Roosbroeck system and a heat flow equation that includes the Joule heating as heat source. The dielectric permittivity is denoted by $\varepsilon = \varepsilon_0\varepsilon_r$, q is the elementary charge, C represents the doping density, and λ is the thermal conductivity.

Additionally we take into account the specialities of organic semiconductors, namely (i) the statistical relation between chemical potential and charge carrier density is given by Gauss–Fermi integrals leading to bounded charge carrier densities and (ii) a mobility function μ_n depending on temperature, density, and electric field strength. The mobility laws are fitted from a numerical solution of the master equation for the hopping transport in a disordered energy landscape with a Gaussian density of states [11]. The charge carrier density n in (1) is given by

$$n = N_{n0}(T)G\Big(\frac{q(\psi-\varphi_n)-E_L(T)}{k_\mathrm{B}T};\frac{\sigma_n(T)}{k_\mathrm{B}T}\Big), \tag{2}$$

where k_B is Boltzmann's constant. We assume that the parameters E_L (lowest unoccupied molecular orbital level), σ_n^2 (its variance), and N_{n0} (total density of transport states) are only weakly temperature dependent such that we neglect this weak temperature dependence in our investigations. We set

$$\eta_n = \eta_n(\psi, \varphi_n, T) := \frac{q(\psi - \varphi_n) - E_L}{k_\mathrm{B} T}, \quad s_n = s_n(\sigma_n, T) := \frac{\sigma_n}{k_\mathrm{B} T}.$$

The function $G : \mathbb{R} \times [0, \infty) \to (0, 1)$ is defined by the Gauss–Fermi integral

$$G(\eta_n, s_n) := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(-\frac{\xi^2}{2}\right) \frac{1}{\exp(s_n \xi - \eta_n) + 1}\, \mathrm{d}\xi.$$

According to [11], the mobility $\mu_n = \mu_n(T, n, |\nabla\psi|)$ is a temperature, density and electric field strength dependent function of the form

$$\mu_n(T, n, F) = \mu_{n0}(T) \times g_1(n, T) \times g_2(F, T), \quad \mu_{n0}(T) = \mu_{n0} c_1 \exp\left\{-c_2 s_n^2\right\}. \tag{3}$$

For the considerations of our paper we set $g_1 = g_2 \equiv 1$ and take only the positive temperature feedback μ_{n0} into account. System (1), (2) is closed by mixed boundary conditions for the drift-diffusion system combined with Robin boundary conditions for the heat flow equation modeling a heat sink with fixed temperature T_a.

In [9] the solvability of the bipolar system including the full mobility functions (weak solutions of continuity equations and Poisson equation, entropy solution of the heat flow equation) is established.

In the non-isothermal case, the modeling of (ideal) Ohmic contacts requires local charge neutrality at the contact for the actual temperature dependent state (ψ, φ_n, T). For an applied voltage V, this leads to the nonlinear relation at the contacts $\Gamma_{D_i} \subset \partial\Omega$ for the prescribed value $\psi_0 = \psi - V$:

$$C_{D_i}(\psi; V, T) := C - N_{n0} G\left(\frac{q(\psi - V) - E_L}{k_\mathrm{B} T}; \frac{\sigma_n}{k_\mathrm{B} T}\right) = 0. \tag{4}$$

A straightforward generalization of the computational approach for the isothermal case would result in the necessity to update ψ_0 for each modification of the temperature T, leading to an additional iterative loop for the determination of each bias solution. To avoid this iteration, we use (4) directly as a nonlinear Dirichlet boundary condition for the electrostatic potential ψ depending on T and treat it with the nonlinear solver along with all other nonlinearities.

3 Discretization Scheme

We use a finite volume method and partition the computational domain Ω by a Voronoi mesh with m Voronoi volumes $\{V_l\}_{l=1,\ldots,m}$ and accompanying collocation points $\{x_l\}$. The potentials ψ, φ_n, and the temperature T are evaluated at each node

x_l. The discretized system corresponding to (1) is derived by integrating the equations over each Voronoi volume V_l, applying Gauss's theorem, and then suitably approximating the boundary and volume integrals. We also add the subscript l in all quantities to denote their corresponding numerical values at x_l. In what follows, we will assume that the material parameters ε, μ_{n0}, N_{n0}, and λ are constant, otherwise, suitable averages have to be used.

The resulting surface integrals are split into two parts: integrals over interfaces between two adjacent Voronoi boxes and integrals over boundary parts of the device:

$$\int_{\partial V_l} -\varepsilon\nabla\psi\cdot\nu\,\mathrm{d}\Gamma = \sum_{V_r\in\mathcal{N}(V_l)}\int_{\partial V_l\cap\partial V_r} -\varepsilon\nabla\psi\cdot\nu\,\mathrm{d}\Gamma + \int_{\partial V_l\cap\partial\Omega} -\varepsilon\nabla\psi\cdot\nu\,\mathrm{d}\Gamma,$$

$$\int_{\partial V_l} -j_n\cdot\nu\,\mathrm{d}\Gamma = \sum_{V_r\in\mathcal{N}(V_l)}\int_{\partial V_l\cap\partial V_r} -j_n\cdot\nu\,\mathrm{d}\Gamma + \int_{\partial V_l\cap\partial\Omega} -j_n\cdot\nu\,\mathrm{d}\Gamma,$$

$$\int_{\partial V_l} -\lambda\nabla T\cdot\nu\,\mathrm{d}\Gamma = \sum_{V_r\in\mathcal{N}(V_l)}\int_{\partial V_l\cap\partial V_r} -\lambda\nabla T\cdot\nu\,\mathrm{d}\Gamma + \int_{\partial V_l\cap\partial\Omega} -\lambda\nabla T\cdot\nu\,\mathrm{d}\Gamma.$$

Here $\mathcal{N}(V_l)$ stands for the set of Voronoi volumes V_r which are adjacent to the Voronoi volume V_l. The integrals over interfaces $\partial V_l\cap\partial V_r$ must be treated specifically in order to maintain the consistency of the numerical solution, whereas the surface integrals over $\partial V_l\cap\partial\Omega$ are evaluated by quadrature rules after replacing the normal flux in the integrand by the corresponding boundary condition.

Numerical fluxes through interfaces $\boldsymbol{\partial V_l\cap\partial V_r}$. Whereas the integrals of $-\varepsilon\nabla\psi\cdot\nu$ and $-\lambda\nabla T\cdot\nu$ over the interface $\partial V_r\cap\partial V_l$ are approximated by the conventional finite difference approximations

$$\int_{\partial V_r\cap\partial V_l} -\varepsilon\nabla\psi\cdot\nu\,\mathrm{d}\Gamma \approx \frac{\mathrm{mes}\,(\partial V_r\cap\partial V_l)}{|x_l-x_r|}\varepsilon\,(\psi_l-\psi_r)$$

(similarly for $-\lambda\nabla T\cdot\nu$), the corresponding integrals in the continuity equations require some extra effort

$$\int_{\partial V_r\cap\partial V_l} j_n\cdot\nu\,\mathrm{d}\Gamma \approx \frac{\mathrm{mes}\,(\partial V_r\cap\partial V_l)}{|x_l-x_r|}J_n^{l;r},$$

where the numerical fluxes $J_n^{l;r}$ are determined by a modification of the Scharfetter-Gummel scheme based on averaging of inverse activity coefficients introduced in [6] and discussed with respect to degenerate semiconductors in [4, 5]. We introduce some notation for the definition of the expressions $J_n^{l;r}$:

$$\psi_{l,r} := \frac{\psi_l+\psi_r}{2},\ \varphi_{n;l,r} := \frac{\varphi_{n;l}+\varphi_{n;r}}{2},\ T_{l,r} := \frac{T_l+T_r}{2},\ U_T^{l,r} := \frac{k_\mathrm{B}T_{l,r}}{q},\ s_n^{l,r} := \frac{\sigma_n}{k_\mathrm{B}T_{l,r}},$$

$$\overline{\eta}_{n;l} := \eta_n\left(\psi_l,\varphi_{n;l},T_{l,r}\right),\overline{\eta}_{n;r} := \eta_n\left(\psi_r,\varphi_{n;r},T_{l,r}\right),\ \overline{\eta}_n^{l,r} := \eta_n\left(\psi_{l,r},\varphi_{n;l,r},T_{l,r}\right),$$

$$n^{l,r} := N_{n0}G\left(\overline{\eta}_n^{l,r};s_n^{l,r}\right),\ \mu_n^{l,r} := \mu_{n0}\left(T_{l,r}\right).$$

With the above definitions and the Bernoulli function, $B\left(x\right)=\frac{x}{\exp(x)-1}$, the numerical fluxes $J_n^{l;r}$ have the form

$$J_n^{l;r}=-qN_{n0}\mu_n^{l,r}U_T^{l,r}\frac{G\left(\overline{\eta}_n^{l,r};s_n^{l,r}\right)}{\exp\left(\overline{\eta}_n^{l,r}\right)}\left[\mathrm{e}^{\overline{\eta}_{n;l}}B\left(\frac{\psi_l-\psi_r}{U_T^{l,r}}\right)-\mathrm{e}^{\overline{\eta}_{n;r}}B\left(-\frac{\psi_l-\psi_r}{U_T^{l,r}}\right)\right].$$

For the discretization of the full bipolar model taking into account the complete mobility functions from organics including the factors g_1 and g_2 we refer to [3].

Numerical treatment of the boundary conditions on $\boldsymbol{\partial V_l \cap \partial \Omega}$. The realization of no-flux and Robin boundary conditions is based on the evaluation of the corresponding surface integrals by a midpoint quadrature rule. Dirichlet boundary conditions are implemented using the Dirichlet penalty method: We replace the Dirichlet boundary conditions for φ_n by $j_n\cdot\nu+\Pi(\varphi_n-V)=0$, and treat them like Robin boundary conditions. The penalty parameter Π is a large number which results in marginalizing the normal flux contributions. In order to approximate the nonlinear Dirichlet boundary condition (4), we use a similar idea. We replace (4) by $-\varepsilon\nabla\psi+\Pi C_{D_i}(\psi;V,T)=0$ and treat the resulting boundary condition as a nonlinear Robin boundary condition. Using this approach, the nonlinearity (4) can be treated without any additional iteration along with all the other nonlinearities in the resulting system of equations by the general Newton solver coupled to a parameter embedding scheme.

Volume integrals. For the integral of the charge density $C-n$ we use the midpoint rule. The Joule heat integral H is approximated using the fluxes $J_n^{l;r}$,

$$\begin{aligned}\int_{V_l}(C-n)\,\mathrm{d}x&\approx\operatorname{mes}\left(V_l\right)\left(C_l-n_l\right),\\ \int_{V_l}H\,\mathrm{d}x&\approx\sum_{V_r\in\mathcal{N}(V_l)}\frac{\operatorname{mes}\left(\partial V_l\cap\partial V_r\right)}{2d}\,J_n^{l;r}\left(\varphi_{n;l}-\varphi_{n;r}\right),\end{aligned}\tag{5}$$

where d denotes the space dimension. Here, we followed the idea proposed in [1] and exploited in [7] allowing to evaluate the Joule heating approximation for electrons and holes by edge contributions.

Path-following method for calculation of S-shaped CV-curves. For a device with two Dirichlet boundary parts Γ_{D_1} and Γ_{D_2}, where on Γ_{D_2} the potential is set to zero and on Γ_{D_1} to the (spatially constant) externally applied voltage V, we determine the CV-relation by calculating the current over Γ_{D_1}. Since organic semiconductors show a pronounced electrothermal feedback that can lead to S-shaped CV-relations, a voltage controlled simulation is unable to cover the full characteristic, since at the lower turning point of the S-curve one would not find a point on the curve with increased voltage and only slightly increased current and related temperature, see e.g. Fig. 1a. For such voltage values, only points on the upper branch of the S-curve are available, related to very different current and temperature values. In other words, for increasing voltage, if at all the method would converge, one could only jump to the upper part of the S-curve and the (unstable) region of negative differential resistance of the S-curve is impossible to resolve. Therefore we implemented a path-following method to trace the S-curve. With the discrete equations for all Voronoi boxes $\{V_l\}$ we arrive at a system of $3m$ coupled nonlinear algebraic equations for $u = (\psi_l, \varphi_{n;l}, T_l)_{l=1,\dots,m}$ of the form $F(u, V) = 0$, $F : \mathbb{R}^{3\mathrm{m}} \times \mathbb{R} \to \mathbb{R}^{3m}$. We adapt the technique described in [7, Sect. 5] which was used in [12] to simulate S-shaped CV-relations for organic LEDs resulting from an electrothermal modeling by p-Laplace thermistor models to the drift-diffusion setting.

4 Simulation Results

The finite volume method has been implemented in the prototype semiconductor device simulator ddfermi [2] which is based on the PDE solution toolbox pdelib [8].

We give a proof of concept that electrothermal drift-diffusion models from Sect. 2 can exhibit S-shaped CV-relations and restrict our simulations to a 340 nm thick, uniformly n-doped layer that is contacted by two metal layers. Due to the high conductivity of the metal layers we assume that the potential is constant here and neglect the metal layer entirely. Instead, we prescribe Dirichlet boundary conditions on the parts Γ_{D_1} and Γ_{D_2}. On Γ_{D_2} the potential is set to zero and on Γ_{D_1} to the externally applied voltage V. We determine the CV-relation of the organic device by calculating the current over Γ_{D_1}. We found a parameter range leading to a pronounced occurrence of S-shaped CV-relations. The used parameters are collected in Table 1.

To discuss the phenomenon of S-shaped CV-relations and their appearance in dependence on physical parameters, we present two types of parameter variations. In Fig. 1a, b we study the influence of the disorder parameter σ_n on the electrothermal interaction in the device. The resulting CV-relations are depicted in Fig. 1a, b shows the maximal device temperature over the applied voltage. Whereas for $\sigma_n = 0.05$ eV no S-shaped CV-relation occurs, although such behavior evolves for higher σ_n. With increasing σ_n the first turning point of the curve moves to a higher applied voltage but the related current density decreases and the ‘S’ becomes more pronounced. Figure 1c contains CV-relations for a variation of the thermal outcoupling conditions realized by Robin boundary conditions of the form $\lambda \nabla T \cdot \nu + \kappa (T - T_a) = 0$ on $\partial \Omega$ for different values of κ. Better cooling broadens the ‘S’, for the two turning

Table 1 Simulation parameters

Parameter	Value	Parameter	Value	Parameter	Value
ε	$4.0\,\varepsilon_0$	E_H	0.0 eV	N_{n0}	$10^{21}\,\mathrm{cm}^{-3}$
λ	$0.4\,\mathrm{Wm}^{-1}\mathrm{K}^{-1}$	T_a	220 K	μ_{n0}	$0.8\,\mathrm{cm}^2\mathrm{V}^{-1}\mathrm{s}^{-1}$
κ	$10^3 \dots 10^5\,\mathrm{Wm}^{-2}\mathrm{K}^{-1}$	c_1	1.0	doping	$5 \cdot 10^{18}\,\mathrm{cm}^{-3}$
σ_n	$0.05 \dots 0.08\,\mathrm{eV}$	c_2	0.4	thickness	340 nm

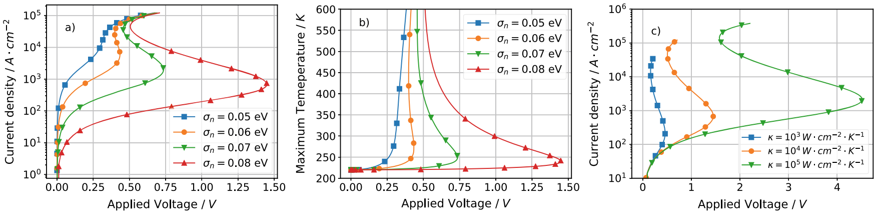


Fig. 1 CV-characteristics using the electrothermal drift-diffusion model for different disorder parameters σ_n **a**, **b** shows the resulting maximal temperature in the device for $\kappa = 10^4$ W/(m^{2}K). **c** Depicts CV-curves for different thermal outcoupling regimes and $\sigma_n = 0.08$ eV

points the applied voltage as well as the current density increase. The exemplary variations of physical parameters show that the complex nonlinear interplay leads to strong variations in the shape of the CV-characteristics.

5 Conclusion and Remarks

We presented a discretization scheme for the electrothermal drift-diffusion model (1) for organic semiconductor devices. We formulated temperature dependent nonlinear Dirichlet boundary conditions for the electrostatic potential (4) at Ohmic contacts, which take into account the shift of the equilibrium potential due to changing device temperature.

We used a finite volume based generalized Scharfetter-Gummel scheme implemented in the prototype semiconductor device simulator `ddfermi` [2] on top of the PDE solver toolbox pdelib [8]. Via a path-following technique, we demonstrated that the model and its discretization for certain parameters exhibit the phenomenon of an S-shaped CV-relation with regions of negative differential resistance. The ability to simulate S-shaped CV-relations using drift-diffusion type electrothermal models is to our knowledge a novelty. Although CV-relations have been observed experimentally in [12], there is a need to be properly modeled in order to understand and optimize the device behavior.

Besides device characteristics, our model (1) and its discretization are capable to describe the spatially resolved electrothermal behavior of real 3D organic semiconductor devices in terms of charge carrier densities, current densities, potentials, temperature distributions. Figure 2 compares the produced Joule heat densities for an organic thin-film transistor with fixed source-drain voltage of 1 V when the channel is opened by raising the gate voltage from 0V (left) to 1V (right).

Simulations for real organic device structures and realistic physical parameters help to estimate the region of a stable working regime guaranteeing the absence of material destruction due to overheating. Furthermore, the description of the spatially resolved electrothermal behavior of real devices is very important for understanding

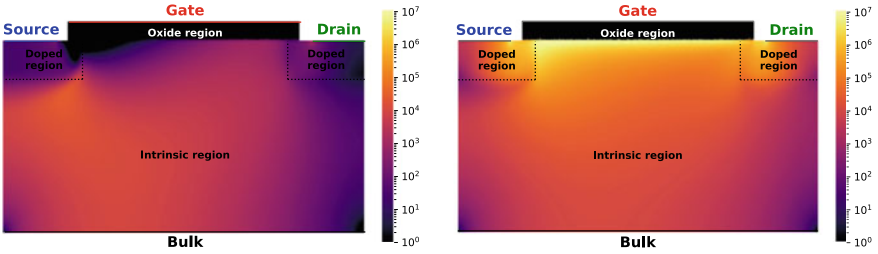


Fig. 2 Simulated Joule heat densities [W/cm^3] (source terms in the heat flow equation) in an organic transistor that demonstrate the change of the electrothermal regime when opening the channel by increasing the gate voltage from 0V (left) to 1V (right) for a fixed source-drain voltage of 1V

the effect of thermal switching, device degradation, device breakdown and local heating (hot spots) in large area devices.

Acknowledgements The authors gratefully acknowledge the funding by the German Research Foundation (DFG) under Germany's Excellence Strategy – The Berlin Mathematics Research Center MATH+ (EXC-2046/1, project ID: 390685689) in project AA2-1 and AA2-6.

References

1. Bradji, A., Herbin, R.: Discretization of coupled heat and electrical diffusion problems by finite-element and finite-volume methods. IMA J. Numer. Anal. **28**, 469–495 (2008)
2. Doan, D.H., Farrell, P., Fuhrmann, J., Kantner, M., Koprucki, T., Rotundo, N.: ddfermi – a drift-diffusion simulation tool (2019). https://doi.org/10.20347/WIAS.SOFTWARE.DDFERMI
3. Doan, D.H., Fischer, A., Fuhrmann, J., Glitzky, A., Liero, M.: Drift-diffusion simulation of S-shaped current-voltage relations for organic semiconductor devices. WIAS-Preprint 2630, Berlin (2019)
4. Farrell, P., Koprucki, T., Fuhrmann, J.: Computational and analytical comparison of flux discretizations for the semiconductor device equations beyond Boltzmann statistics. J. Comput. Phys. **346**, 497–513 (2017)
5. Farrell, P., Rotundo, N., Doan, D., Kantner, M., Fuhrmann, J., Koprucki, T.: Drift-diffusion models. In: Piprek, J. (ed.) Handbook of Optoelectronic Device Modeling and Simulation, chap. 50, vol. 2, pp. 733–771. CRC Press Taylor & Francis (2017)
6. Fuhrmann, J.: Comparison and numerical treatment of generalised Nernst-Planck models. Comput. Phys. Commun. **196**, 166–178 (2015)
7. Fuhrmann, J., Glitzky, A., Liero, M.: Hybrid finite-volume/finite-element schemes for $p(x)$-Laplace thermistor models. In: Cancès, C., Omnes, P. (eds.) Finite Volumes for Complex Applications VIII-Hyperbolic, Elliptic and Parabolic Problems: FVCA 8, Lille, France, June 2017, pp. 397–405. Springer International Publishing, Cham (2017)
8. Fuhrmann, J., Langmach, H., Liero, M., Streckenbach, T., Uhle, M.: pdelib – FVM and FEM toolbox for partial differential equations (2019). http://pdelib.org
9. Glitzky, A., Liero, M., Nika, G.: An existence result for a class of electrothermal drift-diffusion models with Gauss–Fermi statistics for organic semiconductor devices. WIAS-Preprint 2593, Berlin (2019)
10. Klinger, M.P., Fischer, A., Kleemann, H., Leo, K.: Non-linear self-heating in organic transistors reaching high power densities. Sci. Rep. **8**, 9806 (2018)

11. Kordt, P., Bobbert, P., Coehoorn, R., May, F., Lennartz, C., Andrienko, D.: Organic light emitting diodes. In: Piprek, J. (ed.) Handbook of Optoelectronic Device Modeling and Simulation, chap. 15, vol. 1, pp. 473–522. CRC Press Taylor & Francis (2017)
12. Liero, M., Fuhrmann, J., Glitzky, A., Koprucki, T., Fischer, A., Reineke, S.: 3D electrothermal simulations of organic LEDs showing negative differential resistance. Opt. Quantum Electron **49**, 330/1–330/8 (2017)

A Second Order Numerical Scheme for Large-Eddy Simulation of Compressible Flows

B. Gamal, L. Gastaldo, J.-C. Latché, and D. Veynante

Abstract In the context of large eddy simulation of turbulent flows, the control of kinetic energy seems to be an essential requirement for a numerical scheme. We propose in this paper a formally second order non-dissipative scheme dedicated to the numerical simulation of the filtered Naviers-Stokes equations for compressible flows. The spatial discretization is staggered and based on the so-called Marker-And-Cell (MAC) scheme. A MUSCL-like technique is used for convection operators of the mass and the internal energy balance equations in order to preserve the positivity of the density and of the internal energy. Time discretization is performed with the Heun scheme. A kinetic energy conservation identity at discrete level is proved. The good behaviour of the scheme is assessed on the simulation of compressible decaying isotropic turbulence.

Keywords Large eddy simulation · Compressible flows · Explicit scheme

1 Introduction

Large-eddy simulation (LES) has gained a great success in simulating practical flows where the Reynolds numbers are usually very high. In such a method, the large scale fluid motions are computed explicitly from the filtered Navier-Stokes equations, as in DNS, while small-scale effects are modeled.

B. Gamal (✉) · L. Gastaldo · J.-C. Latché
Institut de Radioprotection et de Sûreté Nucléaire (IRSN), Fontenay-aux-Roses, France
e-mail: bassam.gamal@irsn.fr

L. Gastaldo
e-mail: laura.gastaldo@irsn.fr

J.-C. Latché
e-mail: jean-claude.latche@irsn.fr

D. Veynante
Ecole CentraleSupelec, Laboratoire EM2C, CNRS, Paris, France
e-mail: denis.veynante@ecp.fr

R. Klöfkorn et al. (eds.), *Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples*, Springer Proceedings in Mathematics & Statistics 323,
https://doi.org/10.1007/978-3-030-43651-3_60

Let Ω be an open bounded domain of $\mathbb{R}^d$ with $1 \leq d \leq 3$, a flow variable ϕ is decomposed into filtered (large-scale structures) and residual (small structures) terms by means of a filtering operation $\phi = \bar{\phi} + \phi'$ where:

$$\bar{\phi}(\mathbf{x}, t) \equiv \int_{\Omega} G_{\Delta}(\mathbf{r}, \mathbf{x})\phi(\mathbf{x} - \mathbf{r}, t)\, d\mathbf{r} \tag{1}$$

denotes the spatial filtering of ϕ and G_{Δ} is the filter function that determines the scale of the resolved structures. In practice, the filter is usually the grid filter, with the filter width Δ being a measure of local grid size. For compressible flow, the density-weighted (Favre) filtering is applied, i.e., $\widetilde{\phi} = (\overline{\rho\phi})/\overline{\rho}$, $\overline{\rho}$ being the filtered density. When Favre filtered, the spatially filtered Navier-Stokes equations for compressible flows take the form:

$$\partial_t \overline{\rho} + \mathrm{div}(\overline{\rho}\widetilde{\mathbf{v}}) = 0 \tag{2a}$$

$$\partial_t(\overline{\rho}\widetilde{\mathbf{v}}) + \mathrm{div}(\overline{\rho}\widetilde{\mathbf{v}} \otimes \widetilde{\mathbf{v}}) = -\nabla\overline{p} + \mathrm{div}\boldsymbol{\tau} \tag{2b}$$

$$\partial_t(\overline{\rho}\widetilde{e}) + \mathrm{div}(\overline{\rho}\widetilde{\mathbf{v}}\widetilde{e}) + \overline{p}\,\mathrm{div}\widetilde{\mathbf{v}} = \boldsymbol{\tau} : \nabla\widetilde{\mathbf{v}} - \mathrm{div}q \tag{2c}$$

$$\overline{p} = (\gamma - 1)\,\overline{\rho}\widetilde{e} \tag{2d}$$

where t stands for the time, $\overline{p}$, $\widetilde{\mathbf{v}}$ and $\widetilde{e}$ are respectively the filter pressure, velocity and internal energy, γ denotes the heat capacity ratio. Only impermeability conditions are considered for short, and initial conditions $\overline{\rho}_0$, $\widetilde{e}_0$ and $\widetilde{\mathbf{v}}_0$ are such that $\overline{\rho}_0 \geq 0$ and $\widetilde{e}_0 \geq 0$.

The viscous stress tensor could be seen as composed by a computable and an unresolved or subgrid-scale (SGS) part $\boldsymbol{\tau} = \overline{\boldsymbol{\tau}} - \widetilde{\boldsymbol{\sigma}}$. The computable part $\overline{\boldsymbol{\tau}}_{ij}$ is defined as:

$$\overline{\boldsymbol{\tau}}_{ij} = 2\mu\left[\widetilde{\mathbf{S}}_{ij} - \frac{1}{3}\delta_{ij}\sum_k Tr(\widetilde{\mathbf{S}})\right], \quad 1 \leq i, j \leq d \tag{3}$$

where μ is the "computable" turbulent viscosity and $\widetilde{\mathbf{S}}$ is the mean rate-of-strain tensor defined as $\widetilde{\mathbf{S}} = 1/2\left(\nabla\widetilde{\mathbf{v}} + \nabla^t\widetilde{\mathbf{v}}\right)$. The SGS turbulent shear stress $\widetilde{\boldsymbol{\sigma}}$ can not be calculated directly and therefore is modelled in terms of resolved quantities by the Boussinesq's eddy viscosity model:

$$-\widetilde{\sigma}_{ij} = 2\mu_{SGS}\left[\widetilde{\mathbf{S}}_{ij} - \frac{1}{3}\delta_{ij}\sum_k Tr(\widetilde{\mathbf{S}})\right], \quad 1 \leq i, j \leq d \tag{4}$$

where μ_{SGS} is the SGS turbulent viscosity.

Analogously, the heat flux q could be decomposed into a computable part $\widetilde{q}$ and ans SGS part Q:

$$q = \widetilde{q} + Q = \frac{\mu\,\gamma}{\mathrm{Pr}}\nabla\widetilde{e} + \frac{\mu_{SGS}\gamma}{\mathrm{Pr}_t}\nabla\widetilde{e} \tag{5}$$

where Pr and Pr_t are respectively the laminar and the turbulent Prandtl numbers.

The Smagorinsky model is used for the SGS turbulent viscosity computation, in order to close the system:

$$\mu_{SGS} = \rho (C_s \Delta)^2 |\bar{\mathbf{S}}| \tag{6}$$

where C_s is a model parameter.

In the context of LES of turbulent flows, the control of kinetic energy is an essential requirement for a numerical scheme in order to guarantee not only stability but also physical reliability of the results. The aim of this paper is to propose an as less dissipative as possible scheme for the resolution of System (2). The developed scheme is staggered and based on the so-called Marker and Cell (MAC) space discretization. Time-stepping is performed with the Heun scheme. This scheme enjoys some stability properties: the density, the internal energy and the pressure are shown to be non-negative at the discrete level. The conservation of discrete kinetic energy is also proved up to residual terms which may be explicited in Theorem 3.

This paper is organized as follows. The scheme is introduced in Sect. 2, some stability results including kinetic energy identity are given in Sect. 3. Finally, Sect. 4 presents the large eddy simulation of compressible decaying isotropic turbulence.

2 The Numerical Scheme

Mesh and notations—Let Ω be the computational domain (suitable for the discretization by a cartesian grid). A discretization $(\mathcal{M}, \mathcal{E})$ of Ω with a staggered rectangular grid (or MAC grid), involves a primal grid $\mathcal{M}$ which consists in a conforming structured partition of Ω in rectangles ($d = 2$) or rectangular parallelepipeds ($d = 3$), possibly non uniform. A generic cell of this grid is denoted by K. The set of all the edges of this mesh is denoted by $\mathcal{E}$, with $\mathcal{E} = \mathcal{E}_{\text{int}} \cup \mathcal{E}_{\text{ext}}$, where $\mathcal{E}_{\text{int}}$ (resp. $\mathcal{E}_{\text{ext}}$) are the edges of $\mathcal{E}$ that lie in the interior (resp. on the boundary) of the domain. The set of the edges (resp. the internal and boundary edges) that are orthogonal to the ith vector of the orthonormal basis of $\mathbb{R}^d$, $\boldsymbol{e}^{(i)}$, is denoted by $\mathcal{E}^{(i)}$ (resp. $\mathcal{E}^{(i)}_{\text{int}}$ and $\mathcal{E}^{(i)}_{\text{ext}}$), for $1 \leq i \leq d$.

For $\sigma \in \mathcal{E}_{\text{int}}$, we write $\sigma = K|L$ if $\sigma = \partial K \cap \partial L$. The outward normal vector to a face σ of K is denoted by $\boldsymbol{n}_{K,\sigma}$. For $K \in \mathcal{M}$ and $\sigma \in \mathcal{E}$, $|K|$ denotes the measure of K and $|\sigma|$ the $(d-1)$-measure of the face σ. We denote by $d_{K,\sigma}$, $\forall K \in \mathcal{M}$ and $\forall \sigma \in \mathcal{E}$, the Euclidean distance between the center x_K of the cell and the edge σ. We define $d_\sigma = d_{K,\sigma} + d_{L,\sigma}$ if $\sigma \in \mathcal{E}_{\text{int}}$ and $d_\sigma = d_{K,\sigma}$ if $\sigma \in \mathcal{E}_{\text{ext}}$.

A dual cell D_σ associated to an edge $\sigma \in \mathcal{E}$ is defined as follows:

- if $\sigma = K|L \in \mathcal{E}_{\text{int}}$ then $D_\sigma = D_{K,\sigma} \cup D_{L,\sigma}$, where $D_{K,\sigma}$ (resp. $D_{L,\sigma}$) is the half-part of K (resp. L) adjacent to σ (see Fig. 1);
- if $\sigma \in \mathcal{E}_{\text{ext}}$ is adjacent to the cell K, then $D_\sigma = D_{K,\sigma}$.

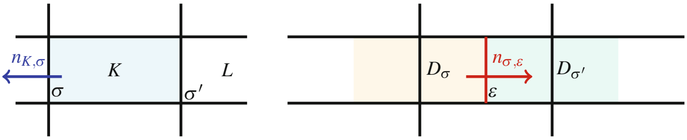


Fig. 1 Notations for control volumes and edges—left: primal mesh, right: dual mesh for the first component of the velocity

For each velocity component i, the domain Ω is thus partitioned in dual cells: $\Omega = \cup_{\sigma\in\mathscr{E}^{(i)}} D_\sigma$. The ith partition is referred to as the ith dual mesh, associated to the ith velocity component, in a sense which is clarified below. The set of the edges of the ith dual mesh is denoted by $\widetilde{\mathscr{E}}^{(i)}$. The dual edge separating two duals cells D_σ and $D_{\sigma'}$ is denoted by $\epsilon = \sigma|\sigma'$. The set of edges of a primal cell K and of a dual cell D_σ are denoted by $\mathscr{E}(K)$ and $\widetilde{\mathscr{E}}(D_\sigma)$ respectively.

The discrete unknowns for the ith component of the velocity are associated to the ith dual mesh and are denoted by $(u_{i,\sigma})_{\sigma\in\mathscr{E}^{(i)}}$. The scalar unknowns (pressure, internal energy, density) are associated to the primal cells and are denoted respectively by $(p_K)_{K\in\mathscr{M}}$, $(e_K)_{K\in\mathscr{M}}$.and $(\rho_K)_{K\in\mathscr{M}}$.

Let notice that, in the following, the filter notations are omitted for the sake of clarity.

Description of the scheme—Let us consider a partition $0 = t_0 < t_1 < \ldots < t_N = T$ of the time interval $(0, T)$, which we suppose uniform, and let $\delta t = t_{n+1} - t_n$ for $n = 0, 1, \cdots, N-1$ be the (constant) time step. The time integration is performed by the second order Heun scheme (which falls in the class of Runge-Kutta schemes), the step n of which may be described throughout three fractional steps described hereafter. The first step reads:

$$\begin{aligned}
&\mathbf{W}^n = (\rho^n,\, e^n,\, p^n,\, \mathbf{v}^n) \text{ being known,}\\
&\quad\textbf{First step}-\text{Compute } \mathbf{W}^{(1)} = (\rho^{(1)},\, e^{(1)},\, p^{(1)},\, \mathbf{v}^{(1)}), \text{ by:}\\
&\quad\quad \mathbf{W}^{(1)} = \mathscr{S}(\mathbf{W}^n) \qquad (7a)
\end{aligned}$$

where the relation $\mathbf{W}^{(1)} = \mathscr{S}(\mathbf{W}^n)$ means that the left-hand side is obtained by applying the standard first-order in time explicit scheme to an initial data given by $\mathbf{W}^n$, which reads:

$$\frac{1}{\delta t}(\rho_K^{(1)} - \rho_K^n) + \mathrm{div}_K(\rho^n\,\mathbf{v}^n) = 0, \quad \forall K \in \mathscr{M} \tag{8a}$$

$$\frac{1}{\delta t}(\rho_K^{(1)} e_K^{(1)} - \rho_K^n e_K^n) + \mathrm{div}_K(\rho^n\, e^n\,\mathbf{v}^n) + p_K^n\,\mathrm{div}_K(\mathbf{v}^n) = (\boldsymbol{\tau}(\mathbf{v}^n) : \nabla\mathbf{v}^n)_K - \mathrm{div}(q)_K, \quad \forall K \in \mathscr{M} \tag{8b}$$

$$\frac{\rho_\sigma^{(1)} v_{\sigma,i}^{(1)} - \rho_\sigma^n v_{\sigma,i}^n}{\delta t} + \mathrm{div}_\sigma(\rho^n \mathbf{v}^n v_i^n) + (\nabla p^n)_{\sigma,i} = \mathrm{div}(\boldsymbol{\tau}(\mathbf{v}^n))_{\sigma,i},\ \forall\sigma \in \mathscr{E}_{\mathrm{int}}^{(i)} \tag{8c}$$

$$p_K^{(1)} = (\gamma - 1)\, \rho_K^{(1)}\, e_K^{(1)}, \quad \forall K \in \mathscr{M} \tag{8d}$$

The terms introduced for each discrete equation will be defined in the following. Note that, to cope with impermeability conditions, the momentum balance equation is not written on the boundary dual cells, and the velocity (in fact, the normal velocity, due to the arrangement of the unknowns) on the boundary edges is just set to zero. The second step of the numerical scheme is analogous to the first one:

$$\textbf{Second step}-\text{Compute } \mathbf{W}^{(2)} = (\rho^{(2)},\, e^{(2)},\, p^{(2)},\, \mathbf{v}^{(2)}), \text{ by:} \tag{9a}$$
$$\mathbf{W}^{(2)} = \mathscr{S}(\mathbf{W}^{(1)})$$

Finally, the last step of the algorithm allows to write the $n+1$ unknows as a linear combination of the n and (2) unknows:

$$\textbf{Last step}-\text{Compute } \rho^{n+1},\, e^{n+1},\, p^{n+1} \text{ and } u_i^{n+1},\ 1 \le i \le d \text{ by:}$$

$$\rho_K^{n+1} = \frac{1}{2}\,(\rho_K^n + \rho_K^{(2)}), \quad \forall K \in \mathscr{M} \tag{10a}$$

$$\rho_K^{n+1}\, e_K^{n+1} = \frac{1}{2}\big(\rho_K^n\, e_K^n + \rho_K^{(2)}\, e_K^{(2)}\big), \quad \forall K \in \mathscr{M} \tag{10b}$$

$$\rho_{D_\sigma}^{n+1}\, v_{i,\sigma}^{n+1} = \frac{1}{2}\big(\rho_{D_\sigma}^n v_{i,\sigma}^n + \rho_{D_\sigma}^{(2)}\, v_{i,\sigma}^{(2)}\big), \quad \forall \sigma \in \mathscr{E}_{\text{int}}^{(i)} \tag{10c}$$

$$p_K^{n+1} = (\gamma - 1)\, \rho_K^{n+1}\, e_K^{n+1} \quad \forall K \in \mathscr{M} \tag{10d}$$

Let us now detail the discrete balance equations involved in (7) (analogously in (9)).

Discrete mass balance—The convection term of Eq. (8a) reads:

$$|K|\, \mathrm{div}_K(\rho\, \mathbf{v}) = \sum_{\sigma \in \mathscr{E}(K)} F_{K,\sigma}, \ \text{ with } F_{K,\sigma} = |\sigma|\, \rho_\sigma\, \mathbf{v}_{K,\sigma} \cdot \boldsymbol{n}_{K,\sigma} \tag{11}$$

where $F_{K,\sigma}$ stands for the mass flux across σ outward K and $\mathbf{v}_{K,\sigma} = v_{\sigma,i}\ \boldsymbol{e}^{(i)}$ for $\sigma \in \mathscr{E}^{(i)}$. The density at the face $\sigma = K|L$, ρ_σ, is approximated by a second order MUSCL-like interpolation: $\forall K \in \mathscr{M}$ and $\forall \sigma \in \mathscr{E}(K) \cap \mathscr{E}_{\text{int}}$, there exists $\alpha_\sigma^K \in [0, 1]$ and $M_\sigma^K \in \mathscr{M}$ such that (see [2] for more details):

$$\rho_\sigma - \rho_K = \left| \begin{array}{ll} \alpha_\sigma^K\,(\rho_K - \rho_{M_\sigma^K}) & \text{if } \mathbf{v}_{K,\sigma}^n \ge 0, \\ \alpha_\sigma^K\,(\rho_{M_\sigma^K} - \rho_K) & \text{otherwise.} \end{array} \right. \tag{12}$$

Discrete internal energy balance—Equation (8b) is an approximation of the internal energy balance over the primal cell K. The convection operator is defined as follows:

$$|K|\, \mathrm{div}_K(\rho\, \mathbf{v}\, e) = \sum_{\sigma \in \mathscr{E}(K)} F_{K,\sigma}\, e_\sigma \tag{13}$$

where the discretization of the internal energy at the primal faces uses the same MUSCL technique as for the density to ensure the positivity of the convection operator.

The viscous dissipation term $(\boldsymbol{\tau}(\mathbf{v}^n) : \nabla \mathbf{v}^n)_K$ and the viscous diffusion term $\mathrm{div}(\boldsymbol{\tau}(\mathbf{v}^n))_{\sigma,i}$ of the momentum balance equation are defined so that they satisfy the following two constraints (see [3] for more details):

- non-negativity of the dissipation: $(\boldsymbol{\tau}(\mathbf{v}^n) : \nabla \mathbf{v}^n)_K \geq 0, \forall K \in \mathscr{M}$;
- consistency of the diffusion and the dissipation, in the following sense:

$$-\sum_{i=1}^{d} \sum_{\sigma \in \mathscr{E}_{\mathrm{int}}^{(i)}} |D_\sigma| \, \mathrm{div}(\boldsymbol{\tau}(\mathbf{v}^n))_{\sigma,i} \; v_{\sigma,i} = \sum_{K \in \mathscr{M}} |K| (\boldsymbol{\tau}(\mathbf{v}^n) : \nabla \mathbf{v}^n)_K \tag{14}$$

i.e., the discrete analogue of the identity $\displaystyle\int_\Omega \mathrm{div}(\boldsymbol{\tau}(\mathbf{v}) \cdot \mathbf{v} = -\int_\Omega \boldsymbol{\tau}(\mathbf{v}) : \nabla \mathbf{v}$.

For the heat diffusion term, the usual finite volume scheme based on a two-point approximation of the fluxes is used, $\forall K \in \mathscr{M}$:

$$\mathrm{div}(q)_K = -\gamma \left(\frac{\mu}{\mathrm{Pr}} + \frac{\mu_{SGS}}{\mathrm{Pr}_t} \right) (\Delta e)_K = \gamma \left(\frac{\mu}{\mathrm{Pr}} + \frac{\mu_{SGS}}{\mathrm{Pr}_t} \right) \sum_{\sigma \in \mathscr{E}(K)} \frac{|\sigma|}{d_\sigma} (e_K - e_L) \tag{15}$$

With this definition, the Laplace operator is monotone [3], i.e.,:

$$\sum_{K \in \mathscr{M}} -\gamma \left(\frac{\mu}{\mathrm{Pr}} + \frac{\mu_{SGS}}{\mathrm{Pr}_t} \right) (\Delta e)_K \; (-e_K^-) \geq 0, \tag{16}$$

where $e_K^- = -\min(e_K, 0)$. This property is necessary to ensure the positivity of the internal energy.

Discrete momentum balance—We now turn to the discrete momentum balance (8c). Following [4], the density on the dual cells is given by the following weighted average:

$$|D_\sigma| \, \rho_{D_\sigma} = |D_{K,\sigma}| \, \rho_K + |D_{L,\sigma}| \, \rho_L, \quad \text{for } \sigma = K|L \in \mathscr{E}(K) \tag{17}$$

The discrete divergence operator on the dual mesh is given by:

$$\mathrm{div}_{D_\sigma}(\rho \, \mathbf{v} \, v_i) = \sum_{\epsilon \in \widetilde{\mathscr{E}}^{(i)}(D_\sigma)} F_{\sigma,\epsilon} v_{\epsilon,i} \tag{18}$$

where $F_{\sigma,\epsilon}$ is the mass flux through the dual face ϵ outward D_σ and the centered choice is made for the approximation of $v_{\epsilon,i}$. The discrete mass flux $F_{\sigma,\epsilon}$ is evaluated as linear combination, with constant coefficients, of the primal mass fluxes at the neighboring faces, in such a way that a discrete mass balance over the dual cells holds [2, 4].

The term $(\nabla p)_{\sigma,i}$ stands for the ith component of the discrete pressure gradient at the face σ. The gradient operator is built as the transpose of the discrete operator for the divergence of the velocity on the primal mesh (i.e., the operator obtained by setting $\rho = 1$ in the $\mathrm{div}_K\,(\rho \mathbf{v})$ operator defined by (11)):

$$(\nabla p)^n_{\sigma,i} = \frac{|\sigma|}{|D_\sigma|}\,(p^n_L - p^n_K)\,\boldsymbol{n}_{K,\sigma}\cdot \boldsymbol{e}^{(i)},\ \text{for}\ \sigma = K|L,\ 1 \le i \le d \tag{19}$$

3 Stability Results

First, we verify that at the discrete level, the numerical scheme preserves the positivity of the density and of the internal energy (and thus of the pressure) [2].

Theorem 1 (Positivity of the density) *Let* $0 \le n \le N-1$*, and let assume that* $\rho^n > 0$ *(i.e., for all* $K \in \mathscr{M}$*,* $\rho^n_K > 0$*) and that the time step satisfies the following condition,* $\forall K \in \mathscr{M}$*:*

$$\delta t \le \min\left[\frac{|K|}{\sum_{\sigma\in\mathscr{E}(K)} |\sigma|\,(1+\alpha^K_\sigma)(v^n_{K,\sigma})^+},\ \frac{|K|}{\sum_{\sigma\in\mathscr{E}(K)} |\sigma|\,(1+\alpha^K_\sigma)(v^{(1)}_{K,\sigma})^+}\right] \tag{20}$$

where, for $a \in \mathbb{R}$*,* $a^+ \ge 0$ *is defined by* $a^+ = \max(a, 0)$ *and* α^K_σ *is introduced in* (12)*. Then a solution to the scheme (7)–(10) satisfies* $\rho^{n+1} > 0$.

Theorem 2 (Positivity of the internal energy) *Let assume that* $e^n > 0$ *(i.e.,* $e^n_K > 0$*,* $\forall K \in \mathscr{M}$*),* $0 \le n \le N-1$*, and that the CFL condition* (20) *holds. In addition, let the the time step satisfy the following condition* $\forall K \in \mathscr{M}$*:*

$$\delta t \le \min\left[\frac{|K|\,\rho^n_K}{(\gamma-1)\,\rho^n_K \sum_{\sigma\in\mathscr{E}(K)} |\sigma|\,(v^n_{K,\sigma})^+ + \sum_{\sigma\in\mathscr{E}(K)} (1+\alpha^K_\sigma)\,(F^n_{K,\sigma})^+},\right.$$
$$\left.\frac{|K|\,\rho^{(1)}_K}{(\gamma-1)\,\rho^{(1)}_K \sum_{\sigma\in\mathscr{E}(K)} |\sigma|\,(v^{(1)}_{K,\sigma})^+ + \sum_{\sigma\in\mathscr{E}(K)} (1+\alpha^K_\sigma)\,(F^{(1)}_{K,\sigma})^+}\right]$$

Then the solution to the scheme (7)–(10) satisfies $e^{n+1} > 0$.

At the continuous level, the kinetic energy balance is obtained by taking the inner product of the momentum balance equation by the velocity and using twice the mass balance equation. At the discrete level, the computation is essentially the same for the convection term, provided that a momentum balance and a mass balance hold on the same cell, which is ensured by construction of the dual densities and mass fluxes ((17) and (18)).

Theorem 3 (Discrete kinetic energy balance) *A solution to the scheme (7)–(10) satisfies the following equality, for* $1 \le i \le d$, $\sigma \in \mathcal{E}_{\mathrm{int}}^{(i)}$ *and* $0 \le n \le N-1$*:*

$$\begin{aligned}
&\frac{1}{2}\,\frac{|D_\sigma|}{\delta t}\Big[\rho_{D_\sigma}^{n+1}(v_{\sigma,i}^{n+1})^2 - \rho_{D_\sigma}^{n}(v_{\sigma,i}^{n})^2\Big] \\
&+ \frac{1}{4}\sum_{\epsilon\in\widetilde{\mathcal{E}}^{(i)}(D_\sigma)} F_{\sigma,\epsilon}^{n} v_{\sigma,i}^{n} v_{\sigma',i}^{n} + |D_\sigma|\,(\nabla p)_{\sigma,i}^{n}\; v_{\sigma,i}^{n} - |D_\sigma|\,\mathrm{div}\boldsymbol{\tau}(\mathbf{v}^{n})_{\sigma,i}\; v_{\sigma,i}^{n} \\
&+ \frac{1}{4}\sum_{\epsilon\in\widetilde{\mathcal{E}}^{(i)}(D_\sigma)} F_{\sigma,\epsilon}^{(1)} v_{\sigma,i}^{(1)} v_{\sigma',i}^{(1)} + |D_\sigma|\,(\nabla p)_{\sigma,i}^{(1)}\; v_{\sigma,i}^{(1)} - |D_\sigma|\,\mathrm{div}\boldsymbol{\tau}(\mathbf{v}^{(1)})_{\sigma,i}\; v_{\sigma,i}^{(1)} = -R_{\sigma,i}^{n+1}
\end{aligned}$$

with

$$\begin{aligned}
R_{\sigma,i}^{n+1} = \frac{1}{4}\frac{|D_\sigma|}{\delta t}\rho_{D_\sigma}^{(2)}\Big[(v_{\sigma,i}^{n+1} - v_{\sigma,i}^{(2)})^2 - (v_{\sigma,i}^{(1)} - v_{\sigma,i}^{(2)})^2\Big] \\
+\frac{1}{4}\frac{|D_\sigma|}{\delta t}\Big[\rho_{D_\sigma}^{n}(v_{\sigma,i}^{n+1} - v_{\sigma,i}^{n})^2 - \rho_{D_\sigma}^{(1)}(v_{\sigma,i}^{n} - v_{\sigma,i}^{(1)})^2\Big]
\end{aligned}$$

The residual terms $R_{\sigma,i}^{n+1}$ may be seen as a numerical dissipation term, which is shown numerically to be second order in time.

4 Numerical Simulation

The scheme under consideration has been developed in the CALIF^{3}S open-source software [1] of the French Institut de Sûreté et de Radioprotection Nucléaire (IRSN).

The test case presented here is the LES of decaying isotropic turbulence.

The numerical simulations are performed in a triply periodic box $[0, 2\pi]$ with 32 cells per axes. The initial velocity field is prescribed using the Random Fourier Method (RFM) that provides a synthetic turbulent velocity field enforcing the Passot-Pouquet model for energy spectrum:

$$E(k) = 16\sqrt{\frac{2}{\pi}}\,\frac{v_{rms}^2}{\kappa_e}\left(\frac{\kappa}{\kappa_e}\right)^4 \exp\left[-2\left(\frac{\kappa}{\kappa_e}\right)^2\right]$$

where κ_e is the wave number at which the most energetic scales occurs and v_{rms} is the root mean square of velocity fluctuations. The initial fields for the internal energy, pressure and density are set uniform. The initial turbulent Mach number is set to $M_t = 0.4$, and the turbulent Reynolds number is set to $Re_T = 2742$.

The time step is computed in order to have a CFL number equal to 1/4. The numerical results are compared to the DNS data of Spyropoulos and Blaisdell [5].

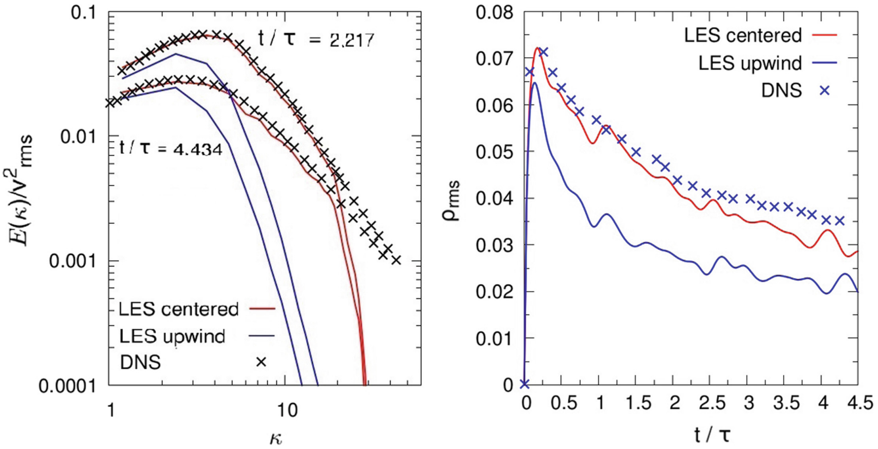


Fig. 2 Energy spectra (left side) and time evolution on density fluctuations (right side). In blue: first-order upwind scheme for comparison

The left part of Fig. 2 shows the comparison between numerical and DNS energy spectra at two different times $t/\tau_t = 2.217$ and $t/\tau = 4.434$, where τ_t is the integral time scale. The right part of Fig. 2 shows the time evolution of density fluctuations. The numerical results are in good agreement with data of literature.

References

1. CALIF^{3}S: A software components library for the computation of fluid flows. https://gforge.irsn.fr/gf/project/califs
2. Gastaldo, L., Herbin, R., Latché, J.C., Therme, N.: A MUSCL-type segregated-explicit staggered scheme for the Euler equations. Comput. Fluids **175**, 91–110 (2018)
3. Grapsas, D.: Staggered fractional step numerical schemes for models for reactive flows. Ph.D. thesis, Aix Marseille University (2017)
4. Herbin, R., Latché, J.C., Nguyen, T.: Consistent segregated staggered schemes with explicit steps for the isentropic and full Euler equations. Math. Model. Numer. Anal. **52**, 893–944 (2018)
5. Spyropoulos, E.T., Blaisdell, G.A.: Evaluation of the dynamic model for simulations of compressible decaying isotropic turbulence. AIAA J. **34**(5), 990–998 (1996)

A Marker-and-Cell Scheme for Viscoelastic Flows on Non Uniform Grids

O. Mokhtari, Y. Davit, J.-C. Latché, R. de Loubens, and M. Quintard

Abstract In this paper, we develop a numerical scheme for the solution of the coupled Stokes and Navier-Stokes equations with constitutive equations describing the flow of viscoelastic fluids. The space discretization is based on the so-called Marker-And-Cell (MAC) scheme. The time discretization uses a fractional-step algorithm where the solution of the Navier-Stokes equations is first obtained by a projection method and then the transport-reaction equation for the conformation tensor is solved by a finite-volume scheme. In order to obtain consistency, the space discretization of the divergence of the elastic part of the stress tensor in the momentum balance equation is derived using a weak form of the MAC scheme. For stability and accuracy reasons, the solution of the transport-reaction equation for the conformation tensor is split into pure convection steps, with a change of variable from $\mathbf{c}$ to $\log(\mathbf{c})$, and a reaction step, which consists in solving one ODE per cell via an Euler scheme with local sub-cycling. Numerical computations for the Stokes flow of an Oldroyd-B fluid in the lid-driven cavity at We = 1 confirm the scheme efficiency.

Keywords Viscoelastic flows · MAC scheme · Projection scheme

O. Mokhtari (✉) · Y. Davit · M. Quintard
Institut de Mécanique des Fluides de Toulouse (IMFT), Université de Toulouse,
Toulouse, France
e-mail: omar.moktari@toulouse-inp.fr

Y. Davit
e-mail: yohan.davit@toulouse-inp.fr

M. Quintard
e-mail: michel.quintard@toulouse-inp.fr

R. de Loubens
Total E&P, CSTJF, Pau, France
e-mail: romain.de-loubens@total.com

J.-C. Latché
Institut de Radioprotection et de Sûreté Nucléaire, Montrouge, France
e-mail: jean-claude.latche@irsn.fr

R. Klöfkorn et al. (eds.), *Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples*, Springer Proceedings in Mathematics & Statistics 323,
https://doi.org/10.1007/978-3-030-43651-3_61

1 Introduction

We consider viscoelastic models for polymeric incompressible liquids. Let Ω be a parallelepiped of $\mathbb{R}^d$, $d \in \{2, 3\}$ and $(0, T)$, $T > 0$, a finite time interval. The fluid is governed by the following system of equations:

$$\rho(\partial_t \mathbf{u} + \xi \mathbf{u} \cdot \nabla \mathbf{u}) = -\nabla p + \operatorname{div} \boldsymbol{\tau}_s(\mathbf{u}) + \operatorname{div} \boldsymbol{\tau}_p, \quad \boldsymbol{\tau}_p = \frac{\eta_p}{\lambda} \mathbf{f}(\mathbf{c})(\mathbf{c} - \mathbf{I_d}), \tag{1a}$$

$$\operatorname{div} \mathbf{u} = 0, \tag{1b}$$

$$\partial_t \mathbf{c} + \mathbf{u} \cdot \nabla \mathbf{c} - (\nabla \mathbf{u})\, \mathbf{c} - \mathbf{c}\, (\nabla \mathbf{u})^t + \frac{1}{\lambda} \mathbf{g}(\mathbf{c})(\mathbf{c} - \mathbf{I_d}) = 0, \tag{1c}$$

where the vector-valued function $\mathbf{u}$ is the velocity of the fluid, p is the pressure, $\boldsymbol{\tau}_s = \eta_s(\nabla \mathbf{u} + (\nabla \mathbf{u})^t)$ is the Newtonian stress tensor for the solvent with η_s its viscosity. The constant coefficients ρ, η_p and λ are the fluid density, the polymer viscosity and the polymer retardation time. The tensor $\boldsymbol{\tau}_p$ is the part of the stress accounting for the presence of polymers and $\mathbf{c}$ is the conformation tensor. The coefficient ξ is zero for the unsteady Stokes equations and $\xi = 1$ for the Navier-Stokes equations. The functions $\mathbf{f}(\mathbf{c})$ and $\mathbf{g}(\mathbf{c})$ depend on the model. For example (see [1] for a review), the Oldroyd-B model is given by $\mathbf{f}(\mathbf{c}) = \mathbf{g}(\mathbf{c}) = \mathbf{I_d}$, and the Fene-CR model corresponds to $\mathbf{f}(\mathbf{c}) = \mathbf{g}(\mathbf{c}) = \frac{b}{b - \operatorname{tr}(\mathbf{c})} \mathbf{I_d}$, with b a real number greater than the space dimension. This system must be complemented by initial conditions for the velocity and the conformation tensor, and by suitable boundary conditions. Here, we suppose for short that the velocity is prescribed over the whole boundary and that the normal velocity vanishes everywhere on the boundary. The dimensionless parameters that characterize these types of flows are the Reynolds number, $Re = \rho U L/(\eta_s + \eta_p)$, and the Weissenberg number, $We = \lambda U/L$, where U and L are the characteristic velocity and length scale.

Here, we develop a numerical scheme for the solution of System (1) based on the following technology. The space discretization is based on the so-called Marker-And-Cell (MAC) scheme. Previous work on MAC schemes for viscoelastic flows can be found in [7], in the context of finite differences and in [4], in the context of finite volumes, both on uniform grids. The time discretization uses a fractional-step algorithm where the solution of the Navier-Stokes equations (1a), (1b) is first obtained by a standard projection method and then the transport-reaction equation for the conformation tensor (1c) is solved by a finite-volume scheme. The development of this scheme faces two essential difficulties. Firstly, we use a weak formulation of (1a) for the discretization of the term $\operatorname{div} \boldsymbol{\tau}_p$, which yields an essential ingredient for the scheme stability and a built-in Lax-Wendroff weak consistency property (see [5]). Secondly, the solution of Equation (1c) requires special care due to the stiffness of the term $(\nabla \mathbf{u})\, \mathbf{c} + \mathbf{c}\, (\nabla \mathbf{u})^t$. In the spirit of [8], the solution procedure for Equation (1c) is split in pure convection steps, with a change of variable from $\mathbf{c}$ to $\log(\mathbf{c})$, and a reaction step, which consists in solving one ODE per cell thanks to the piecewise constant discretization of $\mathbf{c}$. In contrast with [8], these ODEs are solved directly for $\mathbf{c}$, and not $\log(\mathbf{c})$, so as to avoid any artificial introduction of nonlinearities. We further use a local time step for each cell, which ensures the scheme stability and prevents a blow-up of the CPU cost.

2 The Numerical Scheme

Let $\mathscr{M}$ be a MAC mesh (see [6]) of Ω. The discrete pressure and conformation unknowns are associated with the cells of the mesh $\mathscr{M}$ and are denoted by $\{p_K, K \in \mathscr{M}\}$ and $\{\mathbf{c}_K, K \in \mathscr{M}\}$. $\mathscr{E}$ and $\mathscr{E}_{\text{int}}$ are, respectively, the sets of all $(d-1)$-faces σ of the mesh and of the interior faces (*i.e.* the faces which are not included in the boundary). For $1 \leq i \leq d$, we denote by $\mathscr{E}_{\text{int}}^{(i)}$ the subset of the faces that are perpendicular to the ith unit vector of the canonical basis of $\mathbb{R}^d$. The discrete velocity unknowns approximate the normal velocity to the mesh faces. Since the velocity is prescribed on the whole boundary, the degrees of freedom for the ith component of the velocity are associated to $\mathscr{E}_{\text{int}}^{(i)}$ and read $(u_{\sigma,i})_{\sigma\in\mathscr{E}_{\text{int}}^{(i)}}$.

Let us consider a uniform partition $0 = t_0 < t_1 < \ldots < t_N = T$ of $(0, T)$ with a constant time step δt. The pressure correction scheme consists in the following two steps:

Prediction step– Solve for $\tilde{\mathbf{u}}^{n+1}$:

For $1 \leq i \leq d,\ \forall \sigma \in \mathscr{E}_{\text{int}}^{(i)}$,

$$\begin{aligned}\frac{\rho}{\delta t}\,\left(\tilde{u}_{\sigma,i}^{n+1} - u_{\sigma,i}^{n}\right) + \xi\ \rho \operatorname{div}_\sigma(\tilde{u}_i^{n+1}\mathbf{u}^n) - \operatorname{div}_{\sigma,i}\boldsymbol{\tau}_s(\tilde{\mathbf{u}}^{n+1})\\ -\operatorname{div}_{\sigma,i}\boldsymbol{\tau}_p^n + \boldsymbol{\nabla}_{\sigma,i}\,(p^n) = 0.\end{aligned} \tag{2a}$$

Correction step– Solve for p^{n+1} and $\mathbf{u}^{n+1}$:

$$\text{For } 1 \leq i \leq d,\ \forall \sigma \in \mathscr{E}_{\text{int}}^{(i)},\quad \frac{\rho}{\delta t}\,(u_{\sigma,i}^{n+1} - \tilde{u}_{\sigma,i}^{n+1}) + \boldsymbol{\nabla}_{\sigma,i}(p^{n+1} - p^n) = 0, \tag{2b}$$

$$\forall K \in \mathscr{M},\quad \operatorname{div}_K(\mathbf{u}^{n+1}) = 0. \tag{2c}$$

In the prediction step, the tensor $\boldsymbol{\tau}_p^n$ is computed as a function of the conformation tensor by $\boldsymbol{\tau}_{p_K}^n = \frac{\eta_{p_K}}{\lambda_K}\,\mathbf{f}(\mathbf{c}_K^n)\,(\mathbf{c}_K^n - \mathbf{I_d})$, for $K \in \mathscr{M}$.

The discretization of the constitutive equation (1c) is split into pure advection steps and a local ODE, which is a strategy already adopted in [8]. This allows us to preserve the positivity of $\mathbf{c}$ and obtain good accuracy. Furthermore, we use a change of variables for the advection steps and change the conformation tensor into the matrix logarithm of the conformation tensor [4]. The result is the following "Strang-log" scheme:

Advection I– Solve for $\mathbf{c}^{n+\frac{1}{3}}$:

$$\forall K \in \mathscr{M}, \quad \frac{1}{\delta t/2}\left(\log\, \mathbf{c}_K^{n+\frac{1}{3}} - \log\, \mathbf{c}_K^{n}\right) + \mathrm{div}_K(\mathbf{u}^{n+1} \log\, \mathbf{c}_K^{n+\frac{1}{3}}) = 0, \tag{3a}$$

ODE– Set $\mathbf{c}_K(t_n) = \mathbf{c}_K^{n+\frac{1}{3}}$ and solve for $\mathbf{c}^{n+\frac{2}{3}} = \mathbf{c}_K(t_n + \delta t)$:

$$\forall K \in \mathscr{M}, \quad \partial_t \mathbf{c}_K - (\boldsymbol{\nabla}_K \mathbf{u}^{n+1})\, \mathbf{c}_K - \mathbf{c}_K\, \left(\boldsymbol{\nabla}_K \mathbf{u}^{n+1}\right)^t + \frac{1}{\lambda_K}\mathbf{g}(\mathbf{c}_K)(\mathbf{c}_K - \mathbf{I_d}) = 0, \tag{3b}$$

Advection II–Solve for $\mathbf{c}^{n+1}$:

$$\forall K \in \mathscr{M}, \quad \frac{1}{\delta t/2}\left(\log\, \mathbf{c}_K^{n+1} - \log\, \mathbf{c}_K^{n+\frac{2}{3}}\right) + \mathrm{div}_K(\mathbf{u}^{n+1} \log\, \mathbf{c}_K^{n+1}) = 0. \tag{3c}$$

Transport steps are discretized by a standard first-order upwind scheme. The local ODE (3b) is solved using a first-order Euler scheme, with a local sub-time step. Two versions are tested: the fully implicit scheme and a version where the term $(\boldsymbol{\nabla}_K \mathbf{u}^{n+1})\, \mathbf{c}_K + \mathbf{c}_K\, \left(\boldsymbol{\nabla}_K \mathbf{u}^{n+1}\right)^t$ is explicit, while the other ones are still implicit. From a theoretical point of view, both variants seem to have the same stability properties, *i.e.* to preserve the positive definite character of the conformation tensor for a small enough (sub-) time step, depending on the velocity gradient. However, numerical tests show that the implicit version is more stable.

Most discrete operators involved in the scheme are standard and we refer to [6] for their detailed definition. We focus in the next section on the discretization of the divergence of the stress tensor in the momentum balance equation.

3 The Total Stress Divergence Term

The aim of this section is to define the divergence term $\mathrm{div}_{\sigma,i}(\mathbf{T})$ of the total Cauchy stress tensor $\mathbf{T} = -p\mathbf{I_d} + \boldsymbol{\tau}_s(\tilde{\mathbf{u}}) + \boldsymbol{\tau}_p$. We want this quantity to satisfy a discrete analogue of the identity:

$$\int_\Omega \mathrm{div}(\mathbf{T}) \cdot \mathbf{u} = -\int_\Omega \mathbf{T}(\mathbf{u}) : \boldsymbol{\nabla}\mathbf{u}. \tag{4}$$

This relation is crucial to derive a scheme that preserves a free energy estimate at the discrete level [2]. In addition, if the discrete gradient of the interpolation of a regular function converges to the continuous gradient in L^∞-weak $\star$, which is the case here (with, in fact, a strong L^∞ convergence), then the identity (4) readily yields the Lax-Wendroff consistency of the discretization of the term $\mathrm{div}_{\sigma,i}\,(\boldsymbol{\tau}_p^n)$. The strategy to obtain (4), already used in [6] for Newtonian fluids, is to recast the MAC scheme under a weak form. For clarity, we only address the two-dimensional case here. The extension to the three-dimensional case is presented in [6].

The discrete velocity gradient—Here, we detail the discretization of terms associated to the x-component of the velocity, using the notations of Fig. 1. Inside the computational domain, the discrete partial derivatives of this velocity component are defined as follows:

- Let the primal cells be denoted by $K_{i,j} = (x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}) \times (y_{j-\frac{1}{2}}, y_{j+\frac{1}{2}})$. The discrete derivative involved in the divergence (so, for the velocity x-component, only $\partial_x^{\mathcal{M}} u^x$) is defined over the primal cell by, $\forall \mathbf{x} \in K_{i,j}$:

$$\partial_x^{\mathcal{M}} u^x(\mathbf{x}) = \frac{u^x_{i+\frac{1}{2},j} - u^x_{i-\frac{1}{2},j}}{h_i^x}. \tag{5}$$

- For the other derivatives (so, for the velocity x-component, only $\partial_y^{\mathcal{M}} u^x$), we introduce a fourth mesh which is vertex-centred, and we denote by K^{xy} the generic cell of this new mesh, with $K^{xy}_{i-\frac{1}{2},j-\frac{1}{2}} = (x_{i-1}, x_i) \times (y_{j-1}, y_j)$. Then, $\forall \mathbf{x} \in K^{xy}_{i-\frac{1}{2},j-\frac{1}{2}}$:

$$\partial_y^{\mathcal{M}} u^x(\mathbf{x}) = \frac{u^x_{i-\frac{1}{2},j} - u^x_{i-\frac{1}{2},j-1}}{h^y_{j-\frac{1}{2}}}. \tag{6}$$

The only necessary extension of this definition to cope with boundaries concerns the definition of $\partial_y^{\mathcal{M}} u^x$ over a half vertex-centered cell associated with a vertex lying on a horizontal boundary. In this case, we use the usual "fictitious cell trick" in order to apply Relation (6): an external cell, of zero y-dimension, is added to the mesh and the horizontal velocity in this cell is set to the prescribed Dirichlet value, or to zero for the test functions defined below. Extending these definitions to the y-component of the velocity, the discrete diffusion tensor can be defined as:

$$\nabla^{\mathcal{M}} \tilde{\mathbf{u}} = \begin{bmatrix} \partial_x^{\mathcal{M}} \tilde{u}^x & \partial_y^{\mathcal{M}} \tilde{u}^x \\ \partial_x^{\mathcal{M}} \tilde{u}^y & \partial_y^{\mathcal{M}} \tilde{u}^y \end{bmatrix}, \quad \tau^{\mathcal{M}}(\tilde{\mathbf{u}}) = \eta_s \left(\nabla^{\mathcal{M}} \tilde{\mathbf{u}} + (\nabla^{\mathcal{M}} \tilde{\mathbf{u}})^t \right). \tag{7}$$

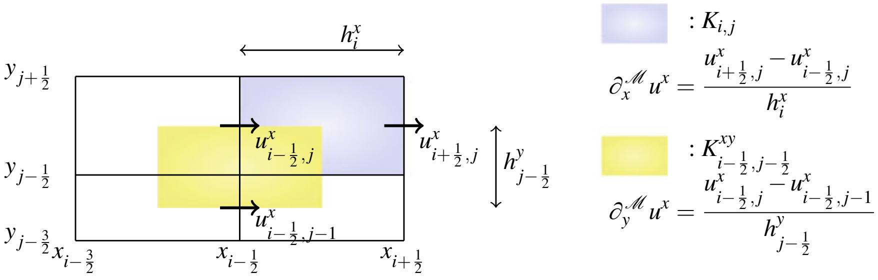

h_i^x
$y_{j+\frac{1}{2}}$
$y_{j-\frac{1}{2}}$
$y_{j-\frac{3}{2}}$
$x_{i-\frac{3}{2}}$
$x_{i-\frac{1}{2}}$
$x_{i+\frac{1}{2}}$
$u^x_{i-\frac{1}{2},j}$
$u^x_{i+\frac{1}{2},j}$
$u^x_{i-\frac{1}{2},j-1}$
$h^y_{j-\frac{1}{2}}$
: $K_{i,j}$
$\partial_x^{\mathcal{M}} u^x = \frac{u^x_{i+\frac{1}{2},j} - u^x_{i-\frac{1}{2},j}}{h_i^x}$
: $K^{xy}_{i-\frac{1}{2},j-\frac{1}{2}}$
$\partial_y^{\mathcal{M}} u^x = \frac{u^x_{i-\frac{1}{2},j} - u^x_{i-\frac{1}{2},j-1}}{h^y_{j-\frac{1}{2}}}$

Fig. 1 Discrete partial derivatives of the x-component of the velocity

Finite-volume test functions— Let us denote by $\mathscr{I}^x \subset \mathbb{N}^2$ (resp. $\mathscr{I}^y \subset \mathbb{N}^2$) the set of pairs (i, j) such that $\mathbf{x}_{i-\frac{1}{2},j}$ (resp. $\mathbf{x}_{i,j-\frac{1}{2}}$) is the mass center of a vertical (resp. horizontal) face of the mesh. For $(i, j) \in \mathscr{I}^x$, we denote by $\boldsymbol{\phi}^{x(i-\frac{1}{2},j)}$ the test function associated with the degree of freedom of the x-component of the velocity located at $\mathbf{x}_{i-\frac{1}{2},j}$. This discrete function is defined by:

$$(\boldsymbol{\phi}^{x,(i-\frac{1}{2},j)})^x_{k-\frac{1}{2},\ell} = \delta^i_k\, \delta^j_\ell,\ \forall (k,\ell) \in \mathscr{I}^x \text{ and } (\boldsymbol{\phi}^{x,(i-\frac{1}{2},j)})^y_{k,\ell-\frac{1}{2}} = 0,\ \forall (k,\ell) \in \mathscr{I}^y.$$

Its non-zero partial derivatives are $\partial^{\mathscr{M}}_x \boldsymbol{\phi}^{x,(i-\frac{1}{2},j)}$ and $\partial^{\mathscr{M}}_y \boldsymbol{\phi}^{x,(i-\frac{1}{2},j)}$ and are given by (5) and (6), respectively. Since the velocity is prescribed on the boundary, no equation is written on the half-dual cells associated to external faces, so no definition is required for the corresponding test functions.

Discrete viscous diffusion and pressure gradient— The discrete divergence of the stress tensor for the solvent is defined by the following weak formulation:

$$\forall (i,j) \in \mathscr{I}^x, \quad -(\mathrm{div}\boldsymbol{\tau}_s(\tilde{\mathbf{u}}))^x_{i-\frac{1}{2},j} = \frac{1}{|K^x_{i-\frac{1}{2},j}|} \int_\Omega \boldsymbol{\tau}^{\mathscr{M}}_s(\tilde{\mathbf{u}}) : \nabla^{\mathscr{M}} \boldsymbol{\phi}^{x,(i-\frac{1}{2},j)}. \qquad (8)$$

Similarly, identifying p with its associated piecewise constant function, we have for the pressure gradient:

$$\forall (i,j) \in \mathscr{I}^x, \quad (\nabla p)^x_{i-\frac{1}{2},j} = \frac{-1}{|K^x_{i-\frac{1}{2},j}|} \int_\Omega p\ \partial^{\mathscr{M}}_x \boldsymbol{\phi}^{x,(i-\frac{1}{2},j)} \qquad (9)$$

It is shown in [6] that Equation (8) yields the usual finite-volume formulation of the MAC scheme. The same holds for the definition (9) of the pressure gradient.

Polymeric stress tensor divergence—This formulation naturally extends to the discretization of the divergence of the polymeric stress tensor. To do so, we first associate the discrete polymeric stress $\boldsymbol{\tau}_p$ to a piecewise function over the primary cells by:

$$\forall \mathbf{x} \in K_{i,j}, \quad \boldsymbol{\tau}_p(\mathbf{x}) = \frac{\eta_p}{\lambda} \mathbf{f}(\mathbf{c}_{i,j})(\mathbf{c}_{i,j} - \mathbf{I_d}).$$

Then we set:

$$\forall (i,j) \in \mathscr{I}^x, \quad -(\mathrm{div}\boldsymbol{\tau}_p)^x_{i-\frac{1}{2},j} = \frac{1}{|K^x_{i-\frac{1}{2},j}|} \int_\Omega \boldsymbol{\tau}_p : \nabla^{\mathscr{M}} \boldsymbol{\phi}^{x,(i-\frac{1}{2},j)}. \qquad (10)$$

An easy computation shows that this relation may be recast as a finite volume formulation, in the sense that the right-hand side may be seen as a sum over the faces of $K^x_{i-\frac{1}{2},j}$ of a discretization of the flux associated to $\mathrm{div}\boldsymbol{\tau}_p$, *i.e.* the integral of the first component of $\boldsymbol{\tau}_p\, \mathbf{n}_{K,\sigma}$. However, as usual when such a duality technique is used, the approximation of the tensor at the horizontal faces may seem strange: indeed, it is a convex combination of the unknown in the two neighbouring cells, but with coefficients which are not those which would be given by a linear interpolation.

4 Numerical Tests

We compare the proposed scheme to results from the literature for the flow of an Oldroyd-B fluid in lid-driven cavity with a Weissenberg number equal to 1. The computational domain is $\Omega = (0, 1)^2$ and the velocity is prescribed on the whole boundary: $\mathbf{u} = \big(8\,x^2\,(1-x)^2\,\big(1+\tanh(8t-4)\big),\ 0\big)^t$ on $(0, 1) \times \{1\}$, $\mathbf{u} = 0$ otherwise. The fluid is initially at rest and the conformation tensor is set to identity. The computation is performed up to $t = 30$. The constant coefficients in System (1) are set to $\rho = 1$, $\eta_s = 0.5$, $\eta_p = 0.5$ and $\lambda = 1$. We use a sequence of successively refined meshes: the coarsest three ones are uniform 64×64, 128×128 and 256×256 cells; the four other ones, denoted by Mn, $n = 1, \ldots, 4$, use a uniform step equal to $1/(256\,n)$ in the x-direction and a splitting in the y-direction with a first step equal to $0.004/n$, a last step equal to $0.001/n$ and a constant ratio between two consecutive steps. The number of cells for the M4 mesh is close to 5.2 million. The sub-time-step for the solution of the ODE (3b) is set to $\delta t/n_e$ with n_e the smallest integer number such that $\delta t/n_e \leq 1/(2\,m\ ||\boldsymbol{\nabla}_K \mathbf{u}^{n+1}||_\infty)$, with $m = 10$ for the uniform meshes and $m = 200$ (to force convergence) for the Mn meshes; this time-step is small enough to preserve the positive definiteness of the conformation tensor when solving (3b) by a backward Euler scheme. Computations are run (in parallel for Meshes Mn) with the open-source CALIF^{3}S software developed at IRSN [3]. The CPU-time used for the solution of the ODE remains almost negligible (less than 3% of the total time), so a more sophisticated algorithm would not enhance the scheme efficiency.

We first describe the results obtained with the three coarsest meshes, with a time-step equal to 0.01. In any case, computations reach a steady state. For the first component of the velocity along the line $x = 0.5$ (Fig. 2, left), the steady state values

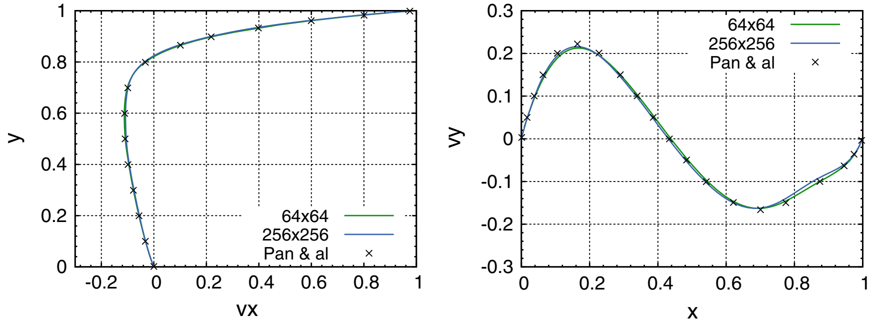


Fig. 2 Left: first component of the velocity along the line $x = 0.5$ -Right: second component of the velocity along the line $y = 0.75$

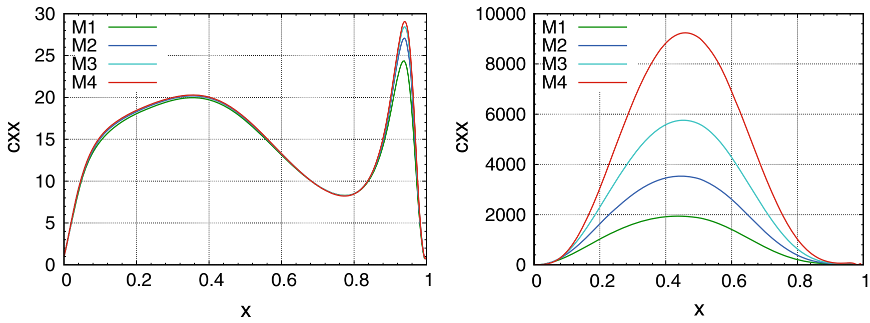


Fig. 3 Conformation tensor component $\mathbf{c}_{xx}$ along the line $y = 0.975$ (left) and $y = 1$ (right)

are almost independent from the mesh, and in close agreement with those given in [8]. The convergence for the second component of the velocity along the line $y = 0.75$ (Fig. 2, right) is a little bit slower: in the eyeball norm, convergence is obtained with the 256×256 mesh and the solution slightly differs from [8]. The most difficult point of this computation consists in obtaining an accurate estimation of the conformation tensor near the lid, and we investigate this issue with the Mn meshes. First of all, we observe that the time-step must be considerably reduced to obtain a stationary solution: $\delta t = 0.001$ for the $M1$, $M2$ and $M3$ meshes, and $\delta t = 0.0005$ for the $M4$ mesh. With a larger times-step, low-frequency instabilities (period in range of 1s) develop from the top-right corner, and remain confined in an area very close to the lid and included in the right half of the cavity. We plot in Fig. 3 the computed value of $\mathbf{c}_{xx}$ along the lines $y = 0.975$ and $y = 1$. At $y = 0.975$, convergence seems to be almost achieved. The picture is completely different at $y = 1$: first, the profile of $\mathbf{c}_{xx}$ dramatically changes from $y = 0.975$; second, the maximum value, obtained close to $x = 0.5$, increases when refining the mesh (multiplication by a 1.6 factor when dividing the space step by 2 for $M2$, $M3$ and $M4$).

References

1. Bird, R.B., Wiest, J.M.: Constitutive equations for polymeric liquids p. 25
2. Boyaval, S., Lelièvre, T., Mangoubi, C.: Free-energy-dissipative schemes for the Oldroyd-B model. Math. Model. Numer. Anal. **43**, 523–561 (2009)
3. CALIF^{3}S: A software components library for the computation of fluid flows. https://gforge.irsn.fr/gf/project/califs
4. Fattal, R., Kupferman, R.: Time-dependent simulation of viscoelastic flows at high Weissenberg number using the log-conformation representation (2005)
5. Gallouët, T., Herbin, R., Latché, J.C., Mallem, K.: Convergence of the MAC scheme for the incompressible Navier-Stokes equations. Found. Comput. Math. **18**(1), 249–289 (2018)
6. Grapsas, D., Herbin, R., Kheriji, W., Latché, J.C.: An unconditionally stable staggered pressure correction scheme for the compressible Navier-Stokes equations. SMAI J. Comput. Math. **2**, 51–97 (2016)

7. Oishi, C., Martins, F., Tom, M., Cuminato, J., McKee, S.: Numerical solution of the extended pom-pom model for viscoelastic free surface flows. J. Non-Newton. Fluid Mech. **166**(3), 165–179 (2011)
8. Pan, T.W., Hao, J., Glowinski, R.: On the simulation of a time-dependent cavity flow of an Oldroyd-B fluid. Int. J. Numer. Methods Fluids **60**, 791–808 (2009)

A Numerical Convergence Study of Some Open Boundary Conditions for Euler Equations

C. Colas, M. Ferrand, J.-M. Hérard, Olivier Hurisse, E. Le Coupanec, and Lucie Quibel

Abstract We discuss herein the suitability of some open boundary conditions. Considering the Euler system of gas dynamics, we compare approximate solutions of one-dimensional Riemann problems in a bounded sub-domain with the restriction in this sub-domain of the exact solution in the infinite domain. Assuming that no information is known from outside of the domain, some basic open boundary condition specifications are given, and a measure of the L^1-norm of the error inside the computational domain enables to show consistency errors in situations involving outgoing shock waves, depending on the chosen boundary condition formulation. This investigation has been performed with Finite Volume methods, using approxi-

C. Colas (✉) · M. Ferrand (✉) · J.-M. Hérard (✉) · O. Hurisse · E. Le Coupanec · L. Quibel
EDF R&D, MFEE, 6 quai Watier, 78400 Chatou, France
e-mail: clement.colas@edf.fr

M. Ferrand
e-mail: martin.ferrand@edf.fr

J.-M. Hérard
e-mail: jean-marc.herard@edf.fr

O. Hurisse
e-mail: olivier.hurisse@edf.fr

E. Le Coupanec
e-mail: erwan.lecoupanec@edf.fr

L. Quibel
e-mail: lucie.quibel@edf.fr

C. Colas · J.-M. Hérard
Aix-Marseille Université, I2M, UMR CNRS 7373, 39 rue Joliot Curie,
13453 Marseille, France

M. Ferrand
CEREA Lab (Ecole des Ponts ParisTech - EDF R&D), 6-8 avenue Blaise Pascal,
Cité Descartes, 77420 Champs-sur-Marne, France

L. Quibel
Université de Strasbourg, IRMA, UMR CNRS 7501, 7 rue René Descartes,
67084 Strasbourg, France

R. Klöfkorn et al. (eds.), *Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples*, Springer Proceedings in Mathematics & Statistics 323,
https://doi.org/10.1007/978-3-030-43651-3_62

mate Riemann solvers in order to compute numerical fluxes for inner interfaces and boundary interfaces.

Keywords Finite volumes · Approximate Riemann solver · Open boundary conditions · Euler equations · Compressible flow

MSC (2010) 65M08 · 65N08 · 76N15

1 Introduction

Concerning computational fluid dynamics, industrial simulations are frequently performed with a partial or total unknown fluid state outside of the computational domain. How are boundary conditions dealt with when no information is known outside? Here the one-dimensional Euler equations governing inviscid compressible fluid flows are considered. The unknowns ρ, u, P respectively denote the density, the velocity and the pressure of the fluid, while the momentum is $Q = \rho u$. The total energy E is such that $E = \rho\left(\frac{u^2}{2} + \varepsilon\right)$. The internal energy $\varepsilon(P, \rho)$ is prescribed by the EOS (Equation Of State). In the sequel, we denote by $\boldsymbol{W} = (\rho, Q, E)^t$ the conservative variable, $\boldsymbol{Y} = (s, u, P)^t$ the non-conservative variable, with s the entropy, and $\boldsymbol{F}(\boldsymbol{W}) = (Q, Qu + P, (E + P)\,u)^t$ the flux function, so that the set of governing equations reads:

$$\partial_t \boldsymbol{W} + \partial_x \boldsymbol{F}(\boldsymbol{W}) = 0. \tag{1}$$

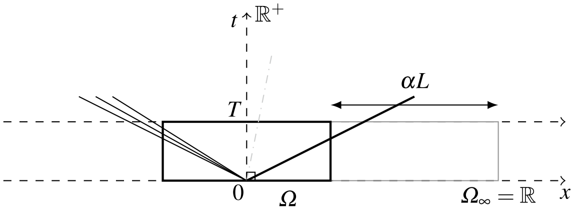
The speed of sound, denoted by c, is such that $c^2 = \left(\frac{P}{\rho^2} - \frac{\partial\varepsilon(P,\rho)}{\partial\rho}\right) / \left(\frac{\partial\varepsilon(P,\rho)}{\partial P}\right)$.

There exists a huge literature on open boundary problems [6, 10–12]. Among these, one pioneering work on boundary conditions for bounded domain may be found in [1]. Actually, the present work addresses the issue of open numerical boundary conditions to get waves outside of the computational domain and can be connected to the work of [7]. The solution of Euler system (1) is sought in $\mathbb{R} \times (0, T)$, with time $T \in \mathbb{R}_+^*$, without boundary conditions, see [14]. This solution, expected to be known and unique, is denoted by $\boldsymbol{W}_{\Omega_\infty}^{exact}(x, t)$ for $(x, t) \in \mathbb{R} \times (0, T)$.

In contrast, the numerical approximations, denoted by $\boldsymbol{W}_{\Omega}^{\Delta x, \Delta t}(x, t)$ for $(x, t) \in \Omega \times (0, T)$, are performed in a bounded computational sub-domain $\Omega \subsetneq \Omega_\infty$ (see Fig. 1) with prescribed open inlet/outlet boundary conditions on $\partial\Omega$.

For this purpose, artificial boundaries are introduced on $\partial\Omega$. Then, numerical boundary conditions, depending on the time and space steps, must be prescribed on $\partial\Omega$. When $(\Delta x, \Delta t) \to (0, 0)$, we assume that some (unique) converged approximation, denoted by $\boldsymbol{W}_{\Omega}^{0,0}(x, t)$ for $(x, t) \in \Omega \times (0, T)$, is obtained. Eventually, we wonder whether $\boldsymbol{W}_{\Omega}^{0,0}(x, t)$ for $(x, t) \in \Omega \times (0, T)$, coincides with the restriction

Fig. 1 Bounded computational domain $\Omega \subsetneq \Omega_\infty$, with Ω_∞ a spatial infinite domain



of the exact solution to Ω, $\boldsymbol{W}^{exact}_{\Omega_\infty}(x,t)$ for $(x,t) \in \Omega \times (0,T)$, or not. In the latter case, the converged approximation $\boldsymbol{W}^{0,0}_{\Omega}$ will be said to be **non-consistent**.

For the Euler system (1), a measure of a subsonic state in the last inner cell N (eigenvalues $\lambda_1(\mathbf{W}^n_N) < 0$ and $\lambda_{2,3}(\mathbf{W}^n_N) > 0$) at a right outlet will require one scalar external information, whereas in the supersonic case ($\lambda_{1,2,3}(\mathbf{W}^n_N) > 0$), the upwind state will be privileged. Actually, we recall that in the subsonic case, the approach of [4, 5] may provide some way to cope with the lack of information.

A first drawback of the latter approach is that the sign of eigenvalues may easily change: signs of eigenvalues $\lambda_k(\mathbf{W}^n_N)$ are not necessarily representative of what happens really at the right boundary when computing true waves associated with the 1D Riemann problem with the initial condition: $\boldsymbol{W}_L = \mathbf{W}^n_N$ and $\boldsymbol{W}_R = \mathbf{W}^n_{ext}$ (unless when $\mathbf{W}^n_{ext} = \mathbf{W}^n_N$). A very instructive example is given in [7] Sect. 3.2, while restricting on a scalar problem (Burgers equation). A second question is: assuming that nothing is known about the exterior state $\mathbf{W}^n_{ext}$, how does the solution, inside the computational sub-domain, depend on the choice of $\mathbf{W}^n_{ext}$?

Herein, the aim consists in testing suitable numerical boundary conditions in the sense that they converge towards the—not necessarily regular—exact solution.

2 Finite Volume Method

We briefly recall the basis of the explicit finite volume scheme VFRoe-ncv, an approximate Godunov scheme using non conservative variables [8, 9]. For the sake of simplicity, regular meshes of the one-dimensional computational domain are considered of size $\Delta x = x_{i+1/2} - x_{i-1/2}$, $i \in \{1, ..., N\}$, and $\Delta t^n = t^{n+1} - t^n$ is the time step, $n \in \mathbb{N}$. The time step is given by some CFL condition in order to gain stability. Let $\mathbf{W}^n_i$ be an approximation of the mean value $\dfrac{1}{\Delta x}\displaystyle\int_{x_{i-1/2}}^{x_{i+1/2}} \boldsymbol{W}(x, t^n)dx$. Time-space integration of system (1) over $\left[x_{i-1/2}, x_{i+1/2}\right] \times \left[t^n, t^{n+1}\right]$ provides the standard following scheme:

$$\Delta x(\mathbf{W}^{n+1}_i - \mathbf{W}^n_i) + \Delta t^n \left(\mathbf{g}^n_{i+\frac{1}{2}} - \mathbf{g}^n_{i-\frac{1}{2}}\right) = 0, \tag{2}$$

where $\mathbf{g}^n_{i+1/2}$ is the numerical flux through the interface $\left\{x_{i+1/2}\right\} \times \left[t^n, t^{n+1}\right]$. For so-called spatially first-order scheme, $\mathbf{g}^n_{i+1/2} = \mathbf{g}(\mathbf{W}^n_i, \mathbf{W}^n_{i+1})$. The numerical flux $\mathbf{g}^n_{i+1/2}$ is obtained by solving the linearized Riemann problem:

$$\begin{cases} \partial_t \boldsymbol{Y} + \boldsymbol{B}(\widetilde{\boldsymbol{Y}})\partial_x \boldsymbol{Y} = 0, \\ \boldsymbol{Y}(x, t^n) = \begin{cases} \mathbf{Y}^n_i & \text{if } x < x_{i+\frac{1}{2}}, \\ \mathbf{Y}^n_{i+1} & \text{if } x > x_{i+\frac{1}{2}}, \end{cases} \end{cases} \tag{3}$$

where $\widetilde{\boldsymbol{Y}} = (\mathbf{Y}^n_i + \mathbf{Y}^n_{i+1})/2$ and $\boldsymbol{B}(\boldsymbol{Y})$ stands for the following matrix:

$$\boldsymbol{B}(\boldsymbol{Y}) = (\partial_{\boldsymbol{Y}} \boldsymbol{W})^{-1} \, \partial_{\boldsymbol{W}} \boldsymbol{F}(\boldsymbol{W}) \partial_{\boldsymbol{Y}} \boldsymbol{W}.$$

Once the exact solution $\boldsymbol{Y}^\star \left(\frac{x - x_{i+1/2}}{t}; \mathbf{Y}^n_i, \mathbf{Y}^n_{i+1}\right)$ of problem (3) is computed, the numerical flux is defined as:

$$\mathbf{g}^n_{i+\frac{1}{2}} = \mathbf{g}(\mathbf{W}^n_i, \mathbf{W}^n_{i+1}) = \boldsymbol{F}(\boldsymbol{W}(\boldsymbol{Y}^\star(0; \mathbf{Y}^n_i, \mathbf{Y}^n_{i+1})). \tag{4}$$

This numerical flux will be used for both inner interfaces and boundary interfaces.

3 Numerical Boundary Conditions for Outgoing Waves

We propose numerical artificial boundary conditions when no information is given on the open boundary of the computational sub-domain. One possible approach is to determine an artificial state $\mathbf{W}^n_{ext}$ in the virtual cell, symmetric of the boundary cell $\mathbf{W}^n_i$, outside of the sub-domain. The numerical boundary flux is then obtained by $\mathbf{g}^n_{1/2} = \mathbf{g}(\mathbf{W}^n_{ext,1}, \mathbf{W}^n_1)$ and $\mathbf{g}^n_{N+1/2} = \mathbf{g}(\mathbf{W}^n_N, \mathbf{W}^n_{ext,N})$. In the following, we assume that the exterior state is connected to the interior state either by a rarefaction wave or a shock wave.

3.1 Outgoing Rarefaction Wave

a. Formulation assuming the invariance of the interior state BC_0

The first boundary condition, widely used in industrial simulations, simply consists in taking the interior state $\mathbf{W}^n_i$ of the boundary cell at each time step t^n

$$\mathbf{W}^n_{ext} = \mathbf{W}^n_N. \tag{5}$$

The numerical boundary flux thus reads $\mathbf{g}^n_{N+1/2} = \mathbf{g}(\mathbf{W}^n_N, \mathbf{W}^n_N) = \boldsymbol{F}(\mathbf{W}^n_N)$. This technique does not need any knowledge about the wave structure.

b. Formulation using the wave structure and an extrapolation of the interior state $\mathrm{BC_r}$

The second boundary condition is built by using the two associated Riemann invariants of the regular wave and a third additional scalar relation. Note that, for an ideal gas, the exact velocity profile is linear w.r.t. x at time t^n. Thus, for an ideal gas EOS such that $\rho\varepsilon = P/(\gamma - 1)$, with $\gamma > 1$, we get:

$$\rho^n_{ext} = \rho^n_N \left(1 - \frac{\gamma - 1}{2}\frac{u^n_{N-1} - u^n_N}{c^n_N}\right)^{\frac{2}{\gamma-1}}, \quad P^n_{ext} = P^n_N \left(1 - \frac{\gamma - 1}{2}\frac{u^n_{N-1} - u^n_N}{c^n_N}\right)^{\frac{2\gamma}{\gamma-1}}$$

and $u^n_{ext} = 2u^n_N - u^n_{N-1}$. The numerical boundary flux is computed by $\mathbf{g}^n_{N+1/2} = \mathbf{g}(\mathbf{W}^n_N, \mathbf{W}^n_{ext})$. This technique connects the interior state with the exterior virtual state by using the rarefaction wave structure.

3.2 Outgoing Shock Wave

c. Formulation assuming the invariance of the interior state $\mathrm{BC_0}$

Same as for rarefaction wave, see case a. (5).

d. Formulation using the far-field state $\mathrm{BC_s}$

The boundary interior cell N is connected with the right initial state $\boldsymbol{W}^0_R$ by a virtual exterior cell of physical size αL, with L the domain length and $\alpha \in \mathbb{R}^*_+$ a parameter, see Fig. 1. Inspired by [3], this exterior state $\mathbf{W}^n_{ext}$ is updated with the numerical flux and the known state $\boldsymbol{W}^0_R$ such that:

$$\alpha L \left(\mathbf{W}^n_{ext} - \mathbf{W}^{n-1}_{ext}\right) + \Delta t^{n-1} \left(\mathbf{g}(\mathbf{W}^{n-1}_{ext}, \boldsymbol{W}^0_R) - \mathbf{g}(\mathbf{W}^{n-1}_N, \mathbf{W}^{n-1}_{ext})\right) = 0. \quad (6)$$

This technique gives the following asymptotic update of the exterior state $\mathbf{W}^n_{ext}$ when $\alpha \to +\infty$ for a finite time step Δt^{n-1}: $\lim\limits_{\alpha\to+\infty} \mathbf{W}^n_{ext} = \mathbf{W}^{n-1}_{ext}$. The exterior state is steady and therefore equal to its initial state $\mathbf{W}^0_{ext}$, which is the right state $\boldsymbol{W}^0_R$. The numerical boundary flux thus yields: $\mathbf{g}^n_{N+1/2} = \mathbf{g}(\mathbf{W}^n_N, \boldsymbol{W}^0_R)$. This asymptotic boundary condition amounts to impose, in the virtual exterior cell, the right state $\boldsymbol{W}^0_R$ known from the initial condition of the Cauchy problem.

4 Numerical Results

We discuss below some results of this preliminary study. Other results with distinct EOS are available in [2]. Two subsonic test cases, corresponding to 1D Riemann problems with a diatomic ideal gas EOS ($\gamma = \frac{7}{5}$), are performed with CFL$=$0.5. The first one is a pure left outgoing 1-rarefaction wave with the initial condition:

$$\begin{cases} (\rho_L, u_L, P_L) = \left(1\,\mathrm{kg/m^3}, 0\,\mathrm{m/s}, 10^5\,\mathrm{Pa}\right), \\ (\rho_R, u_R, P_R) = \left(0.5\,\mathrm{kg/m^3}, 242.2\,\mathrm{m/s}, 3.789 \times 10^4\,\mathrm{Pa}\right). \end{cases}$$

The second one is a pure right outgoing 3-shock wave with the initial condition:

$$\begin{cases} (\rho_L, u_L, P_L) = \left(1\,\mathrm{kg/m^3}, 418.3\,\mathrm{m/s}, 2.75 \times 10^5\,\mathrm{Pa}\right), \\ (\rho_R, u_R, P_R) = \left(0.5\,\mathrm{kg/m^3}, 0\,\mathrm{m/s}, 10^5\,\mathrm{Pa}\right). \end{cases}$$

The numerical convergence of the scheme, when waves are gone out of the bounded computational domain $\Omega = (-200\,\mathrm{m}, 200\,\mathrm{m})$, is measured with the L^1-norm of the error.

For smooth waves, the boundary conditions $\mathrm{BC_0}$ and $\mathrm{BC_r}$ enable to guarantee consistency when waves are going out ($t_0 < t < t_1$) or are gone out ($t > t_1$) of Ω. The numerical errors and the rates of convergence are collected in Table 1 and Fig. 2 for an outgoing rarefaction wave, and in Table 2 and Fig. 3 when the whole rarefaction wave has left the computational domain. As expected for an ideal gas EOS [8], the numerical rates of convergence for variables (u, P) are approximately 0.85—close to 1—when $t < t_1$ (see Table 1), and thus similar to those arising for $t < t_0$, see [8, 9]. Table 2 shows greater orders of convergence which may be due to the fact that the exact solution becomes fully constant for $t > t_1$. The $\mathrm{BC_r}$ condition gives very similar errors and does not provide more accurate approximations.

In contrast, the $\mathrm{BC_0}$ condition does not ensure the consistency of the scheme for an outgoing shock wave (at $t > t_0$, shock is outside of Ω), see Fig. 4: clearly, approximate solutions converge towards another solution when $(\Delta x, \Delta t) \to (0, 0)$.

Table 1 $\mathrm{BC_0}$: L^1 convergence orders for the rarefaction wave at $t_0 < t < t_1$

Δx (m)	N	ρ L^1-error	ρ cnv. order	u L^1-error	u cnv. order	P L^1-error	P cnv. order
5e−1	800	5.172e−3		8.868e−3		2.371e−3	
2.5e−1	1600	2.925e−3	0.8221	5.009e−3	0.8241	1.335e−3	0.8243
1.25e−1	3200	1.631e−3	0.8426	2.798e−3	0.8403	7.478e−4	0.8402
6.25e−2	6400	8.984e−4	0.8605	1.550e−3	0.8518	4.194e−4	0.8516
3.125e−2	12800	4.891e−4	0.8774	8.548e−4	0.8587	2.379e−4	0.8582
1.5625e−2	25600	2.691e−4	0.8621	4.714e−4	0.8588	1.386e−4	0.8579
7.8125e−3	51200	1.489e−4	0.8533	2.617e−4	0.8491	8.461e−5	0.8474

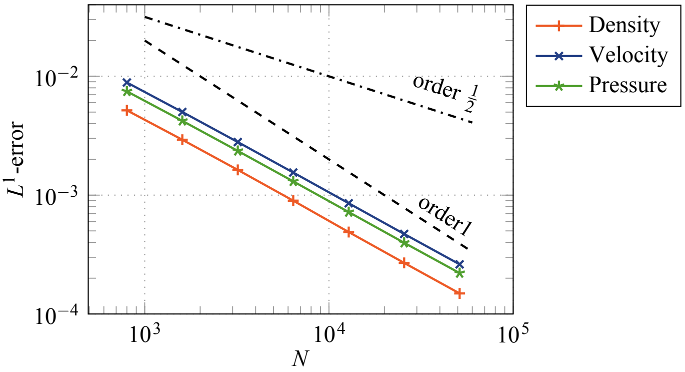


Fig. 2 BC$_0$: L^1 convergence curves for the rarefaction wave at $t_0 < t < t_1$

Table 2 BC$_0$: L^1 convergence orders for the rarefaction wave at $t > t_1$

Δx (m)	N	ρ L^1-error	ρ cnv. order	u L^1-error	u cnv. order	P L^1-error	P cnv. order
5e−1	800	1.279e−3		2.462e−4		2.562e−4	
2.5e−1	1600	6.755e−4	0.9211	1.284e−4	0.9384	1.337e−4	0.9383
1.25e−1	3200	3.522e−4	0.9395	6.557e−5	0.9700	6.826e−5	0.9700
6.25e−2	6400	1.823e−4	0.9502	3.265e−5	1.0061	3.399e−5	1.0061
3.125e−2	12800	9.423e−5	0.9521	1.565e−5	1.0608	1.629e−5	1.0609
1.5625e−2	25600	4.904e−5	0.9420	6.962e−6	1.1687	7.247e−6	1.1687
7.8125e−3	51200	2.604e−5	0.9134	2.551e−6	1.4486	2.655e−6	1.4486

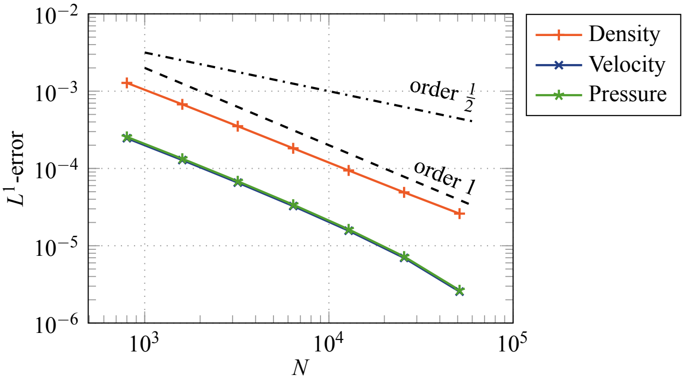


Fig. 3 BC$_0$: L^1 convergence curves for the rarefaction wave at $t > t_1$

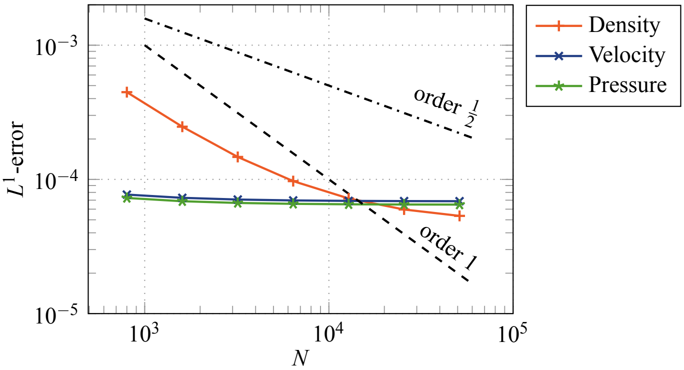


Fig. 4 BC_0: L^1 convergence curves for the shock wave at $t > t_0$

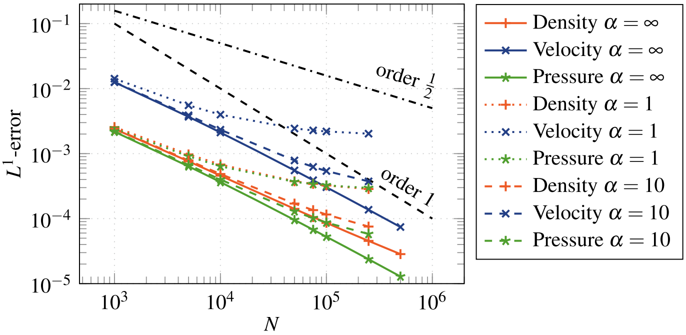


Fig. 5 BC_s: L^1 convergence curves for the shock tube at $t > t_0$

The BC_s boundary condition, for a finite value of the parameter $\alpha > 0$, is still not consistent, see Fig. 5. At the limit $\alpha \to +\infty$, the asymptotic condition BC_s allows to retrieve the consistency of the approximate solution with the exact solution.

Further works aim at considering another boundary condition for outgoing shock waves based on an imposed scalar value outside and the Rankine-Hugoniot relations. The issue of the supersonic shock wave case and of the dependence on the scheme [13] are being examined. To our knowledge, this measured loss of consistency has not been pointed out before.

Acknowledgements The first author received a financial support by ANRT through an EDF-CIFRE contract 2016/0728. The last author also receives a financial support by ANRT through an EDF-CIFRE contract 2017/0476. All computational facilities were provided by EDF R&D.

References

1. Bardos, C., LeRoux, A.Y., Nédélec, J.C.: First order quasilinear equations with boundary conditions. Commun. Part. Differ. Equ. **4**(9), 1017–1034 (1979)
2. Colas, C.: Time-implicit integral formulation for fluid flow modelling in congested media. Ph.D. thesis, Aix-Marseille Université (2019). https://tel.archives-ouvertes.fr/tel-02382958
3. Deininger, M., Iben, U., Munz, C.D.: Coupling of three- and one-dimensional hydraulic flow simulations. Comput. Fluids **190**, 128–138 (2019)
4. Dubois, F.: Boundary conditions and the Osher scheme for the Euler equations of gas dynamics. Internal Report CMAP 170, Ecole Polytechnique, Palaiseau, France (1987)
5. Dubois, F., Le Floch, P.: Boundary conditions for nonlinear hyperbolic systems of conservation laws. J. Diff. Equ. **71**(1), 93–122 (1988)
6. Engquist, B., Majda, A.: Absorbing boundary conditions for the numerical simulation of waves. Math. Comput. **31**(139), 629–651 (1977)
7. Gallouët, T.: Boundary conditions for hyperbolic equations or systems. In: Feistauer, M., Dolejší, V., Knobloch, P., Najzar, K. (eds.) Numerical Mathematics and Advanced Applications, pp. 39–55. Springer, Berlin, Heidelberg (2004)
8. Gallouët, T., Hérard, J.M., Seguin, N.: Some recent Finite Volume schemes to compute Euler equations using real gas EOS. Int. J. Numer. Methods Fluids **39**, 1073–1138 (2002)
9. Gallouët, T., Hérard, J.M., Seguin, N.: On the use of symmetrizing variables for vacuum. Calcolo **40**(3), 163–194 (2003)
10. Hedstrom, G.W.: Nonreflecting boundary conditions for nonlinear hyperbolic systems. J. Comput. Phys. **30**, 222–237 (1979)
11. Orlanski, I.: A simple boundary condition for unbounded hyperbolic flows. J. Comput. Phys. **21**(3), 251–269 (1976)
12. Poinsot, T.J., Lele, S.K.: Boundary conditions for direct simulations of compressible viscous flows. J. Comput. Phys. **101**(1), 104–129 (1992)
13. Quibel, L.: Simulation d'écoulements diphasiques eau-vapeur avec un modèle homogène. Ph.D. thesis in preparation, Université de Strasbourg. http://www.theses.fr/s188859
14. Smoller, J.: Shock Waves and Reaction-Diffusion Equations, A Series of Comprehensive Studies in Mathematics, vol. 258. Springer, New York (1994)

Simulation of a Liquid-Vapour Compressible Flow by a Lattice Boltzmann Method

Philippe Helluy, Olivier Hurisse, and Lucie Quibel

Abstract This work is devoted to the numerical resolution of a compressible three-phase flow with phase transition by a Lattice-Boltzmann Method (LBM). The flow presents complex features and large variations of physical quantities. The LBM is a robust numerical method that is entropy stable and that can be extended to second order accuracy without additional numerical cost. We present preliminary numerical results, which confirm its competitiveness compared to other Finite Volume methods.

Keywords Lattice Boltzmann method · Compressible flow · Phase transition

MSC (2010) 35Q79 · 76M12 · 76M28

1 Introduction

In this work, we are interested in the numerical resolution of a hyperbolic system arising in thermohydraulics. The objective is to compute a three-phase flow made of liquid water, vapour and an inert gas (such as air, for instance). Because of the envisaged range of pressure and temperature, there can be phase transition between the liquid and its vapour.

The Equation Of State (EOS) is complex and presents large variations of the thermodynamical parameters. It can be obtained from physical experiments and tabulations. It generally leads to very costly numerical methods, where most of the time is spent in the evaluation of the EOS. In addition, if because of the approximation

P. Helluy (✉)
Inria Tonus, IRMA UMR 7501, Université de Strasbourg, Strasbourg, France
e-mail: philippe.helluy@unistra.fr

O. Hurisse · L. Quibel
EDF R&D, 6 quai Watier, 78400 Chatou, France
e-mail: olivier.hurisse@edf.fr

L. Quibel
e-mail: lucie.quibel@edf.fr

R. Klöfkorn et al. (eds.), *Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples*, Springer Proceedings in Mathematics & Statistics 323,
https://doi.org/10.1007/978-3-030-43651-3_63

the EOS does not satisfy some convexity properties, the resulting system of conservation laws may not be hyperbolic and thus unstable. Here we use a simplified pressure law obtained from an entropy optimization procedure. The pressure law was first described in [1]. By construction, it ensures a convex hyperbolic domain and thus stability of some classical Finite Volume (FV) schemes such as Godunov-type schemes [7] or the Bouchut kinetic scheme [2].

The standard FV method is only first order. Its accuracy can be improved by slope reconstruction/limitation techniques. But this induces a cost and a more difficult parallelization because the computation stencil is enlarged.

In this work, we replace the FV scheme by a Lattice Boltzmann Method (LBM). The LBM is based on an abstract kinetic representation of the hyperbolic system. Then the scheme is a succession of free transport steps solved by an exact characteristic shift and relaxation operations that are local to the cell. This makes the LBM very efficient and easy to parallelize. In addition, by simply changing the relaxation parameter, it is possible to adjust the numerical viscosity of the LBM and to achieve second order with no additional cost.

We apply the whole approach for computing a vapour explosion test case.

2 Kinetic Approximation of Conservation Laws

2.1 Vectorial Kinetic Approximation with Over-Relaxation

In this work, we are interested in the numerical resolution of a hyperbolic system arising in thermohydraulics. The vector of unknown is denoted $\mathbf{u}(x,t) \in \mathbb{R}^m$. The system has the general form

$$\partial_t \mathbf{u} + \partial_x \mathbf{f}(\mathbf{u}) = 0. \tag{1}$$

The flux $\mathbf{f}$ is a smooth function $\mathbb{R}^m \mapsto \mathbb{R}^m$ satisfying the hyperbolicity property: its jacobian matrix $\mathbf{f}'(\mathbf{u})$ is diagonalizable with real eigenvalues for all $\mathbf{u}$ in the hyperbolicity domain $\mathscr{C}$, which is assumed to be convex. The relaxation approach, introduced by Jin and Xin [9], consists in replacing (1) by an extended system of the form

$$\partial_t \mathbf{u} + \partial_x \mathbf{z} = \mathbf{0}, \tag{2}$$

$$\partial_t \mathbf{z} + \lambda^2 \partial_x \mathbf{u} = \boldsymbol{\mu}. \tag{3}$$

The speed λ is a positive constant. The new vector $\mathbf{z}$ is called the approximated flux. The source term $\boldsymbol{\mu}$ is designed in such a way that $\mathbf{z} \simeq \mathbf{f}(\mathbf{u})$. We introduce a time step $\Delta t > 0$ and the Dirac comb:

$$\Psi(t) = \sum_{i \in \mathbb{Z}} \delta(t - i\Delta t).$$

The source term $\boldsymbol{\mu}$ is then defined by

$$\boldsymbol{\mu}(x,t) = \boldsymbol{\Omega}\Psi(t)\left(\mathbf{f}(\mathbf{u}(x,t)) - \mathbf{z}(x,t^-)\right), \quad \mathbf{I} \le \boldsymbol{\Omega} \le \mathbf{2I}.$$

In the more general case, $\boldsymbol{\Omega}$ is a matrix called the relaxation matrix. Inequalities on matrices have to be understood, as usual, in the sense of the associated quadratic forms. From the distribution theory, we see that at time $t = i\Delta t$, $\mathbf{z}$ is discontinuous: $\mathbf{z}(x,t^+) \neq \mathbf{z}(x,t^-)$, and

$$\mathbf{z}(x,t^+) = \boldsymbol{\Omega}\mathbf{f}(\mathbf{u}(x,t)) + (\mathbf{I} - \boldsymbol{\Omega})\mathbf{z}(x,t^-).$$

If the relaxation matrix $\boldsymbol{\Omega} = \mathbf{I}$, we recover in this way the classical first order splitting Jin-Xin algorithm, where $\mathbf{z} = \mathbf{f}(\mathbf{u})$ at the end of each time step. The **over-relaxation** corresponds to $\boldsymbol{\Omega} = 2\mathbf{I}$. It can be proved that the resulting scheme is a second order $O(\Delta t^2)$ approximation of (1). See [3, 5], for instance, and included references.

We can diagonalize the linear hyperbolic operator arising from the left-hand side of (2)–(3). In this way, we obtain a kinetic interpretation of the Jin-Xin approximation. For this, we consider the change of variables

$$\mathbf{k}^+ = \frac{\mathbf{u}}{2} + \frac{\mathbf{z}}{2\lambda}, \quad \mathbf{k}^- = \frac{\mathbf{u}}{2} - \frac{\mathbf{z}}{2\lambda}.$$

$$\mathbf{u} = \mathbf{k}^+ + \mathbf{k}^-, \quad \mathbf{z} = \lambda\mathbf{k}^+ - \lambda\mathbf{k}^-.$$

Then we get

$$\partial_t\mathbf{k}^+ + \lambda\partial_x\mathbf{k}^+ = \mathbf{r}^+, \quad \partial_t\mathbf{k}^- - \lambda\partial_x\mathbf{k}^- = \mathbf{r}^-, \tag{4}$$

where

$$\mathbf{r}^\pm(x,t) = \boldsymbol{\Omega}\Psi(t)\left(\mathbf{k}^{eq,\pm}(\mathbf{u}(x,t^-)) - \mathbf{k}^\pm(x,t^-)\right),$$

and the “Maxwellian” states $\mathbf{k}^{eq,\pm}$ are given by

$$\mathbf{k}^{eq,\pm}(\mathbf{u}) = \frac{\mathbf{u}}{2} \pm \frac{\mathbf{f}(\mathbf{u})}{2\lambda}.$$

In other words, from these calculations, we see that most of the time, the kinetic variables $\mathbf{k}^+$ and $\mathbf{k}^-$ satisfy free transport equations at velocity ±λ, with relaxation to equilibrium at each time step.

2.2 Equivalent Equation

The equivalent equation allows to better understand the effect of the relaxation matrix $\boldsymbol{\Omega}$. Let us introduce the “flux error” $\mathbf{y} := \mathbf{z} - \mathbf{f}(\mathbf{u})$. The following result holds:

Theorem 1 *If the relaxation matrix satisfies* $\mathbf{I} < \boldsymbol{\Omega} < 2\mathbf{I}$ *and if* $\mathbf{y} = 0$ *at the initial time, then, up to second order terms in* $O(\Delta t^2)$*,* $\mathbf{u}$ *is a solution of the following system of conservation laws*

$$\partial_t \mathbf{u} + \partial_x \mathbf{f}(\mathbf{u}) = \Delta t \partial_x \left((\boldsymbol{\Omega}^{-1} - \frac{1}{2}\mathbf{I})(\lambda^2 \mathbf{I} - \mathbf{f}'(\mathbf{u})^2)\partial_x \mathbf{u} \right) + O(\Delta t^2).$$

Remark 1 The proof is based on standard Taylor expansions. For a rigorous formulation and proof, we refer to [4]. The approach is classical in the analysis of the Lattice Boltzmann Method (LBM). See also for instance [5, 6, 10].

Remark 2 The above analysis allows to recover formally the so-called sub-characteristic condition. Assuming that $\mathbf{I} < \boldsymbol{\Omega} < 2\mathbf{I}$, the second order ("viscous") terms have the good sign, which ensures stability of the model, if the following matrix is positive:

$$\mathbf{V}(\mathbf{u}) = \lambda^2 \mathbf{I} - \mathbf{f}'(\mathbf{u})^2 > 0. \tag{5}$$

3 Numerical Methods

Our objective is to design a specific Lattice Boltzmann Method (LBM) for approximating three-phase flow. For comparison, we need a classical finite volume method, which we describe now.

3.1 Finite Volume Method

The finite volume scheme (FV scheme in the sequel) is constructed for approximating the solutions of (1). We denote by Δx the space step and by Δt the time step. We assume that the space step and the time step are related by a Courant–Friedrichs–Lewy (CFL) relation $\Delta t = \beta \frac{\Delta x}{\lambda}$, where $\beta > 0$ is the CFL number. We use the same velocity λ in the FV and LBM methods for defining the CFL number. Because of the sub-characteristic condition (5), λ is larger than the wave speeds of (1). We thus expect that the FV scheme will be stable at least for $\beta < 1$.

We look for an approximation

$$\mathbf{u}_i^n \simeq \frac{1}{\Delta x} \int_{x_{i-1/2}}^{x_{i+1/2}} \mathbf{u}(x, t_n) dx \simeq \mathbf{u}(x_i, t_n), \quad x_i = i\Delta x, \quad t_n = n\Delta t.$$

We consider the FV scheme

$$\frac{\mathbf{u}_i^{n+1} - \mathbf{u}_i^n}{\Delta t} + \frac{\mathbf{f}(\mathbf{u}_i^n, \mathbf{u}_{i+1}^n) - \mathbf{f}(\mathbf{u}_{i-1}^n, \mathbf{u}_i^n)}{\Delta x} = 0.$$

The numerical flux $\mathbf{f}(\cdot,\cdot)$ is the Rusanov flux given by

$$\mathbf{f}(\mathbf{u},\mathbf{v}) = \frac{\mathbf{f}(\mathbf{u})+\mathbf{f}(\mathbf{v})}{2} - \frac{\max(\rho(\mathbf{f}'(\mathbf{u})),\rho(\mathbf{f}'(\mathbf{v})))}{2}(\mathbf{v}-\mathbf{u}),$$

where $\rho(\mathbf{M})$ is the spectral radius of matrix $\mathbf{M}$.

3.2 Lattice Boltzmann Method (LBM)

In the LBM scheme, we assume that the CFL number $\beta = 1$. This allows to solve the free transport steps exactly. More precisely, if we also denote by $\mathbf{u}_i^n$, $\mathbf{z}_i^n$, $\mathbf{k}_i^{\pm,n}$ the approximation of $\mathbf{u}$, $\mathbf{z}$ and $\mathbf{k}^{\pm}$ at points x_i and time t_n, the transport step is given by simple shift operations, which solve the free transport equations (4) exactly

$$\mathbf{k}_i^{-,n+1-} = \mathbf{k}_{i+1}^{-,n}, \quad \mathbf{k}_i^{+,n+1-} = \mathbf{k}_{i-1}^{+,n}.$$

Then, one takes

$$\mathbf{u}_i^{n+1} = \mathbf{k}_i^{-,n+1-} + \mathbf{k}_i^{+,n+1-}, \quad \mathbf{z}_i^{n+1-} = -\lambda\mathbf{k}_i^{-,n+1-} + \lambda\mathbf{k}_i^{+,n+1-}.$$

The relaxation step is then

$$\mathbf{z}_i^{n+1} = \mathbf{z}_i^{n+1-} + \boldsymbol{\Omega}(\mathbf{u}_i^{n+1})\left(\mathbf{f}(\mathbf{u}_i^{n+1}) - \mathbf{z}_i^{n+1-}\right).$$

4 Application to a Three-Phase Flows

We wish to apply the above theory to a compressible three-phase flow model (two gases and a liquid). Because of strong variations in pressure and temperature, the liquid will undergo phase transition, which requires a proper mathematical model. The unknowns are the density ρ, the velocity u, the pressure p, the internal energy e and the mass fraction of the inert gas $\varphi = \varphi_3$. The total energy E is the sum of the internal energy and the kinetic energy: $E = \rho e + \frac{1}{2}\rho u^2$. The pressure Equation Of State (EOS) is of the form $p = p(\rho, e, \varphi)$. The three-phase flow model is a system of conservation laws of the form (1) with

$$\mathbf{u} = (\rho, \rho u, \rho E, \rho\varphi)^{\mathsf{T}}, \quad \mathbf{f}(\mathbf{u}) = (\rho u, \rho u^2 + p, (\rho E + p)u, \rho u\varphi)^{\mathsf{T}}.$$

Now we sketch the practical construction of the three-phase pressure law. This construction has to be done with care in order to ensure that the hyperbolicity domain $\mathcal{C}$ is convex. The general principles are mainly given in [1, 8]. We consider a mixture of three phases (1), (2) and (3) representing the vapour, the liquid and the non-

Table 1 Left: physical parameters for the three phases. Right:initial data for the vapor explosion test case

param.	vapour (1)	liquid (2)	gas (3)
γ_i	1.3	3	1.4
π_i (Pa)	0	8533×10^5	0
C_i (J.kg^{-1}.K^{-1})	1615.38	1400	719.28
Q_i (J.kg^{-1})	1.892×10^6	-1.1148×10^6	0
s_i^0	583.46	16658.99	263.62

	liquid (L)	air (R)
ρ	554.09	1.186245
e	1161999.729	210749.040
φ	10^{-6}	$1 - 10^{-6}$

condensable gas (air), respectively. The liquid is not miscible with the two others, while the vapour and the gas are miscible. We only admit phase transition between vapour (1) and liquid (2). Each phase obeys a stiffened gas Equation Of State (EOS), where the entropy function is defined by

$$s_i(\tau_i, e_i) = C_i \ln((e_i - Q_i - \pi_i \tau_i)\tau_i^{\gamma_i - 1}) + s_i^0 \quad i = 1, 2, 3. \tag{6}$$

In this formula, C_i is the specific heat at constant volume, Q_i is the heat of formation, π_i is the reference pressure and s_i^0 the reference entropy. The specific energy is noted e_i and the specific volume τ_i is the inverse of the density $1/\rho_i$. Some possible parameters are given in Table 1.

The mass fractions of the phases are noted φ_i, the volume fractions, α_i and the energy fractions, ζ_i. The phase specific volumes τ_i and energies e_i are related to the mixture specific volume τ and energy e by

$$\tau_i = \frac{\alpha_i}{\varphi_i}\tau, \quad e_i = \frac{\zeta_i}{\varphi_i}e.$$

The mass fraction φ_3 of the inert gas is supposed to be fixed and given. We thus introduce the vector of the unknown fractions $Y = (\varphi_1, \varphi_2, \alpha_1, \alpha_2, \alpha_3, \zeta_1, \zeta_2, \zeta_3)$. The unknown fractions satisfy the following constraints

$$Y \in Q := [0, 1]^8 \cap \{\alpha_1 = \alpha_3, \alpha_1 + \alpha_2 = 1, \varphi_1 + \varphi_2 + \varphi_3 = 1, \zeta_1 + \zeta_2 + \zeta_3 = 1\}.$$

These constraints are justified by the fact that the two gases are perfectly miscible (Dalton's law) and that the liquid and the gases are non-miscible. The mixture entropy is then given by a convex optimization problem:

$$s(\tau, e, \varphi_3) = \max_{Y \in Q} \sum_{i=1}^{3} \varphi_i s_i(\frac{\alpha_i}{\varphi_i}\tau, \frac{\zeta_i}{\varphi_i}e).$$

Once the optimization problem is solved, the temperature T and the pressure p of the mixture are then given by

$$T = 1/\frac{\partial s}{\partial e} \quad p = T\frac{\partial s}{\partial \tau}.$$

We have no place to detail the computations. We refer to [1]. The major advantage of the above construction is that it ensures that the hyperbolicity domain is convex.

5 Vapour Explosion Test

We consider a test case relevant for thermohydraulics. This is quite a realistic modelling of a sudden depressurization of a heated liquid in a pipe. The left (L) part of the pipe is filled with pressurized heated water. The right (R) part of the pipe is filled with air at ambient temperature and pressure. The numerical parameters are summed up in Table 1.

At time $t = 0$, the liquid-air separation is removed. We plot several physical quantities at time $t = 1.2\,\text{ms}$. We observe a complex wave structure. From left to right: a rarefaction wave running into the liquid, a slower vaporization wave running into the liquid, a contact wave, and finally a shock wave running into the air. Let us remark the presence of a non-standard split wave made of two simple waves. This is a typical feature of Riemann problems with non-convex equations of state arising from phase transition problematics.

On Fig. 1, we compare the numerical solutions obtained by the FV and the LBM schemes. The LBM is tested with an over-relaxation parameter $\boldsymbol{\Omega} = \mathbf{I}$ (first order) and $\boldsymbol{\Omega} = 1.9\mathbf{I}$ (improved precision). The second order LBM scheme with $\boldsymbol{\Omega} = 2\mathbf{I}$ is unstable here, which is not surprising because there is a shock wave to capture. The results of the LBM scheme with $\boldsymbol{\Omega} = \mathbf{I}$ are not plotted because they are almost superimposed with the results of the first order FV scheme. We observe a better precision of the improved LBM scheme with $\boldsymbol{\Omega} = 1.9\mathbf{I}$: the simple waves are better resolved. We observe small oscillations in the discontinuities. It is not surprising because $\boldsymbol{\Omega} = 1.9\mathbf{I}$ corresponds almost to a second order scheme without limiters. We are currently working on a better strategy for adapting locally the value of $\boldsymbol{\Omega}$ for suppressing oscillations.

6 Conclusion

We have constructed a numerical scheme based on the LBM. This scheme is faster and more precise than a classical FV method. It has been successfully validated on a complex three-phase flow with phase transition. It is possible to adjust its precision and stability thanks to the over-relaxation parameter $\boldsymbol{\Omega}$, with no additional computational cost. In future works we will study strategies for completely avoiding numerical oscillations in shock waves. This can certainly be achieved because the LBM scheme with $\boldsymbol{\Omega} = \mathbf{I}$ is free of oscillations and entropy-dissipative.

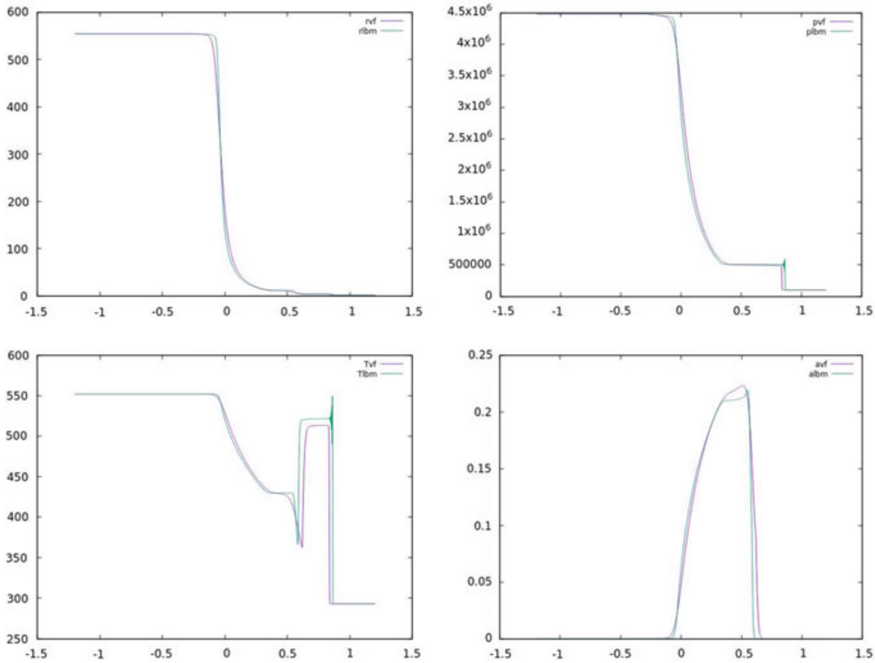


Fig. 1 Numerical solution of the Riemann problem described in Table 1. Top left: density, top right: pressure, bottom left: temperature, bottom right: vapour mass fraction. Comparison between the Finite Volume and Lattice Boltzmann Method with $\boldsymbol{\Omega} = 1.9\mathbf{I}$ on a mesh with 2000 cells

References

1. Bachmann, M., Müller, S., Helluy, P., Mathis, H.: A simple model for cavitation with non-condensable gases. In Hyperbolic Problems: Theory, Numerics and Applications (In 2 Volumes), pp. 289–296. World Scientific, Singapore (2012)
2. Bouchut, F.: Construction of BGK models with a family of kinetic entropies for a given system of conservation laws. J. Stat. Phys. **95**(1–2), 113–170 (1999)
3. Coulette, D., Franck, E., Helluy, P., Mehrenberger, M., Navoret, L.: High-order implicit palindromic discontinuous galerkin method for kinetic-relaxation approximation. Comput. Fluids **190**, 485–502 (2019)
4. Courtès, C., Coulette, D., Franck, E., Navoret, L.: Vectorial kinetic relaxation model with central velocity. Application to implicit relaxations schemes (2018)
5. Drui, F., Franck, E., Helluy, P., Navoret, L.: An analysis of over-relaxation in a kinetic approximation of systems of conservation laws. Comptes Rendus Méc. **347**(3), 259–269 (2019)
6. Dubois, F.: Equivalent partial differential equations of a lattice boltzmann scheme. Comput. Math. Appl. **55**(7), 1441–1449 (2008)
7. Harten, A., Lax, P.D., van Leer, B.: On upstream differencing and godunov-type schemes for hyperbolic conservation laws. SIAM Rev. **25**(1), 35–61 (1983)
8. Helluy, P., Mathis, H.: Pressure laws and fast legendre transform. Math. Model. Methods Appl. Sci. **21**(04), 745–775 (2011)

9. Jin, S., Xin, Z.: The relaxation schemes for systems of conservation laws in arbitrary space dimensions. Commun. Pure Appl. Math. **48**(3), 235–276 (1995)
10. Otomo, H., Boghosian, B.M., Dubois, F.: Two complementary lattice-boltzmann-based analyses for nonlinear systems. Phys. A: Stat. Mech. Its Appl. **486**, 1000–1011 (2017)

Discontinuous Galerkin Method for Incompressible Two-Phase Flows

Janick Gerstenberger, Samuel Burbulla, and Dietmar Kröner

Abstract In this contribution we present a local discontinuous Galerkin (LDG) pressure-correction scheme for the incompressible Navier–Stokes equations. The scheme does not need penalty parameters and satisfies the discrete continuity equation exactly. The scheme is especially suitable for two-phase flow when used with a piecewise-linear interface construction (PLIC) volume-of-fluid (VoF) method and cut-cell quadratures.

Keywords Two-phase flow · Incompressible Navier–Stokes Equations · Discontinuous galerkin

MSC (2010) 76T10 · 76D45 · 76D05 · 65M60

1 Introduction

Sharp interface models for incompressible two-phase flows have gained in popularity in recent years. These models combine incompressible flows in the bulk domains with jump conditions along the interfaces that separate the fluids, which model fluid interactions and surface effects, like surface tension.

Discontinuous Galerkin methods are a popular choice for solving incompressible flow problems due to their local mass conservation property and potentially high order of convergence.

J. Gerstenberger (✉) · D. Kröner
AAM, Albert-Ludwigs-Universität Freiburg, Hermann-Herder-Str. 10,
79104 Freiburg, Germany
e-mail: janick.gerstenberger@mathematik.uni-freiburg.de

D. Kröner
e-mail: dietmar@mathematik.uni-freiburg.de

S. Burbulla
IANS, Universität Stuttgart, Pfaffenwaldring 57, 70569 Stuttgart, Germany
e-mail: samuel.burbulla@mathematik.uni-stuttgart.de

R. Klöfkorn et al. (eds.), *Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples*, Springer Proceedings in Mathematics & Statistics 323,
https://doi.org/10.1007/978-3-030-43651-3_64

Designing schemes for solving such two-phase flows presents several challenges. The choice of method for the phase transport has implications for the interface representation and conservation properties. Volume-of-fluid [11] methods are conservative but their interface representations is in general non-continuous, while level-set methods have continuous interface representations, but are not conservative by themselves. There are several further methods that combine features from or generalize volume-of-fluid and level-set methods, but these are comparatively more complex. Incompressible flow problems have a saddle-point structure, which can make them computationally difficult/expensive to solve. Splitting methods like the various projection methods where introduced to decouple the problems into simpler ones. Projections methods replace the saddle-point problem with a advection-diffusion equation for the velocity and Poisson problem for the pressure, which are computationally simpler to solve. But for most discontinuous Galerkin discretizations the discrete continuity equation is not satisfied without some postprocessing techniques like the $H(div)$ reconstruction presented in [9]. Stability in regards to high coefficient ratios and strong surface effects are further issues that need to be addressed.

We present a discontinuous Galerkin pressure-correction method for incompressible two-phase flow that is robust in regards to coefficient ratios. The scheme is simple to implement and has shown good results for benchmarks problems and some numerical experiments with realistic data.

This contribution is structured as follows: We first briefly present the model for incompressible two-phase flow without phase transitions. Next we give some notation, present the modified LDG method without penalization introduced in [8] and present the new LDG pressure-correction scheme for incompressible two-phase flows based on it. Then we present numerical experiments for benchmark problems from the literature [7]. Lastly we give some concluding remarks and an outlook on further work.

2 Model

The model we consider is a simplification of the sharp interface model presented in [10] without phase transitions. Let $\Omega \subset \mathbb{R}^n$, $n = 2, 3$ be a bounded domain, which is divided into two disjunct phases: at time t, phase $i = 1, 2$ occupies subdomain $\Omega_i(t) \subset \Omega$. The boundary between the phase is the (phase) interface $\Gamma(t) := \partial\Omega_1 \cap \partial\Omega_2$.

Let $\mathbf{u}$ denote the velocity field, p the pressure field, ν the outer normal on Ω_1, κ the mean curvature of $\Gamma(t)$ and $[\![q]\!] := q_2 - q_1$ the jump of the variable q across $\Gamma(t)$. In addition the constants $\varrho_1, \varrho_2 > 0$ denote the densities, $\mu_1, \mu_2 > 0$ the viscosities of the phases, $\mathbf{g}$ the gravitational acceleration and σ the coefficient of surface tension. In the following we drop the indices on density and viscosity and keep in mind that these coefficients depend on the phase.

Then the model is given by the incompressible Navier–Stokes equations in each phase,

$$\left.\begin{aligned} \partial_t(\varrho\mathbf{u}) + \nabla\cdot(\varrho\mathbf{u}\otimes\mathbf{u} + \mathsf{T}) &= \varrho\mathbf{g} \\ \nabla\cdot\mathbf{u} &= 0 \end{aligned}\right\} \text{ in } \Omega_1, \Omega_2\,, \tag{1}$$

with the stress tensor $\mathsf{T} = p\mathsf{I} - 2\mu\,\mathcal{D}\mathbf{u}$, $\mathcal{D}\mathbf{u} = S(\nabla\mathbf{u}) := (\nabla\mathbf{u} + \nabla\mathbf{u}^\top)/2$, augmented with the following jump conditions

$$\left.\begin{aligned} [\![\mathsf{T}]\!]\boldsymbol{\nu} &= -\sigma\kappa\boldsymbol{\nu} \\ [\![\mathbf{u}]\!] &= 0 \end{aligned}\right\} \text{ on } \Gamma. \tag{2}$$

Additionally we impose either no-slip or free-slip conditions on the boundary of Ω.

3 Discretization

To discretize the model above we employ a primal LDG method [2]. It can be formulated in terms of a discrete gradient operator that is composed of the elementwise gradient and a lifting of the jumps into the piecewise discrete space. By constructing the liftings one order higher than the used discrete space the method is rendered stable without penalty parameters [8]. From these building blocks we construct an incremental pressure-projection scheme that satisfies the discrete continuity equation.

Since we require strict mass-conservation we choose to use a PLIC-VoF method [5, 11]. These methods have the disadvantage that the reconstructed interface is in general not continuous and its curvature needs further approximations [3], but is mass conservaftive.

We chose this LDG method because with the discrete gradients and no penalty terms no (explicit) evaluations of integrals over element boundaries are needed. In our experience the non-continuous discrete interface reconstructions on the inter-element boundaries lead to issues around the interface.

3.1 Notation and Liftings

To derive the discretization we first introduce some notation. Let $\mathcal{T}_h$ be a triangulation of Ω into elements E. By Σ_h^I we denote the set of all interior intersections e of elements $E^-, E^+ \in \mathcal{T}_h$ with $e = E^- \cap E^+ \neq \emptyset$, by Σ_h^D the set of all intersections with Dirichlet boundary values, by Σ_h^N the set of intersections with Neumann boundary values and by $\Sigma_h = \Sigma_h^I \cup \Sigma_h^D \cup \Sigma_h^N$ the set of all intersections.

For $e \in \Sigma_h^I$, with $e = E^- \cap E^+$, we introduce the jump and the average

$$[\![u]\!] = u_{|_{E^-}} - u_{|_{E^+}}\,, \tag{3a}$$

$$\langle\!\langle u \rangle\!\rangle = \tfrac{1}{2}\left(u_{|_{E^-}} + u_{|_{E^+}}\right) \tag{3b}$$

We define the discrete spaces of piecewise polynomials of degree $\leq k$

$$\mathcal{V}_k^d = \left\{\mathbf{v} \in [L^2(\Omega)]^d \mid \mathbf{v}_{|_E} \in [\mathcal{P}_k(E)]^d,\ E \in \mathcal{T}_h\right\} \tag{4}$$

and $\mathcal{V}_k = \mathcal{V}_k^1$ in the scalar case. The piecewise L^2 inner product over $\mathcal{T}_h$ is given by

$$\langle \mathbf{v}, \mathbf{w}\rangle = \sum_{E \in \mathcal{T}_h} \langle \mathbf{v}, \mathbf{w}\rangle_E = \sum_{E \in \mathcal{T}_h} \int_E \mathbf{v} \cdot \mathbf{w} \tag{5}$$

The gradient lifting operators $R : \mathcal{V}_k \mapsto \mathcal{V}_{k+1}^d$ and $R_a : \mathcal{V}_k \mapsto \mathcal{V}_{k+1}^d$ are defined by

$$\left\langle R([\![v]\!]), \mathbf{w}\right\rangle = \sum_{e \in \Sigma_h^I} \int_e [\![v]\!] \langle\!\langle \mathbf{w} \rangle\!\rangle \cdot \mathbf{n}_e\,, \tag{6a}$$

$$\left\langle R_a([\![v]\!]), \mathbf{w}\right\rangle = \left\langle R([\![v]\!]), \mathbf{w}\right\rangle + \sum_{e \in \Sigma_h^D} \int_e (v - a)\mathbf{w} \cdot \mathbf{n}_e\,, \tag{6b}$$

for all $\mathbf{w} \in \mathcal{V}_{k+1}^d$. Similarly, the divergence lifting operators $M : \mathcal{V}_{k+1}^d \mapsto \mathcal{V}_k$, $M_{\mathbf{b}} : \mathcal{V}_{k+1}^d \mapsto \mathcal{V}_k$ are defined by

$$\left\langle M([\![\mathbf{v}]\!]), w\right\rangle = \sum_{e \in \Sigma_h^I} \int_e [\![\mathbf{v}]\!] \cdot \mathbf{n}_e \,\langle\!\langle w \rangle\!\rangle\,, \tag{7a}$$

$$\left\langle M_{\mathbf{b}}([\![\mathbf{v}]\!]), w\right\rangle = \left\langle M([\![\mathbf{v}]\!]), w\right\rangle + \sum_{e \in \Sigma_h^D} \int_e (\mathbf{v} - \mathbf{b}) \cdot \mathbf{n}_e \, w\,, \tag{7b}$$

for all $w \in \mathcal{V}_k$. Here a, $\mathbf{b}$ are the respective Dirichlet boundary conditions.

With the liftings we can now define the lifted DG gradient and divergence

$$\overline{\nabla} v = \nabla_h v - R([\![v]\!]), \qquad \overline{\nabla}_g v = \nabla_h v - R_g([\![v]\!]), \qquad v \in \mathcal{V}_k, \tag{8a}$$

$$\overline{\nabla} \cdot \mathbf{v} = \nabla_h \cdot \mathbf{v} - M([\![\mathbf{v}]\!]), \qquad \overline{\nabla}_g \cdot \mathbf{v} = \nabla_h \cdot \mathbf{v} - M_g([\![\mathbf{v}]\!]), \qquad \mathbf{v} \in \mathcal{V}_k^d. \tag{8b}$$

The lifted derivatives with homogeneous Dirichlet boundary ($a = 0$, $\mathbf{b} = 0$) satisfy the following discrete integration-by-parts identities, as shown in [6]:

$$\left\langle \overline{\nabla}_0 \cdot \mathbf{v}, w \right\rangle = -\left\langle \mathbf{v}, \overline{\nabla} w \right\rangle \tag{9a}$$

$$\left\langle \overline{\nabla} \cdot \mathbf{v}, w \right\rangle = -\left\langle \mathbf{v}, \overline{\nabla}_0 w \right\rangle \tag{9b}$$

for all $\mathbf{v} \in \mathcal{V}_{k+1}^d$, $w \in \mathcal{V}_k$. This means the lifted derivatives are adjoint to each other. These identies are useful for defining of projection methods with respect to the lifted derivatives.

3.2 Unpenalized LDG Scheme

As a simple example we now consider the discretization of the Poisson equation $-\Delta u = f$ with homogeneous Dirichlet boundary data. The modified LDG method [8] reads: Find $u \in \mathcal{V}_k$ such that

$$\left\langle \overline{\nabla}_0 u, \overline{\nabla}_0 v \right\rangle = \left\langle f, v \right\rangle \qquad \forall v \in \mathcal{V}_k. \tag{10}$$

This scheme with order $k+1$ liftings is stable without adding penalty terms and can also be written in a "strong"-form by using the integration-by-parts identity (9b)

$$-\left\langle \underline{\Delta}_0 u, v \right\rangle := -\left\langle \overline{\nabla} \cdot \overline{\nabla}_0 u, v \right\rangle = \left\langle f, v \right\rangle \qquad \forall v \in \mathcal{V}_k. \tag{11}$$

3.3 Two-Phase LDG Scheme

We now present our primal LDG pressure-correction scheme for two-phase flow.

The scheme first split into an explicit (linearized) advection step for the momentum/velocity and the phase transport and an implicit Stokes step. Because the momentum and the phase interface are transported at the same rate, we can formulate the explicit step in the velocity, which also means this part of the scheme does not depend on the phase interface. The Stokes step is then further split into an implicit momentum step, a pressure Poisson step and an update step.

To sharply resolve the phases the terms containing a phase dependent coefficient are integrated using cut-cell quadratures. The cut-cell quadratures are constructed by cutting the elements containing an interface reconstruction with its interface reconstruction, subtriangulating the part of each phase and using standard simplex quadratures in the subtriangulations. Surface tension effect are included by integration of the jump condition over segments of the interface reconstruction. Using the lifted derivatives we eliminate (explicit) evaluations of integrals containing phase dependent coefficients along element boundaries, which in our experience can cause instabilities.

To simplify the presentation we restrict ourself to first order time stepping (IMEX Euler) and assume the phase transport/interface reconstruction is given: Let Γ_h^{n+1}

denote the set of all interface reconstruction at time t^{n+1} and κ^{n+1} the approximated interface mean curvature at time t^{n+1}.

The complete scheme for the velocity and pressure at time t^{n+1} then reads as: Find $\mathbf{u}^{n+1} \in \mathcal{V}^d_{k+1}$, $p^{n+1} \in \mathcal{V}_k$ such that

$$\frac{1}{\Delta t}\big\langle \hat{\mathbf{u}}^* - \mathbf{u}^n, \mathbf{v}\big\rangle - \big\langle \mathbf{u}^n \otimes \mathbf{w}^n, \nabla \mathbf{v}\big\rangle + \sum_{e\in\Sigma_h}\int_e F_e(\mathbf{u}^n; \mathbf{w}^n)\cdot \mathbf{v} = \big\langle \mathbf{g}, \mathbf{v}\big\rangle, \tag{12a}$$

$$\frac{1}{\Delta t}\big\langle \varrho(\mathbf{u}^* - \hat{\mathbf{u}}^*), \mathbf{v}\big\rangle - \big\langle 2\mu\, S(\overline{\nabla}_0\mathbf{u}^*), \overline{\nabla}_0\mathbf{v}\big\rangle = -\big\langle \overline{\nabla} p^n, \mathbf{v}\big\rangle + \sum_{\gamma\in\Gamma_h^{n+1}}\int_\gamma \sigma\kappa^{n+1}\mathbf{v}\cdot\mathbf{n}_\gamma\,, \tag{12b}$$

$$\big\langle \frac{1}{\varrho}\overline{\nabla} p^*, \overline{\nabla} q\big\rangle = \frac{1}{\Delta t}\big\langle \overline{\nabla}_0\cdot\mathbf{u}^*, q\big\rangle, \tag{12c}$$

$$\mathbf{u}^{n+1} = \mathbf{u}^* + \frac{\Delta t}{\varrho}\overline{\nabla} p^*\,, \qquad p^{n+1} = p^n + p^*\,, \tag{12d}$$

for all $\mathbf{v} \in \mathcal{V}^d_{k+1}$, $q \in \mathcal{V}_k$. Here F_e is a suitable numerical flux (e.g. local Lax-Friedrichs flux), $\mathbf{w}^n = \mathbf{u}^n$ is the transport velocity and all terms in Eqs. (12b), (12c) containing ϱ, μ are integrated with cut-cell quadratures with respect to Γ_h^{n+1}.

We note that here the lifted gradient/divergence maps into the dg space with increased/decreased polynomial order of its argument function space (e.g. $\overline{\nabla}_0\mathbf{u} \in \mathcal{V}^{d\times d}_{k+2}$, $\overline{\nabla}_0\cdot\mathbf{u} \in \mathcal{V}_k$,).

The resulting discrete velocity field is exactly divergence-free with respect to the discrete lifted divergence operator for arbitrary density ratios (up to the accurracy of the solver used). Applying the discrete DG divergence to the velocity field update and using Eq. (9a) recovers Eq. (12c).

4 Numerical Experiments

The scheme has been implemented in DUNE-FEM [4], which is part of the DUNE (Distributed and Unified Numerics Environment) framework [1]. The discretization used here are orthonormal $\mathcal{P}_2$ (resp. $\mathcal{P}_1$) dg elements for the velocity (resp. pressure) on a cartesian grid for space and a BDF2 scheme for time stepping. The implicit substeps are solved with a GMRES preconditioned with either an incomplete LU- or $LDL^\top$-decomposition. The phase transport is solved using a VoF method with a geometric flux, also implemented in DUNE.

We consider the benchmark problems presented in [7]. The general setup is the following: The computational domain is given by $\Omega = [0, 1] \times [0, 2] \subset \mathbb{R}^2$, Ω_2 is a circular subdomain with diameter $d = 0.5$ centered at $(0.5, 0.5)^\top$ and $\Omega_1 := \Omega \setminus \Omega_2$. The no-slip boundary condition is prescribed at the top and bottom boundaries, the

Table 1 Physical parameters and dimensionless numbers defining the test cases. $Re = \varrho_1 U_g d/\mu_1$, $Eo = \varrho_1 U_g^2 d/\sigma$, $U_g = \sqrt{gd}$

Case	ϱ_1	ϱ_2	μ_1	μ_2	g	σ	Re	Eo	ϱ_1/ϱ_2	μ_1/μ_2
1	1000	100	10	1	0.98	24.5	35	10	10	10
2	1000	1	10	0.1	0.98	1.96	35	125	1000	100

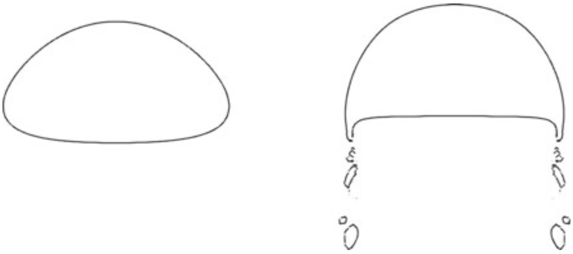


Fig. 1 Bubble shape of test case 1 (left) and test case 2 (right) at the final time $t = 3$ on a cartesian mesh with 80 by 160 cells

free-slip condition is prescribed on the vertical boundaries and the system is initially at reset.

The fluid coefficients for the test problems are given in Table 1. Test case 1 results in a ellipsoidal bubble, while in test case 2 the bubble gets deformed significantly and eventually break ups occur. Our results are in good agreement for the shape (see Fig. 1), the rise velocity and the center of mass of the bubble with the results presented in [7]. The 'circularity' can not be compared directly since our interface representation makes measuring the perimeter of the bubble difficult, nonetheless it is in general agreement.

Figure 2 shows an experiment of the effect of droplet merging on a super-hydrophobic surface in a water-steam system in free-fall. The initial configuaration consist of two equal-sized droplets, two unequal-sized droplets and a single droplet as reference. When droplets of similar size merge the force of surface tension is strong enough that the droplet "jumps".

5 Conclusions and Outlook

In this contribution we presented a primal local discontinuous Galerkin pressure correction scheme suitable for two-phase flows with high density and viscosity ratios. The interface is sharply resolved by using cut-cell quadratures. Numerical experiments show that the scheme agrees with other results presented in the literature.

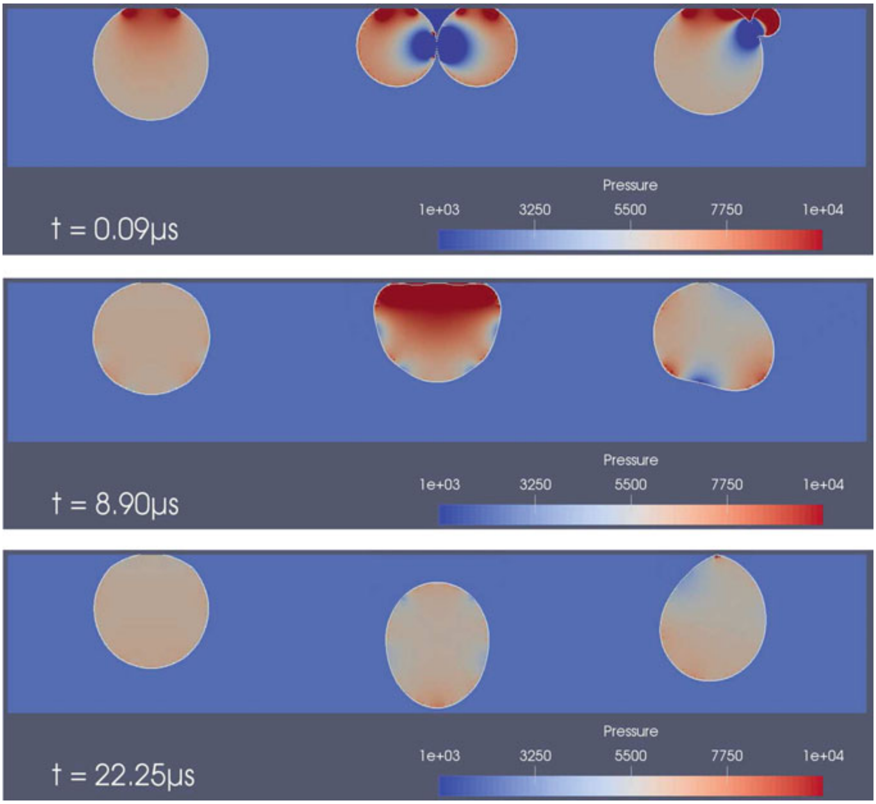


Fig. 2 Pressure profile of water droplets in a steam atmosphere on super-hydrophobic surface (contact angle $\alpha = 165°$) in free fall on a cartesian mesh with 300 by 100 cells, $h = 10^{-4}$m

Further work includes extending the model and the scheme to include the effects of phase transition and investigating if other, more compact, DG methods can recast in the same form as the presented LDG method.

Acknowledgements The first author would like to thank Tobias Malkmus for providing the initial code base for the fluid solver and explaining its subtleties, Martin Nolte for his ideas and help in modifying said code.

References

1. Blatt, M., Burchardt, A., Dedner, A., Engwer, C., Fahlke, J., Flemisch, B., Gersbacher, C., Gräser, C., Gruber, F., Grüninger, C., Kempf, D., Klöfkorn, R., Malkmus, T., Müthing, S., Nolte, M., Piatkowski, M., Sander, O.: The distributed and unified numerics environment, version 2.4. Arch. Numer. Softw. **4**(100), 13–29 (2016)

2. Cockburn, B., Kanschat, G., Schötzau, D.: A locally conservative LDG method for the incompressible Navier–Stokes equations. Math. Comput. **74**(251), 1067–1095 (2005)
3. Cummins, S.J., Francois, M.M., Kothe, D.B.: Estimating curvature from volume fractions. Comput. Struct. **83**(6), 425–434 (2005)
4. Dedner, A., Klöfkorn, R., Nolte, M., Ohlberger, M.: A generic interface for parallel and adaptive discretization schemes: abstraction principles and the DUNE-FEM module. Computing **90**(3–4), 165–196 (2010)
5. Diot, S., François, M.M.: An interface reconstruction method based on an analytical formula for 3d arbitrary convex cells. J. Comput. Phys. **305**, 63–74 (2016)
6. Feng, X., Lewis, T., Neilan, M.: Discontinuous galerkin finite element differential calculus and applications to numerical solutions of linear and nonlinear partial differential equations. J. Comput. Appl. Math **299**, 68–91 (2016)
7. Hysing, S., Turek, S., Kuzmin, D., Parolini, N., Burman, E., Ganesan, S., Tobiska, L.: Quantitative benchmark computations of two-dimensional bubble dynamics. Int. J. Numer. Methods Fluids **60**(11), 1259–1288 (2009)
8. John, L., Neilan, M., Smears, I.: Stable discontinuous Galerkin FEM without penalty parameters. In: Numerical Mathematics and Advanced Applications—ENUMATH 2015, pp. 165–173. Springer, Cham (2016)
9. Piatkowski, M., Müthing, S., Bastian, P.: A stable and high-order accurat splitting method for the incompressible Navier–Stokes equations. J. Comput. Phys. **356**, 220–239 (2018)
10. Prüss, J., Shimizu, S., Wilke, M.: Qualitative behaviour of incompressible two-phase flows with phase transitions: the case of non-equal densities. Commun. Part. Differ. Equ. **39**(7), 1236–1283 (2014)
11. Rider, W.J., Kothe, D.B.: Reconstructing volume tracking. J. Comput. Phys. **141**(2), 112–152 (1998)

High-Order Numerical Methods for Compressible Two-Phase Flows

Ksenia Kozhanova, Eric Goncalves, and Yannick Hoarau

Abstract We study the numerical methods to solve stiff two-phase flow problem which involves strong shock and expansion waves. In particular we focus the present study on high order reconstruction techniques coupled with HLLC and KNP numerical flux formulations associated to a four-equation model. These numerical methods are first tested on 1-D expansion tube case to investigate the accuracy of the schemes. The originality of our project is to construct a high-order numerical tool for solving the 2-D problem of two-phase shock-interface interaction with high density ratio between the phases. This paper presents the intermediate results with tests of low density ratio.

Keywords Two-phase flows · Hyperbolic system · Shock-interface interaction · High-order numerical schemes

MSC (2010) 65M99 · 76M12 · 76T99

1 Introduction

The importance of two-phase fluid flow modelling arises from many practical applications, from hydraulic turbines to power generation plants. However, the hyperbolic nature of the system describing such a flow make it extremely complicated task for numerical methods, especially for the cases with high density ratio and strong shock waves. There is a variety of methods to solve these problems, e.g. the sharp interface

K. Kozhanova (✉) · E. Goncalves
ISAE-ENSMA, Institut Pprime, UPR, 3346 CNRS, Poitiers, France
e-mail: ksenia.kozhanova@ensma.fr

E. Goncalves
e-mail: eric.goncalves@ensma.fr

Y. Hoarau
Université de Strasbourg, ICUBE, UMR 7357 CNRS, Strasbourg, France
e-mail: hoarau@unistra.fr

R. Klöfkorn et al. (eds.), *Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples*, Springer Proceedings in Mathematics & Statistics 323,
https://doi.org/10.1007/978-3-030-43651-3_65

or the diffuse interface methods. In the latter, the material interface can be captured by introducing a non-conservative transport equation for the void fraction. Yet, these methods can lead to spurious oscillations of the solution near the interface [14]. In this paper we consider a four-equation model [6, 7] and aim to establish the effect of numerical scheme on the basis of inviscid applications. The solver is based on an explicit finite volume method with different numerical schemes. Previous work have shown that simple MUSCL-based schemes do not perform well in the liquid-gas problems with strong shocks and expansion waves [6]. Thus, our focus is on high-order spatial resolution techniques coupled with HLLC and KNP flux approximations other than classical MUSCL methods. Only intermediate result of air-helium bubble interaction is presented here.

2 Mathematical Model

This section discusses the mathematical representation of the homogeneous compressible approach based on a 4-equation system. This system consists of three conservation laws for the mixture quantities and a fourth equation for the void-ratio α. Using conservative variables representation, i.e. $w = (\rho, \rho\overrightarrow{V}, \rho E, \alpha)$, in 2D it can be written as

$$\begin{aligned} \frac{\partial \rho}{\partial t} + \mathrm{div}(\rho\overrightarrow{V}) = 0 \\ \frac{\partial(\rho\overrightarrow{V})}{\partial t} + \mathrm{div}(\rho\overrightarrow{V}\otimes\overrightarrow{V} + PId) = 0 \\ \frac{\partial(\rho E)}{\partial t} + \mathrm{div}(\rho\overrightarrow{V}H) = 0 \\ \frac{\partial \alpha}{\partial t} + \overrightarrow{V}.\mathrm{grad}(\alpha) = \mathrm{Kdiv}(\overrightarrow{V}) \end{aligned} \tag{1}$$

where $\overrightarrow{V} = (u, v)$ denotes the centre of mass velocity, $E = e + V^2/2$ is the total energy of mixture and $H = h + V^2/2$ is the enthalpy of this mixture. The reflection of the change in each phase volume and speed of sound c of pure phases l, v are included into the term K,

$$K = \frac{\rho_l c_l^2 - \rho_v c_v^2}{\frac{\rho_l c_l^2}{1-\alpha} + \frac{\rho_v c_v^2}{\alpha}}$$

where ρ is the density of corresponding pure phase. We use the equations of state (EOS) for stiffened gas to close the system and relate the pressure and temperature to the internal energy and density. This system (1) is of hyperbolic nature and the mixture speed of sound follows the Wallis formulation [8].

3 Numerical Discretization

The matrix form of 4-equation model (1) in 1D is,

$$\frac{\partial w}{\partial t} + \frac{\partial (F(w))}{\partial x} = S(w) \tag{2}$$

where F is the convective flux, $S(w)$ is the source term and w is the vector of conserved variables and the void ratio. Using the finite volume method, we discretize the computational domain into regular meshes, i.e. Δx for space and Δt for time. Thus, Eq. (2) can be reformulated into its general discreet form,

$$\Delta x \frac{w_j^{n+1} - w_j^n}{\Delta t} + F_{j+1/2}^n - F_{j-1/2}^n = S_j^n \Delta x \tag{3}$$

where j and n stand for discretization in space and time, respectively.

We are looking to approximate the numerical flux $F_{j+1/2}^n$ and $F_{j-1/2}^n$ through the cell interface by using the solution to the Riemann problem or any other fully numerically resolved technique. The difficulty arises due to the non-conservative form of Eq. (1) and the existence of the source term. Two formulations of numerical flux are proposed in the present paper: a modification of classical HLLC numerical flux formulation with special treatment of the source term, see Eqs. 3.15–3.16 in [6], and KNP scheme [13].

The second-order accuracy in space and time is achieved by using MUSCL Hancock (MH) [9] predictor-corrector scheme. The predictor step for the cell values $\mathbf{W}_j$ is performed by using half the full time step. Thus, the new cell values $\tilde{\mathbf{W}}_j$ are computed as following,

$$\tilde{\mathbf{W}}_j = \mathbf{W}_j - \frac{\Delta t}{2\Delta x}\mathbf{A}_W \delta \mathbf{W}_j - \frac{\Delta t}{2}\mathbf{S}, \tag{4}$$

where $\mathbf{W}$ is a vector of reconstruction variables, $\mathbf{A}_W$ is a coefficient matrix, which depends on the choice of variables set, $\mathbf{S}$ is a source term and $\delta \mathbf{W}_j$ is a slope limiter. Importantly, the conservative or characteristic sets of reconstruction variables require the analytical derivation of Jacobian matrices in place of $\mathbf{A}$. The predictor step (4) is followed by the classical computation of time-centered interface values, which are then used to derive interface fluxes (see Eqs. 14–16 in [9]). Finally, the solution is advanced over the full time-step.

Our strategy is to improve the spatial order by finding suitable high order reconstruction method, i.e. by changing the calculation of the slope limiter $\delta \mathbf{W}_j$. The first candidate is piecewise parabolic method (PPM) [10], which is based on piecewise parabolic interpolation and has 4th order of accuracy for smooth solutions. The interface value $w_{j+\frac{1}{2}}$ is computed as,

$$w_{j+\frac{1}{2}} = \frac{7}{12}(w_j + w_{j+1}) - \frac{1}{12}(w_{j+2} + w_{j-1}), \tag{5}$$

This value is then applied to $w_{L,j}$ and $w_{R,j-1}$ for almost all j (where underscripts L and R are standing for the left and right interface neighbours, respectively, for the cell value w_j). Two particular cases may occur in the areas where the interpolation function takes on the values outside the interval between $w_{L,j}$ and $w_{R,j}$. The first one arises when w_j is a local minimum or maximum, where the interpolation function will be set constant. The second case is when w_j is inside the required interval, but close enough to the left or right variable. This situation is treated by monotonicity preserving technique, which is based on resetting one or both values of w_L and w_R. The resetting strategy is applied according to,

$$\begin{aligned}
&w_{L,j} \leftarrow w_j, w_{R,j} \leftarrow w_j \text{ if } (w_{R,j} - w_j)(w_j - w_{L,j}) \leq 0 \\
&w_{L,j} \leftarrow 3w_j - 2w_{R,j} \text{ if } (w_{R,j} - w_{L,j})(w_j - \tfrac{1}{2}(w_{L,j} + w_{R,j})) > \tfrac{(w_{R,j} - w_{L,j})^2}{6} \\
&w_{R,j} \leftarrow 3w_j - 2w_{L,j} \text{ if } -\tfrac{(w_{R,j} - w_{L,j})^2}{6} > \\
&(w_{R,j} - w_{L,j})(w_j - \tfrac{1}{2}(w_{L,j} + w_{R,j}))
\end{aligned} \tag{6}$$

Furthermore, we can improve this procedure by introducing the so-called narrow profile method, where we reset the values for $w_{L,j}$ and $w_{R,j}$ if jth zone is inside discontinuity. In this case formulation (5) is replaced by

$$w_{L,j} \leftarrow w^d_{L,j} = w_{j-1} + \frac{1}{2}\Delta_m w_{j-1}, \quad w_{R,j} \leftarrow w^d_{R,j} = w_{j+1} - \frac{1}{2}\Delta_m w_{j+1} \tag{7}$$

Here $\Delta_m w_j$ ensures that the discontinuities have sharper representation and $w_{j+1/2}$ falls inside the interval between w_j and w_{j+1},

$$\begin{aligned}
\Delta_m w_j &= \min(|\Delta w_j|,\ 2|w_j - w_{j-1}|,\ 2|w_j - w_{j-1}|)\operatorname{sgn}(\Delta w_j) \\
&\qquad \text{if } (w_{j+1} - w_j)(w_j - w_{j-1}) > 0 \\
&= 0 \ \text{ otherwise}
\end{aligned} \tag{8}$$

The alternative approach is weighted essentially non-oscillatory class of schemes (WENO). The present work considers only 3rd and 5th order of WENO, with several recent developments which allow to keep constant order of the scheme at all nodes. Among the advantages of WENO method we are mostly interested in smoother data dependence, which is expected to result in less oscillations, but at the same time sharper representation.

We denote $\omega_{0,1,2}$ and $q_{0,1,2}$ as weights and third-order linear reconstruction in three stencils of chosen set of variables, respectively. The fifth-order reconstruction then states [5],

$$w_{j+\frac{1}{2}} = \sum_{k=0}^{2} \omega_k q_k, \text{ with } \omega_k = \frac{\alpha_k}{\alpha_0 + \alpha_1 + \alpha_2} \text{ and } \alpha_k = \frac{C_k}{(\varepsilon + IS_k)^2} \tag{9}$$

where C_k ($C_0 = 0.1,\ C_1 = 0.6,\ C_2 = 0.3$) is the optimal weight parameter to achieve fifth-order upstream central difference approximation. Quantity ε is introduced to avoid zero in denominator and IS_k is the smoothness indicator, computed as a sum of L2 norms. The above formulation has been a subject to many research papers concerning the way to choose ε and the computation of smoothness indicator function.

The non-sensitivity to the ε number in the interval between 10^{-5} and 10^{-7} has been demonstrated by [5]. However, [1] showed that in the optimal weights calculation, the choice of ε becomes crucial due to the convergence to zero of other terms in denominator in smooth regions of the flow. More precisely, ε has a strong effect on the order of convergence of the resulting scheme and, moreover, the ENO behaviour of the scheme can be diminished by choosing ε to be too large.

The alternative derivation of weights along with appropriate value for ε has been proposed by [2] to achieve fifth order of reconstruction. A modified smoothness indicator formulation is,

$$IS_k^Z = \frac{IS_k + \varepsilon}{IS_k + |IS_0 - IS_2| + \varepsilon} \tag{10}$$

Due to the non-uniformity of the flow solution and/or finite grid size, we always have variation between IS_k and IS_k^Z, which can lead to the relatively large deviation of the weights values ω_k from optimal weights C_k. In fact, it can be proven that the better accuracy of the scheme achieved by having a smallest possible difference between ω_k and C_k.

This problem has been addressed in [3], where authors propose another alternative derivation of smoothness indicator function, based on the analysis of uniformity of IS_k,

$$IS_k^{SZ} = R_0 A\min(IS_0, \ldots, IS_{r-1}) + IS_k, \text{ where } R_0 = \frac{\min(\beta_k)}{\max(\beta_k) + \varepsilon'} \tag{11}$$

with $\varepsilon' = 10^{-10}$ set to avoid zero in denominator. R_0 is chosen in a way to ensure the uniform nature of IS_k and A can be chosen up to 100 to preserve ENO property. We set $A = 10$.

Using this improvement, we have the possibility to choose between approaches of [2, 5].

4 Numerical Results

This section presents the numerical results obtained using methods presented above. All reconstructions have been applied to the primitive variables, i.e. $\mathbf{W} = (\rho, u, v, P, \alpha)$, except for the PPM, where characteristic variables have been used. This choice is based on our previously performed study which demonstrated stable non-oscillating behaviour by using these variables. The directional splitting method has been used in order to tackle two-dimensional case.

The first case is the double rarefaction case proposed by [15], which has been used for the numerical schemes validation, namely, its accuracy and robustness. This test consists in a one meter long tube filled of water with a small fraction of gas $\alpha = 0.01$. An initial velocity discontinuity is located in the middle of the tube (± 2 m/s). The obtained numerical solution presents two expansion waves, which corresponds to the physical nature of the test and has been modelled by using different mesh size, i.e. 1000, 2000 and 4000 cells. The reference solution has been computed on 32000 nodes. The results has been compared at the final time $t = 3.2$ ms.

The solution computed on the coarsest mesh demonstrated fairly sharp approximation in good agreement with the reference solution, except by using WENO5 JS [5], which has symmetrical oscillations. On the other hand, the grid refinement revealed the oscillating nature of PPM reconstruction. However, we observed good robust performance by using 5th order WENO scheme coupled with HLLC MH, particularly with improved derivation of smoothness indicator as in formulation (11), where we set $\varepsilon = 10^{-20}$ (WENO SZ). This configuration provided the solution profiles slightly sharper than WENO with smoothness indicators computed as per (10), (WENO5 Z). The WENO3 scheme led to the diffusive results in all tests.

We continue the numerical study by moving to 2D shock-bubble interaction, using the air-helium configuration. This problem has been investigated by many authors since the experimental study of Haas and Sturtevant in 1987. The helium bubble has the initial diameter of 4 cm and is impacted by the normal shock wave moving at the Mach number 1.175. The EOS parameters are presented in Table 1:

The volume fraction α in this case is the volume fraction of the lighter gas in a carrier gas. We perform the calculations in a half-domain due to the symmetry of the problem. Simulations have been done using the uniform mesh discretization of 4000×400 cells and the time step of 2.510^{-9} s. All computations led to the correct bubble shape evolution compared to the experimental results.

The bubble is first flattened in shock propagation direction and becomes kidney shaped due to the formation of a high speed air jet at the upstream interface. The jet impingement on the downstream interface induces the formation of the counter-rotating vortical structures responsible for the bubble elongation. The density gradient modulus presented on the Fig. 1 at the time 0.05ms compares the performance of the

Table 1 Air-Helium EOS parameters and post-shock condition

	γ	P_∞	ρ
Air	1.4	0 Pa	1.163 kg/m^3
Helium	1.648	0 Pa	0.16 kg/m^3
	P	ρ	u
Post-shock	1.444 10^5 Pa	1.51 kg/m^3	93.65 m/s

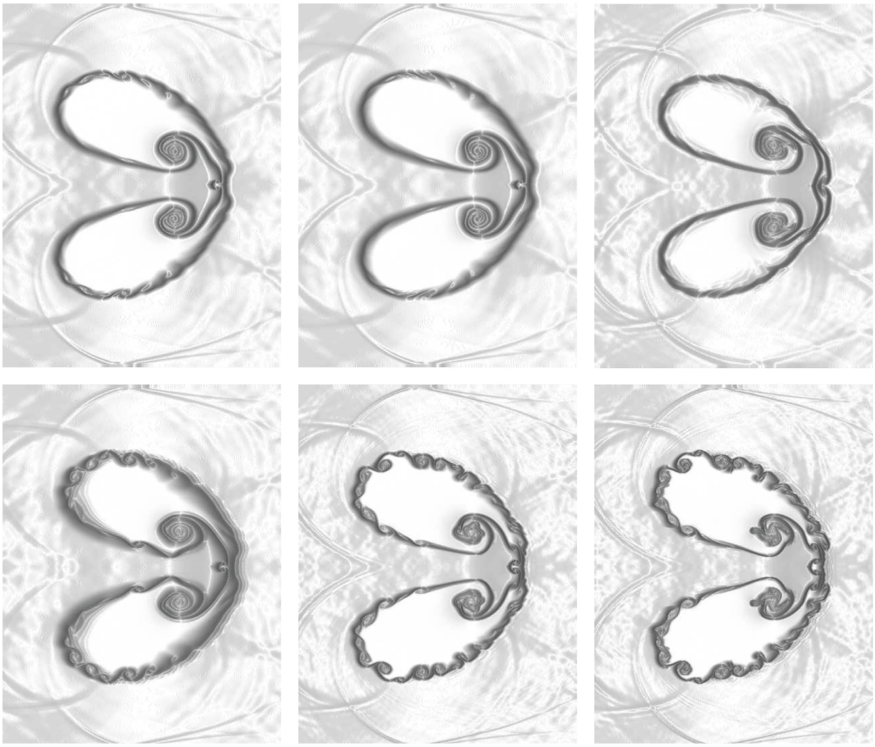


Fig. 1 Air-helium bubble interaction, density gradient, numerical schemes comparison (left to right, top to bottom): HLLC MH vanAlbada, HLLC MH WENO3, KNP MH WENO5 SZ, HLLC MH PPM, HLLC MH WENO5 JS, HLLC MH WENO5 SZ, mesh 4000 × 400, dt = 2.5e–9, T = 0.05 ms

schemes. The strong diffusive effect of WENO3 reconstruction is observed immediately, where the vortices of the bubble interface have been smoothed considerably. Slightly less diffusive result has been obtained by using KNP flux formulation, however, this led to the oscillations inside the bubble interface. The same result has been observed by applying other reconstruction techniques to KNP. The PPM with MH method showed notably sharper bubble interface but with yet even stronger oscillations. On the other hand, WENO5 techniques provided accurate sharp reconstruction of the bubble with reproducing the detailed vortices along the bubble surface when coupled with HLLC flux approximation, particularly WENO SZ scheme.

The extension of these methods to the high density ratio case is currently being studied. We have achieved third-order accuracy to this date.

5 Conclusion

The present work has been focused on high-order numerical schemes with application to two-phase flows, which are known to be difficult to solve numerically. This paper presents intermediate results with low density ratio between the phases. Particularly, we performed the implementation and computations using PPM, WENO numerical reconstructions with HLLC and KNP flux approximations. Our simulations demonstrated very diffusive results of KNP flux formulation and WENO3 reconstruction. PPM method led to sharper but oscillating results. High order WENO5 strategies provided detailed correct interface reconstruction.

Further work is going to be based on the implementation of different classes of flux formulation. The additional modifications of high order schemes are going to be studied, which would allow the computation for the cases with high density ratio. A detailed convergence analysis is going to be presented.

Acknowledgements This work has been fully supported by French ANR (ANR-18-CE46-0009).

References

1. Henrick, A.K., Aslam, T.D., Powers, J.M.: Mapped weighted essentially non-oscillatory schemes: achieving optimal order near critical points. J. Comput. Phys. **208**, 206 (2005)
2. Borges, R., Carmona, M., Costa, B., Don, W.S.: An improved weighted weighted essentially non-oscillatory scheme for hyperbolic conservation laws. J. Comput. Phys. **227**, 3191 (2008)
3. Shen, Y., Zha, G.: Improvement of the WENO scheme smoothness estimator. Int. J. Num. Meth. Fluids **64**, 653 (2009)
4. Shu, C.-W.: High order weighted essentially non-oscillatory schemes for convection dominated problems, III. Soc. Ind. Appl. Math. **51**, 82 (2009)
5. Jiang, G., Shu, C.-W.: "Efficient implementation of weighted ENO schemes. J. Comput. Phys. **126**, 202 (1996)
6. Goncalves, E., Zeidan, D.: Simulation of Compressible two-phase flows using a void ratio transport equation. Commun. Comput. Phys. **24**(1), 167 (2018)
7. Goncalves, E., Parnaudeau, P.: Comparison of multiphase models for computing shock-induced bubble collapse. Int. J. Num. Meth. Heat Fluid Flow. https://doi.org/10.1108/HFF-05-2019-0399
8. Wallis, G.: One-Dimensional Two-Phase Flow. McGraw-Hill, New York (1967)
9. van Leer, B.: On the relation between the upwind-differencing schemes of Godunov, Engquist Osher and Roe. SIAM J. Sci. Stat. Comput. **5**(1), 1 (1984)
10. Colella, P., Woodward, P.R.: The piecewise parabolic method (PPM) for gas-dynamical simulations. J. Comput. Phys. **54**(1), 174 (1984)
11. van Leer, B.: Towards the ultimate conservative difference scheme. V. A second-order sequel to Godunov's method. J. Comput. Phys. **32**, 101 (1979)
12. Jiang, G.-S., Shu, C.-W.: Efficient implementation of weighted ENO schemes. J. Comput. Phys. **126**, 202 (1996)
13. La Spina, G., Vitturi, M.: High-resolution finite volume central schemes for a compressible two-phase model. SIAM J. Sci. Comput. **34**(6), B861 (2012)

14. Abgrall, R.: How to prevent pressure oscillations in multicomponent flow calculations: a quasi conservative approach. J. Comput. Phys. **125**(1), 150 (1996)
15. Saurel, R., Le Metayer, O.: A multiphase model for compressible flows with interfaces, shocks, detonation waves and cavitation. J. Fluid Mech. **431**, 239 (2001)

A Python Framework for Solving Advection-Diffusion Problems

Andreas Dedner and Robert Klöfkorn

Abstract This paper discusses a Python interface for the recently published DUNE-FEM-DG module which provides highly efficient implementations of the Discontinuous Galerkin (DG) method for solving a wide range of non linear partial differential equations (PDE). Although the C++ interfaces of DUNE-FEM-DG are highly flexible and customizable, a solid knowledge of C++ is necessary to make use of this powerful tool. With this work easier user interfaces based on Python and the Unified Form Language are provided to open DUNE-FEM-DG for a broader audience. The Python interfaces are demonstrated for both parabolic and first order hyperbolic PDEs.

Keywords DUNE · DUNE-FEM · Finite volume · Discontinuous Galerkin · Python

MSC (2010) 65M08 · 65M60 · 35Q31 · 35Q90 · 68N99

In this paper we introduce a Python layer for the DUNE-FEM-DG[1] module [6] which is available open-source. The DUNE-FEM-DG module is based on DUNE [3] and DUNE-FEM [10] in particular and makes use of the infrastructure implemented by DUNE-FEM for seamless integration of parallel-adaptive Finite Element based discretization methods. DUNE-FEM-DG focuses exclusively on Discontinuous Galerkin (DG) methods for various types of problems. The discretizations used in this module are described by two main papers, [8] where we introduced a generic stabilization for convection dominated problems that works on generally unstructured and nonconforming grids and [5] where we introduced a parameter independent DG flux discretization for diffusive operators.

[1] https://gitlab.dune-project.org/dune-fem/dune-fem-dg.git.

A. Dedner (✉)
University of Warwick, Warwick, UK
e-mail: A.S.Dedner@warwick.ac.uk

R. Klöfkorn
NORCE Norwegian Research Centre AS, Nygaardsgaten 112, 5008 Bergen, Norway
e-mail: robert.kloefkorn@norceresearch.no

R. Klöfkorn et al. (eds.), *Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples*, Springer Proceedings in Mathematics & Statistics 323,
https://doi.org/10.1007/978-3-030-43651-3_66

DG methods have been studied intensively by many other groups and many software packages exist. However, most of these packages do not combine the following features: unstructured grids for 2, and 3 space dimensions, grid adaptivity, parallel computing capabilities, and open-source licenses. Besides DUNE-FEM-DG a few alternatives exists, for example, deal.II ([2]), feel++ ([13]), or Nektar++ ([12]).

DUNE-FEM-DG has been used in several applications (see [6] for a detailed list), most notably a comparison with the production code of the German Weather Service COSMO has been carried out for test cases for atmospheric flow ([4]). The focus of the implementation is on Runge–Kutta DG methods using mainly a matrix-free approach to handle implicit time discretizations which is a method especially used for convection dominated problems.

A strength of the DUNE-FEM-DG module is the general application area, i.e. convection dominated as well as diffusion dominated problems, for 1, 2, and 3d models, including parallelization and local grid adaptivity. A shortcoming so far has been the template heavy and relatively complicated C++ user interfaces for implementing new models and applications or coupling of such. Recent development has therefore been focused on adding a Python layer on top of DUNE-FEM-DG allowing user friendly model description based on the Unified Form Language (UFL) [1] and code generation. Low level Python bindings were introduced for the DUNE grid interface in [11] and a detailed tutorial providing high level access to DUNE-FEM is also available [9]. These bindings can now be used together with the efficient and flexible DG methods available in DUNE-FEM-DG making it easy to solve very complex coupled nonlinear PDEs.

The paper is organized as follows. In Sect. 1 we describe the DG discretizations. In Sect. 2 we introduce the newly developed Python based model interface. In Sect. 3 we investigate the performance impact of using Python scripting and conclude with discussing the extensibility of the approach in Sect. 4.

1 Governing Equations and Discretization

We consider a general class of time dependent nonlinear advection-diffusion-reaction problems for a vector valued function $\boldsymbol{U}\colon (0,T)\times\Omega\to\mathbb{R}^r$ with $r\in\mathbb{N}^+$ components of the form

$$\partial_t \boldsymbol{U} = \mathscr{L}(\boldsymbol{U}) := -\nabla\cdot\big(F_c(\boldsymbol{U}) - F_v(\boldsymbol{U},\nabla\boldsymbol{U})\big) + S_i(\boldsymbol{U}) + S_e(\boldsymbol{U}) \quad \text{in } (0,T]\times\Omega \tag{1}$$

in $\Omega\subset\mathbb{R}^d$, $d=1,2,3$. Suitable initial and boundary conditions have to be added. F_c describes the convective flux, F_v the viscous flux, S_i a stiff source term and S_e a non-stiff source term. Note that all the coefficients in the partial differential equation are allowed to depend explicitly on the spatial variable x and on time t but to simplify the presentation we suppress this dependency in our notation. Also note that any one of these terms is also allowed to be zero.

For the discretization we use a method of lines approach based on first discretizing the differential operator in space using a DG approximation and then solving the resulting system of ODEs using a time stepping scheme.

Given a tessellation $\mathscr{T}_h$ of the domain Ω with $\cup_{K\in\mathscr{T}_h} K = \Omega$ we introduce a piecewise polynomial space $V_h = \{\boldsymbol{v} \in L^2(\Omega, \mathbb{R}^r) : \boldsymbol{v}|_K \in [\mathscr{P}_k(K)]^r,\ K \in \mathscr{T}_h\}$ for some $k \in \mathbb{N}$, where on simplicial elements $\mathscr{P}_k(K)$ is a space containing polynomials up to degree k. Other basis functions and element types could be chosen as well but we focus on this case for simplicity of the presentation. In addition, we denote with Γ_i the set of all intersections between two elements of the grid $\mathscr{T}_h$ and accordingly with Γ the set of all intersections, also with the boundary of the domain Ω.

We discretize the spatial operator $\mathscr{L}(\boldsymbol{U})$ in (1) with either Dirichlet, Neumann, or Robin type boundary conditions by defining for all basis functions $\boldsymbol{\varphi} \in V_h$, $\langle \boldsymbol{\varphi}, \mathscr{L}_h(\boldsymbol{U}_h)\rangle := \langle \boldsymbol{\varphi}, K_h(\boldsymbol{U}_h)\rangle + \langle \boldsymbol{\varphi}, I_h(\boldsymbol{U}_h)\rangle$ with the element integrals

$$\langle \boldsymbol{\varphi}, K_h(\boldsymbol{U}_h)\rangle := \sum_{K\in\mathscr{T}_h} \int_K \Big((\widehat{F}_c(\boldsymbol{U}_h) - \widehat{F}_v(\boldsymbol{U}_h, \nabla\boldsymbol{U}_h)) : \nabla\boldsymbol{\varphi} + S(\boldsymbol{U}_h)\cdot\boldsymbol{\varphi}\Big), \quad (2)$$

with $S(\boldsymbol{U}_h) = S_i(\boldsymbol{U}_h) + S_e(\boldsymbol{U}_h)$ and the surface integrals (by introducing appropriate numerical fluxes $\widehat{F}_c$, $\widehat{F}_v$ for the convection and diffusion terms, respectively)

$$\begin{aligned}\langle \boldsymbol{\varphi}, I_h(\boldsymbol{U}_h)\rangle := &\sum_{e\in\Gamma_i} \int_e \Big(\{\widehat{F}_c(\boldsymbol{U}_h, \llbracket \boldsymbol{U}_h \rrbracket_e)^T : \nabla\boldsymbol{\varphi}\}_e + \{\widehat{F}_v(\boldsymbol{U}_h, \nabla\boldsymbol{U}_h)\}_e : \llbracket \boldsymbol{\varphi} \rrbracket_e \Big) \\ &- \sum_{e\in\Gamma} \int_e \Big(\widehat{F}_c(\boldsymbol{U}_h) - \widehat{F}_v(\boldsymbol{U}_h, \nabla\boldsymbol{U}_h) \Big) : \llbracket \boldsymbol{\varphi} \rrbracket_e, \end{aligned} \quad (3)$$

with $\{\boldsymbol{V}\}_e, \llbracket \boldsymbol{V} \rrbracket_e$ denoting the average and jump of $\boldsymbol{V}$ over e, respectively. The convective numerical flux $\widehat{F}_c$ can be any appropriate numerical flux e.g. the local Lax–Friedrichs flux. A wide range of diffusion fluxes $\widehat{F}_v$ can be found in the literature and many of these fluxes are available in DUNE-FEM-DG (cf. [5, 6]).

A range of different ODE solvers are available most based around *Strong Stability Preserving* Runge–Kutta methods (SSP-RK) (for details see [6]). The results and implementation techniques presented in this paper can be applied to explicit, implicit, or semi-implicit methods and mostly a **matrix-free** implementation of the discrete operator $\mathscr{L}_h$ is used. In addition, assembled operators are available.

When using semi-implicit time stepping, the operator $\mathscr{L}$ is split such that $\mathscr{L}(\boldsymbol{U}) = \mathscr{L}_e(\boldsymbol{U}) + \mathscr{L}_i(\boldsymbol{U})$ with

$$\mathscr{L}_e(\boldsymbol{U}) := -\nabla\cdot F_c(\boldsymbol{U}) + S_e(\boldsymbol{U}) \quad \text{and} \quad \mathscr{L}_i(\boldsymbol{U}) := \nabla\cdot F_v(\boldsymbol{U}, \nabla\boldsymbol{U})\big) + S_i(\boldsymbol{U})$$

where $\mathscr{L}_e$ is treated explicitly and is $\mathscr{L}_i$ is treated implicitly.

2 Python Model Interface

In this section we describe the interface to implement a new problem by providing functions describing the analytical model in its strong form (1). We start with a conservation law using the example of the Euler equations of gas dynamics with an ideal gas law were the pressure is given by $p = (\gamma - 1)(\epsilon - \frac{1}{2}\rho|\mathbf{v}|^2)$ with ϵ being the internal energy, ρ the density, $\mathbf{v}$ the velocity, and the adiabatic constant $\gamma = 1.4$. The model is given by a class with some static methods, i.e., `F_c(t,x,U)` returning the analytic flux as a matrix for given state vector U. For the DG approximation the numerical flux is also required. The simplest flux is given by the local Lax–Friedrichs scheme which requires (in addition to F_c) the maximum wave speed in a given direction $\mathbf{n}$. In the case of the Euler equations the maximum wave speed is $|\mathbf{v} \cdot \mathbf{n}| + c$ where $\mathbf{v}$ is the velocity and $c = \sqrt{\gamma \frac{p}{\rho}}$ is the speed of sound. The corresponding class method is `maxLambda`. In addition boundary conditions have to be provided in the model class. Below is an example implementation:

```
class CompressibleEuler:
    def toPrim(U): # auxiliary function
        v = as_vector( [U[i]/U[0] for i in range(1,3)] )
        pressure = 0.4*(U[3]-dot(v,v)*U[0]/2)
        return U[0], v, pressure
    def F_c(t,x,U):
        rho, v, p = Model.toPrim(U)
        return as_matrix([ [rho*v[0], rho*v[1]],
           [rho*v[0]*v[0]+p, rho*v[0]*v[1]], [rho*v[0]*v[1],
               rho*v[1]*v[1]+p],
           [(U[3]+p)*v[0], (U[3]+p)*v[1]] ])
    boundary = {range(1,5): lambda t,x,U: U}
    def maxLambda(t,x,U,n):
        rho, v, p = Model.toPrim(U)
        return abs(dot(v,n)) + sqrt(1.4*p/rho)
```

To set up the grid, space, and DG operator is straightforward as shown in the following code snippet:

```
gridView = structuredGrid([-1,-1],[1,1],[40,40])
space    = dgonb( gridView, order=3, dimRange=4)
uh       = space.interpolate([1.4,0,0,1], name='uh')
operator = femDGOperator(Model, space, limiter=None)
```

Finally DUNE-FEM implements a number of time stepping algorithms, e.g., explicit, implicit, and IMEX RK methods. For the above model (containing no diffusive flux) the default is an explicit SSP-RK method. A complete time loop is shown next:

```
stepper = femdgStepper(order=3, operator=operator)
t = 0
while t < 0.1:
    operator.setTime(t)
    t += stepper(uh, dt=0.001)
```

Instead of using a fixed time step $\Delta t = 0.001$ we can let the scheme choose a time step based on the CFL condition which is chosen using the `maxSpeed` method. The only required change is to remove the `dt` parameter in the `stepper` call.

To stabilize the DG method we use limiters and to reduce computational cost these are only applied in cells flagged by a *troubled cell indicator*. We use a indicator based on jumps over inflow edges of the cell [8]. This requires some additional methods: the velocity `velocity(t,x,U)` and an interfacial quantity `jump(U,V)`—we use the relative pressure jump to correctly detect shock waves. Finally, we also apply the limiter in cells where the solution takes on non physical values i.e. $\rho \leq 0$ and $p \leq 0$:

```
def velocity(t,x,U):
    return Model.toPrim(U)[1]
def jump(U,V):
    pL = Model.toPrim(U)[3]
    pR = Model.toPrim(V)[3]
    return (pL - pR)/(0.5*(pL + pR))
def physical(U):
    rho, _, p = Model.toPrim(U)
    return conditional( rho>1e-8, conditional( p>1e-8, 1, 0 ),
        0 )
```

To solve a problem with a discontinuous solution (here a radially symmetric Riemann problem) we just need to change the initial conditions and change the value of the `limiter` parameter when setting up the operator:

```
x = SpatialCoordinate(space)
uh.interpolate( conditional(dot(x,x)<0.1,
    as_vector([1,0,0,2.5]), as_vector([0.125,0,0,0.25])) )
operator = femDGOperator(Model, space, limiter="MinMod")
operator.applyLimiter(uh)
```

The code for the time loop remains the same as above.

To add diffusion and source terms only requires some additional methods on the model class: `def S_e(t,x,U,DU)`, `def S_i(t,x,U,DU)`, and `def F_v(t,x,U,DU)`. Here is an example of an advection diffusion reaction problem with three chemical components c_1, c_2, c_3 each satisfying $\partial_t c_i + \nabla \cdot (\mathbf{v} c_i) = \Delta c_i + S_e(c_1, c_2, c_3)$. The reaction rates are not very high so that we can treat the source term explicitly:

```
class Model:
    transportVelocity = computeVelocity()
    def S_e(t,x,U,DU):
        P1 = as_vector([0.1,0.1]) # midpoint of first source
        P2 = as_vector([0.9,0.9]) # midpoint of second source
        f1 = conditional(dot(x-P1,x-P1) < 0.1**2, 1, 0)
        f2 = conditional(dot(x-P2,x-P2) < 0.1**2, 1, 0)
        f  = conditional(t<5, as_vector([f1,f2,0]),
            as_vector([0,0,0]))
        r = 10*as_vector([U[0]*U[1], U[0]*U[1], -2*U[0]*U[1]])
        return f - r
    def F_c(t,x,U): return as_matrix([ [*
        (Model.velocity(t,x,U)*u)] for u in U ])
    def maxLambda(t,x,U,n): return
        abs(dot(Model.velocity(t,x,U),n))
    def velocity(t,x,U): return Model.transportVelocity
    def F_v(t,x,U,DU): return 0.02*DU
    def physical(U): return
        conditional(U[0]>=0,1,0)*conditional(U[1]>=0,1,0)*\
                            conditional(U[2]>=0,1,0)
    boundary = {range(1,5): as_vector([0,0,0])}
```

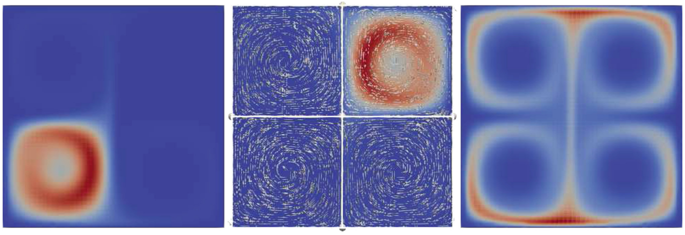


Fig. 1 The three components of the chemical reaction system (left to right) at $t = 10$

The velocity field is given by $\mathbf{v} = \nabla \times \Psi$ where Ψ solves $-\Delta\Psi = f$ with zero boundary conditions. We use a standard finite element scheme to compute Ψ

```
def computeVelocity():
    psiSpace = lagrange(gridView, order=1, dimRange=1)
    Psi   = psiSpace.interpolate(0,name="streamFunction")
    u,v,x = TrialFunction(psiSpace), TestFunction(psiSpace),
        SpatialCoordinate(psiSpace)
    form  = ( inner(grad(u),grad(v)) - 2*sin(x[0])*sin(x[1]) *
        v[0] ) * dx
    streamScheme = galerkin([form == 0,
        DirichletBC(streamSpace,[0]) ])
    streamScheme.solve(target=Psi)
    return as_vector([-Psi[0].dx(1),Psi[0].dx(0)])
```

Setting up the spatial operator and the stepper is done as above:

```
operator = femDGOperator(Model, space, limiter="scaling")
stepper  = femdgStepper(order=3, operator=operator)
```

where we use a scaling limiter to maintain positivity of all three reactants [7]. The velocity field consists of four rotors and in the top right and bottom left corners the concentration of c_1, c_2 is slowly increased over time up to $t = 5$. Through the flow and diffusion process c_1, c_2 both increase in the center of the domain leading to a production of c_3 as shown in Fig. 1.

3 Efficiency of Python Based Auto-Generated Models

While Python is easy to use, its flexibility can lead to some deficiencies when it comes to performance. In DUNE-Python and DUNE-FemPy a just-in-time compilation concept is used to create Python modules based on the static C++ type of every object used [11]. This way we avoid virtualization of the DUNE interfaces. It is therefore interesting to investigate the performance of this approach by comparing it to the previously hand-coded pure C++ version described in [6].

As a test example we choose a standard Riemann problem for the Euler equations solved on a series of different grid resolutions using quadratic basis functions, a

Table 1 Performance comparison of the C++ and the Python code for a simple test example solving the Euler equations in 2d with an explicit time stepping

SPGrid				ALUGrid			
code \ #el	1024	4096	16384	code \ #el	1024	4096	16384
C++	7.19	57.45	464.85	C++	12.72	106.08	884.28
Python	7.04	56.29	457.23	Python	13.32	110.48	924.81
C++ / Python	1.02	1.02	1.017	C++ / Python	0.955	0.96	0.956

minmod limiter, and explicit RK3 time stepping. We use two different grid implementation, a dedicated Cartesian grid (SPGrid) and a fully unstructured grid (ALUGrid) (Table 1).

We observe that for the Cartesian grid, SPGrid, we consistently achieve a small improvement of 2% while for the unstructured grid, ALUGrid, we observe a performance decrease of about 4%. This can be explained with the fact that for SPGrid all code can be inlined in the just-in-time compiled Python module. For ALUGrid, where a library exists, this is not so straight forward. In the future we will experiment with link time optimization and try to reduce implementation of small code snippets in the ALUGrid library.

4 Extensibility

In this paper we could only sketch the concepts behind the new Python bindings for DUNE-FEM-DG. The DUNE-FEM framework on which this is based, provides a significant amount of flexibility which we could not describe here in any detail. For example, changing the underlying grid implementation to a locally adaptive triangular grid or even a polygonal grid is straightforward as is changing the discrete function space. We are currently in the process of extending our DG implementation to work with polygonal grids. In a first step we show in Fig. 2 results for the radial Riemann problem mentioned in Sect. 2. In the Python code one only needs to change the grid implementation and use the `finiteVolume` space instead of the `dgonb` space.

```python
from dune.polygongrid import voronoiDomain, polygonGrid
boundingBox = numpy.array([ [-1,-1], [1,1] ])
gridView = polygonGrid( voronoiDomain(160000, boundingBox) )
space = finiteVolume( gridView, dimRange=4)
```

Other fluxes for discretizing both the advection and diffusion terms are also available and can be easily used by providing suitable parameters during the construction of the DG operator. In addition the Python bindings provided for DUNE-FEM [9] can be used to solve additional problems as shown in chemical reaction example.

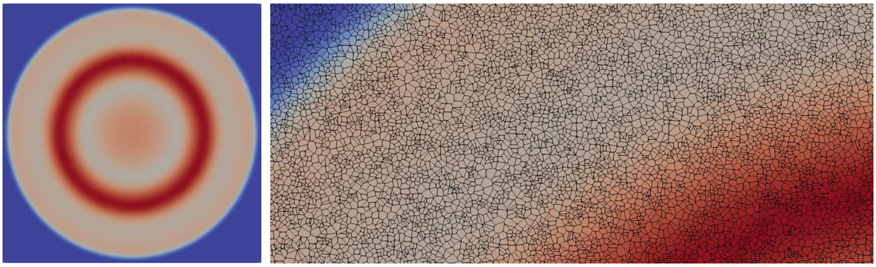


Fig. 2 Finite volume scheme for the Euler equations on a polygonal grid

We plan to improve the extensibility of the package further by providing simple hooks for users to implement their own numerical fluxes and to customize the available limiters, e.g., implementing their own troubled cell indicators. This will either be achieved through additional code generation where this is possible or through user written simple C++ functions for more complex problems. The use of code generation (or direct C++ implementations) is crucial for the parts of the discretization described above since these are used during the evaluation of the operator and are thus time critical. Other parts of the algorithm can be easily implemented on the Python side with only a very minor reduction to computational efficiency. This is for example the case for the time stepping scheme. It is already now straightforward to implement additional time stepping schemes directly within Python as shown below using the example of an explicit Runge–Kutta (RK) method:

```
# given: operator, cfl, explicit RK (A,b,c), discrete functions rhs, k[:]
operator.stepTime(0,0)
operator(uh,k[0])
dt = cfl*operator.timeStepEstimate[0]
for i in range(1,len(A)): rhs.assign(uh)
    for j in range(i): rhs.axpy(dt*A[i][j],k[j])
    operator.stepTime(c[i],dt)
    operator(rhs,k[i])
for i in range(len(b)): uh.axpy(dt*b[i],k[i])
operator.applyLimiter(uh)
```

In the future we plan to investigate bindings for other ODE solvers available through Python.

Acknowledgements Robert Klöfkorn acknowledges the support of the Research Council of Norway through the INTPART project INSPIRE (274883).

References

1. Alnæs, M.S., Logg, A., Ølgaard, K.B., Rognes, M.E., Wells, G.N.: Unified Form Language: A domain-specific language for weak formulations of partial differential equations. CoRR abs/1211.4047 (2012). http://arxiv.org/abs/1211.4047
2. Bangerth, W., Hartmann, R., Kanschat, G.: deal.II—a general purpose object oriented finite element library. ACM Trans. Math. Softw. **33**(4), 24/1–24/27 (2007). http://dealii.org/
3. Bastian, P., Blatt, M., Dedner, A., Engwer, C., Klöfkorn, R., Kornhuber, R., Ohlberger, M., Sander, O.: A generic grid interface for parallel and adaptive scientific computing. Part II: implementation and tests in DUNE. Computing **82**(2–3), 121–138 (2008)
4. Brdar, S., Baldauf, M., Dedner, A., Klöfkorn, R.: Comparison of dynamical cores for NWP models: comparison of COSMO and DUNE. Theor. Comput. Fluid Dyn. **27**(3–4), 453–472 (2013). https://doi.org/10.1007/s00162-012-0264-z
5. Brdar, S., Dedner, A., Klöfkorn, R.: Compact and stable discontinuous Galerkin methods for convection-diffusion problems. SIAM J. Sci. Comput. **34**(1), 263–282 (2012)
6. Dedner, A., Girke, S., Klöfkorn, R., Malkmus, T.: The DUNE-FEM-DG module. ANS **5**(1), 21–62 (2017). https://doi.org/10.11588/ans.2017.1.28602
7. Dedner, A., Kane, B., Klöfkorn, R., Nolte, M.: Python framework for hp-adaptive discontinuous Galerkin methods for two-phase flow in porous media. AMM **67**, 179–200 (2019)
8. Dedner, A., Klöfkorn, R.: A generic stabilization approach for higher order discontinuous Galerkin methods for convection dominated problems. J. Sci. Comput. **47**(3), 365–388 (2011)
9. Dedner, A., Klöfkorn, R.: The Dune-Fempy module (2019). https://dune-project.org/sphinx/content/sphinx/dune-fem/
10. Dedner, A., Klöfkorn, R., Nolte, M., Ohlberger, M.: A generic interface for parallel and adaptive scientific computing: abstraction principles and the DUNE-FEM module. Computing **90**(3–4), 165–196 (2010)
11. Dedner, A., Nolte, M.: The Dune-Python module. CoRR abs/1807.05252 (2018). http://arxiv.org/abs/1807.05252
12. Karniadakis, G., Sherwin, S.: Spectral/HP Element Methods for Computational Fluid Dynamics. Oxford University Press, Oxford (2005). http://www.nektar.info/
13. The Feel++ Consortium: The Feel++ book (2015). https://www.gitbook.com/book/feelpp/feelpp-book

3-Dimensional Particulate Flow Modelling Using a Viscous Penalty Combined with a Stable Projection Scheme

L. Batteux, J. Laminie, J.-C. Latché, and P. Poullet

Abstract We introduce a strategy for the simulation a particulate flow in a 3-dimensional domain. The particles are assumed to be rigid, and the homogeneous fluid flow to be governed by the incompressible Navier–Stokes equations. The system is solved using a predictor-corrector scheme for the Navier–Stokes equations with variable density. The latter scheme is adapted to take into account the solid domain by adopting a volume penalization method. In order to advect efficiently the particles, the approximation of the mass balance equation is carried out by an anti-dissipative scheme similar to the Ultra-Bee scheme. We conclude with numerical tests in the context of particulate flows.

Keywords 65M08 · 76D05 · 76T20

1 Introduction

This work is focused on the modelling of fluid–solid systems in a 3-dimensional domain. To reproduce faithfully the fluid–solid interactions is a problem of large interest due to numerous processes in industrial applications. There are several methods to attempt to model such a problem, but in this study, one considers rigid solid inclusions in an incompressible viscous fluid flow and one enforces a strong coupling between both phases. The motion of the solid domain may then be described using Newton laws for rigid bodies. As we are concerned with the efficiency and compu-

L. Batteux (✉) · J. Laminie · P. Poullet
LAMIA, Université des Antilles, Campus de Fouillole, 97157
Pointe-à-Pitre, Guadeloupe FWI, France
e-mail: lea.batteux@univ-antilles.fr

P. Poullet
e-mail: pascal.poullet@univ-antilles.fr

J.-C. Latché
IRSN, BP13115 St-Paul-lez-Durance Cedex, France
e-mail: jean-claude.latche@irsn.fr

R. Klöfkorn et al. (eds.), *Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples*, Springer Proceedings in Mathematics & Statistics 323,
https://doi.org/10.1007/978-3-030-43651-3_67

tational costs, we resort to an Eulerian formulation for the fluid flow and extend the fluid problem inside the solid domain in the manner of fictitious domain methods. In our case we enforce some kind of Brinkmann law inside the solid domain by adopting the H1-penalty method [1]. In practice, it will come down to penalizing the tensor of deformation term where the particles are in the non-homogeneous Navier–Stokes equations. Let us denote by Ω the domain containing the particulate flow. The continuous problem is given by:

$$\begin{cases} \dfrac{\partial \rho}{\partial t} + \boldsymbol{\nabla} \cdot (\rho \mathbf{u}) = 0 & \text{in } \mathbb{R}^+ \times \Omega \\[2ex] \dfrac{\partial (\rho \mathbf{u})}{\partial t} + \boldsymbol{\nabla} \cdot (\rho \mathbf{u} \otimes \mathbf{u}) - 2\boldsymbol{\nabla} \cdot (\mu(\rho)\mathbf{D}(\mathbf{u})) + \nabla p = \mathbf{f} & \text{in } \mathbb{R}^+ \times \Omega \\[2ex] \boldsymbol{\nabla} \cdot \mathbf{u} = 0 & \text{in } \mathbb{R}^+ \times \Omega \end{cases} \tag{1}$$

with the unknowns being the density ρ, velocity $\mathbf{u}$ and pressure p. The source term is denoted by $\mathbf{f}$ and contains the forces applied to the particles among other exterior forces. The advection of the rigid particles is carried out by the mass balance equation rather than Newton laws. The viscosity μ is continuously dependent on ρ, and taking $\mu \to \infty$ inside the solid domain allows for the penalization of the tensor of deformation rate $\mathbf{D}(\mathbf{u})$. We aim to tend towards $\|\mathbf{D}(\mathbf{u})\|_{\mathbf{L}^2(\Omega_s(t))} = 0$ which is equivalent to a rigid motion velocity field inside the particles $\Omega_s(t)$. The system of Eq. (1) is complemented with initial conditions for $(\rho, \mathbf{u})$.

The flow being incompressible, one often resorts to a projection scheme [4, 7] to solve problem (1). Many variants can be found in the literature especially when considering multiphase flows. However a lot of problems remain open regarding the stability or convergence of those schemes. Again, for efficiency reasons we will rely on the scheme introduced in [8]. In this article, the authors circumvent the relatively expensive computational cost of the elliptic pressure problem of standard projection problem for incompressible variable density flows, by switching to an approximate and more efficient formulation of the latter. In this paper, we adapt the scheme presented in [8] for finite elements to the MAC discretization, while keeping the same stability properties. Additionally, the discontinuity and the jumps of viscosity along the fluid/solid interface require an accurate tracking of the surface of the particles. To this end, we replace the discrete mass balance equation with an anti-diffusive advection scheme introduced in [3] that is similar to the Ultra-Bee method. The scheme is adapted to the dimension $d = 3$ by considering an alternate direction variant.

In the following section we introduce the notations, meshes as well as the full discrete scheme. In a final section we carry out and comment on the simulation of the fall of a rigid sphere.

2 Numerical Method

Let $0 = t_0 < t_1 < \cdots < t_N = T$ be a uniform partition of the time interval $[0, T]$. We note $\delta t = T/N$ the time step so that $t_n = n\delta t$, for $n \in [\![0, N]\!]$. The incremental projection scheme from [8] reads,

$$\frac{\rho^{n+1} - \rho^n}{\delta t} + \boldsymbol{\nabla} \cdot (\rho^{n+1}\overline{\mathbf{u}}^{\mathbf{n}}) = 0 \text{ in } \Omega \tag{2a}$$

$$\begin{aligned} \frac{\rho^{n+1}\mathbf{u}^{n+1} - \rho^n\mathbf{u}^n}{\delta t} &+\boldsymbol{\nabla} \cdot (\rho^{n+1}\mathbf{u}^{n+1} \otimes \overline{\mathbf{u}}^{\mathbf{n}}) \\ &-\boldsymbol{\nabla} \cdot ((\mu^{n+1}\mathbf{D}(\mathbf{u}^{n+1}))) + \nabla(2p^n - p^{n-1}) = \mathbf{f}^{n+1} \quad \text{in } \Omega \\ &\qquad\qquad\qquad\qquad\qquad\qquad \mathbf{u}^{n+1} = 0 \qquad \text{on } \partial\Omega \end{aligned} \tag{2b}$$

$$-\frac{\delta t}{\chi}\Delta(p^{n+1} - p^n) = -\boldsymbol{\nabla} \cdot \mathbf{u}^{\mathbf{n+1}} \text{ in } \Omega \tag{2c}$$

$$\overline{\mathbf{u}}^{\mathbf{n+1}} = \mathbf{u}^{n+1} - \frac{\delta t}{\chi}\nabla(p^{n+1} - p^n) \text{ in } \Omega \tag{2d}$$

for any time increment t^{n+1}. We denote $\mathbf{f}(t^{n+1}) = \mathbf{f}^{n+1}$ and we take $\chi = \min_{\mathbf{x}\in\Omega}\rho_0$. We aim to compute the sequence of discrete solution $(\mathbf{u}^{n+1}, \overline{\mathbf{u}}^{n+1}, p^{n+1}, \rho^{n+1})$ for $n \geq 0$. For each time step, the density is advected in (2a) by the divergence free velocity from the previous step, $\overline{\mathbf{u}}^n$. It follows with the calculation of a tentative velocity $\mathbf{u}^{n+1}$ by solving Eq. (2b). Using the Helmholtz decomposition of $\mathbf{L}^2(\Omega)$, the substep (2c) acts as the projection of $\mathbf{u}^{n+1}$ on $\mathbf{H} = \{\mathbf{u} \in \mathbf{L}^2(\Omega),\ \boldsymbol{\nabla} \cdot \mathbf{u} = 0,\ \mathbf{u} \cdot \mathbf{n}_{|\partial\Omega} = 0\}$ to get a corrected divergence–free velocity $\overline{\mathbf{u}}^{n+1}$ in (2d). The specificity (and interest) of the present scheme is to replace the actual density by the constant coefficient χ; indeed, a direct adaptation of the classic incremental projection scheme [7] to nonhomogeneous fluids would result in a variable Poisson problem of the form $\boldsymbol{\nabla} \cdot ((\rho^{n+1})^{-1}\nabla\Phi^{n+1}) = 0$ with Neumann boundary conditions. For discontinuous densities with a high $\rho_{\max}/\rho_{\min}$ ratio, this problem can be expensive to solve due to its ill-conditioned status.

2.1 Notations, Mesh and Discrete Projection Scheme

The domain Ω is discretized according to a staggered MAC mesh $\mathscr{D} = (\mathscr{M}, \mathscr{E})$ so the scheme (2) can benefit from the infsup stability property. Let the primal grid $\mathscr{M}$ consist in a conforming structured partition of Ω using rectangular parallelepipeds elements. The parallelepipeds are defined as primal cells and noted K. Therefore we have $\overline{\cup_{K\in\mathscr{M}} K} = \overline{\Omega}$. We may assume that the faces of the primal cells are normal to the vectors of the standard basis of $\mathbb{R}^3$, denoted by $(\mathbf{e}_1, \ldots, \mathbf{e}_d)$. A face of the primal cell $K \in \mathscr{M}$ will be noted $\sigma \in \mathscr{E}(K)$, $\mathscr{E}(K)$ referring to the set of all faces of K.

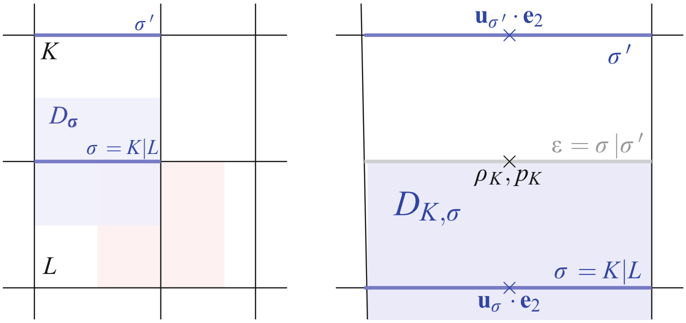

σ'
K
D_σ
$\sigma = K|L$
L
$\mathbf{u}_{\sigma'} \cdot \mathbf{e}_2$
σ'
$\varepsilon = \sigma|\sigma'$
ρ_K, p_K
$D_{K,\sigma}$
$\sigma = K|L$
$\mathbf{u}_\sigma \cdot \mathbf{e}_2$

Fig. 1 2-dimensional representation of $(\mathscr{M}, \mathscr{E})$

The staggered grid is completed by defining the dual grid $\mathscr{E}$ as the set of all edges of $\mathscr{M}$: $\{\sigma \in \mathscr{E}(K)|K \in \mathscr{M}\}$. We note $\mathscr{E} = \mathscr{E}_{\text{int}} \cup \mathscr{E}_{\text{ext}}$, where $\mathscr{E}_{\text{int}}$ (resp. $\mathscr{E}_{\text{ext}}$) are the edges of $\mathscr{E}$ that lie in the interior (resp. on the boundary) of the domain. The set of faces that are orthogonal to the ith unit vector $\mathbf{e}_i$ of the canonical basis of $\mathbb{R}^d$ is denoted by $\mathscr{E}^{(i)}$, for $i = 1, \ldots, d$. Correspondingly we introduce $\mathscr{E}^{(i)}_{\text{int}} = \mathscr{E}_{\text{int}} \cap \mathscr{E}^{(i)}$, $\mathscr{E}^{(i)}_{\text{ext}} = \mathscr{E}_{\text{ext}} \cap \mathscr{E}^{(i)}$ so $\mathscr{E}^{(i)} = \mathscr{E}^{(i)}_{\text{int}} \cup \mathscr{E}^{(i)}_{\text{ext}}$.

For $\sigma \in \mathscr{E}_{\text{int}}$, we note $\sigma = K|L$ for $(K, L) \in \mathscr{M}^2$ such that $\partial K \cap \partial L = \sigma$ and we associate the dual cell D_σ with $D_{K,\sigma} \cup D_{L,\sigma}$, where $D_{K,\sigma}$ (resp. $D_{L,\sigma}$) is the half-part of K (resp. L) adjacent to σ. If $\sigma \in \mathscr{E}_{\text{ext}}$ is adjacent to the cell K, then $D_\sigma = D_{K,\sigma}$. We can define Ω from the dual mesh: $\Omega = \cup_{\sigma \in \mathscr{E}_i} D_\sigma$, $i = 1, \ldots, d$. A dual face separating two duals cells D_σ and $D_{\sigma'}$ is denoted by $\epsilon = \sigma|\sigma'$. In agreement with the staggered MAC scheme, we will define the unknowns (ρ, p) on the primal grid and the ith component of the velocity on $\mathscr{E}^{(i)}$ in such a way that we deal with quantities $(\rho_K, p_K)_{K \in \mathscr{M}}$ and $(\mathbf{u}_\sigma)_{\sigma \in \mathscr{E}}$. The grids and notations for $d = 2$ are illustrated in Fig. 1.

The discretization of problem (2) begins with the discrete approximation of the mass balance Eq. (2a). For $K \in \mathscr{M}$, we resort in the classic way to the divergence formula for the computation of (2a) integrated over the primal cell K. This yields:

$$\frac{\rho_K^{n+1} - \rho_K^n}{\delta t} + \frac{1}{|K|} \sum_{\substack{\sigma \in \mathscr{E}(K) \\ \sigma \in \mathscr{E}^{(i)}_{\text{int}}}} F^{n+1}_{K,\sigma} = 0, \tag{3}$$

where $F^{n+1}_{K,\sigma}$ refers to the mass flux across the primal face σ outward K. In Sect. 2.2, we introduce the two techniques we adopt to compute $F^{n+1}_{K,\sigma}$; namely the classic upwind scheme and an antidiffusive scheme inspired by [3].

Let us focus on the discretization of the prediction step. For $\sigma = K|L \in \mathscr{E}^{(i)}_{\text{int}}$, the approximation of (2b) is obtained by integrating the ith prediction equation over the associated dual cell D_σ. In order for the scheme to be stable and provide the desired estimates for the unknowns, we must pay a particular attention to the discretization of the convective term in (2b):

$$\frac{1}{|D_\sigma|}\int_{D_\sigma}\Big((\rho^{n+1}\mathbf{u}^{n+1}-\rho^n\mathbf{u}^n)/\delta t+\boldsymbol{\nabla}\cdot(\rho^{n+1}\mathbf{u}^{n+1}\otimes\bar{\mathbf{u}}^n)\Big)\cdot\mathbf{e}_i\quad\mathbf{dx}$$

so that the discretization of the prediction step is compatible with the the discrete mass balance. This yields an approximation of the form:

$$(\rho^{n+1}_{D_\sigma}u^{n+1}_\sigma-\rho^n_{D_\sigma}u^n_\sigma)/\delta t+\frac{1}{|D_\sigma|}\sum_{\epsilon\in\mathscr{E}(D_\sigma)}F^{n+1}_{\sigma,\epsilon}u^{n+1}_{\epsilon,i}$$

where the values for the density on D_σ, denoted ρ_{D_σ} (resp. the fluxes on the dual faces σ, noted as $F^{n+1}_{\sigma,\epsilon}$) are given as functions of the density on the primal cells K, L (resp. the fluxes on the primal faces). This is achieved by averaging the discrete mass balance over K and L to obtain a consistent mass balance equation on D_σ. Therefore we define $|D_\sigma|\rho^n_{D_\sigma}=|D_{K,\sigma}|\rho^n_K+|D_{L,\sigma}|\rho^n_L$. The dual fluxes $F^{n+1}_{\sigma,\epsilon}$ are defined as the average of the fluxes on matching primal faces-the primal faces with coinciding outward normals [5, 9]. The approximation $u^{n+1}_{\epsilon,i}$ of the ith component of the velocity valued on the dual face ϵ is obtained by the upwind scheme. Finally the discrete ith prediction step is given by

$$\begin{aligned}(\rho^{n+1}_{D_\sigma}u^{n+1}_\sigma&-\rho^n_{D_\sigma}u^n_\sigma)/\delta t+\frac{1}{|D_\sigma|}\sum_{\epsilon\in\mathscr{E}(D_\sigma)}F^{n+1}_{\sigma,\epsilon}u^{n+1}_{\epsilon,i}\\&-(\boldsymbol{\nabla}\cdot(\mu^{n+1}\mathbf{D}(\mathbf{u}^{n+1})))_{D_\sigma}+(\nabla(2p^n-p^{n-1}))_{D_\sigma}=f^{n+1}_\sigma\end{aligned}$$

with $|D_\sigma|f^{n+1}_\sigma=\int_{D_\sigma}\mathbf{f}^{n+1}\mathbf{dx}$. The remaining terms are discretized in a straightforward way, with $(\nabla(2p^n-p^{n-1}))_{D_\sigma}$ approximated by $(|\sigma|/|D_\sigma|)(\varphi^n_L-\varphi^n_K)\mathbf{n}_{K,\sigma}\cdot\mathbf{e}_i$, if we note $\varphi^n=2p^n-p^{n-1}$ and define $\mathbf{n}_{K,\sigma}$ as the normal to the face σ outward K. For the viscous term, precautions must be taken given the discontinuous nature of the viscosity. It comes down to the discretization of the following term:

$$-\int_{\partial D_\sigma}\mu^{n+1}\mathbf{D}(\mathbf{u}^{n+1}):(\mathbf{n}_{\epsilon,\sigma}\otimes\mathbf{e}_i)\mathbf{dx}$$

Therefore involving the value of μ_ϵ, the viscosity evaluated at the faces of D_σ. In our case we average the viscosity over D_σ and D'_σ associated to the dual edge $\epsilon=\sigma|\sigma'$, and denoted μ_{D_σ} and $\mu_{D'_\sigma}$. We define μ_{D_σ} the same way ρ_{D_σ} was defined in the predictive step. However one may resort to VOF (Volume of Fluid) techniques-among others-for the computation of μ_ϵ [10]. We refer to [6] for the detailed approximation of the viscous term.

2.2 Antidiffusive Transport Scheme for the Particles

For the advection of the density we resort on one hand to the classic upwind scheme, and on the other hand to an antidiffusive transport technique we introduce below. As stated in the previous section, the difference in those methods primarily involves the computation of $F^{n+1}_{K,\sigma}$ in (3).

For the upwind scheme the latter is defined as $F^{n+1}_{K,\sigma} = |\sigma|\rho^{n+1}_{\sigma}\overline{u}^{n}_{K,\sigma}$ with $\overline{u}^{n}_{K,\sigma} = \overline{u}^{n}_{\sigma}\mathbf{n}_{K,\sigma}\cdot\mathbf{e}_i$ when $\sigma\in\mathscr{E}^{(i)}$. The updated density at the face $\sigma = K|L$, denoted ρ^{n+1}_{σ}, is given by:

$$\rho^{n+1}_{\sigma} = \begin{cases}\rho^{n+1}_{K}, & \overline{u}^{n}_{K,\sigma}\geq 0\\ \rho^{n+1}_{L}, & \text{otherwise}\end{cases}$$

However in the context of the non-homogeneous Navier–Stokes equations penalized by the H1–penalty method, this approximation of the mass balance generates a large numerical diffusion around the solid phase (as observed in Sect. 3). We replace the diffusive upwind technique for the transport of the density with an antidiffusive scheme (AD–scheme) based on [2, 3] and adapted to the dimension $d = 3$ of the problem by considering an alternate directions variant.

Let us focus on the transport of the density by the AD–scheme in the direction $\mathbf{e}_i$. For $K\in\mathscr{M}$ we note ρ^{*}_{K} the updated value of the density on K computed from its previous value ρ_K. Let $(K^-, K^+)\in\mathscr{M}^2$ so that the primal cells K^-, K, K^+ are consecutive and such that $\sigma^- = K^-|K$ and $\sigma^+ = K|K^+$ are in $\mathscr{E}^{(i)}$. We reorder the cells by imposing $\mathbf{n}_{K,\sigma^+}\cdot\mathbf{e}_i\geq 0$ and $\mathbf{n}_{K,\sigma^-}\cdot\mathbf{e}_i\leq 0$. For the time being, Let us assume that the velocities are positive. The transport of the density in ith direction is carried out by:

$$\rho^{*}_{K} = \rho_K - \frac{\delta t|\sigma|}{|K|}((\rho_{\sigma^+}u_{\sigma^+} - \rho_{\sigma^-}u_{\sigma^-}) - (u_{\sigma^+} - u_{\sigma^-})\rho_K)$$

where u_{σ^+} (resp. u_{σ^-}) is the value of the ith component of the velocity on σ^+ (resp. σ^-). The density on the faces σ^+ and σ^-, noted ρ_{σ^+} and ρ_{σ^-}, are to be determined. An equivalent formulation yields:

$$\rho^{*}_{K} = \rho_K + \nu_{\sigma^+}(\rho_K - \rho_{\sigma^+}) + \nu_{\sigma^-}(\rho_{\sigma^-} - \rho_K)$$

by defining $\nu_{\sigma^+} = \delta t|\sigma|u_{\sigma^+}/|K|$ and $\nu_{\sigma^-} = \delta t|\sigma|u_{\sigma^-}/|K|$ as local Courant numbers. We compute the value ρ_{σ^+} in such a way that ρ^{*}_{K} is a convex combination of $\rho_{K^-}, \rho_K, \rho_{K^+}$. Let us note $[\![a, b]\!] = [\min(a, b), \max(a, b)]$. It then comes down to the projection of the downwind value on σ^+ (ρ_{K^+} in this case) on $[\![\rho_K, \rho_{K^+}]\!]\cap[\![\rho_K, \rho_K + (1-\nu_{\sigma^-})/\nu_{\sigma^+}(\rho_K - \rho_{K^-})]\!]$ using the classic minmod formula. We carry out a similar process for any values of $\mathbf{u}$ on faces σ^+, σ^-[2]. We extend the advection to other directions to obtain the alternate direction variant of the AD–scheme.

3 Numerical Test—Dropping a Ball in a Viscous Fluid

We drop a rigid heavy sphere in a viscous fluid and observe it reaching its terminal velocity. We define the fluid by setting $\rho_f = \mu_f = 1$. The gravity constant $g = 9.81$ is applied to the ball. The sphere with radius $r = 0.08$ and density $\rho_s = 100$ is falling down the rectangular domain $[0, 1] \times [0, 1] \times [0, 3]$ to which we applied channel–flow boundary conditions. We take $\mu_s = 10^4$ for the penalty viscosity. For the time step we will be using $\delta t = 0.001$. The spatial step h is such that $h = \max_{i=x,y,z} h_i = 1/50$. For the initial data at $t = 0$, the fluid is considered at rest and the particle located at its initial position $(0.5, 0.5, 1)$. We take $p^{-1} = 0$ and compute p^0 and $\vec{\mathbf{u}}^{\mathbf{0}}$ by projection of the initial velocity on $\mathbf{H}$.

In this particular test we resort to the upwind scheme for the approximation of fluxes on the primal faces. However this technique can produce a large diffusion (Fig. 2) of the discontinuous quantities and a deformability of the particle (that is ensured by a well-advected viscosity), thus resulting in an incorrect terminal velocity. The advection of the solid phase requires an antidiffusive scheme as introduced in Sect. 2.2, for which we carry out the following 1-dimensional numerical test; the transport of a discontinuous density along the dimensional domain $[0, 2]$ and over the time interval $[0, 2]$. For the discretization steps we take $h = 0.005$ and $\delta t = 0.02$. The initial density is defined as $0.3(\mathbb{1}_{[0.1,0.5]}(x) + \mathbb{1}_{[1.0,1.5]}(x))$ and is advected by the velocity $u(t, x) = \frac{3}{2\pi} \max(\arctan(10^3(x - t - 1)), \arctan(-10^3(x - t - 1/10)))$. The former has been chosen to obtain sudden sign changes, and thus should advect the step-function back and forth. The shape of step function of the density is modified only by the reduction of the plateau (see Fig. 3). No numerical diffusion is added by the scheme and the bounds of the density remain the same.

In a second experiment, we study the behaviour of a droplet carried by the incompressible velocity field $(-\sin(x)\cos(y), \sin(y)\cos(x))$ for (x, y) in the

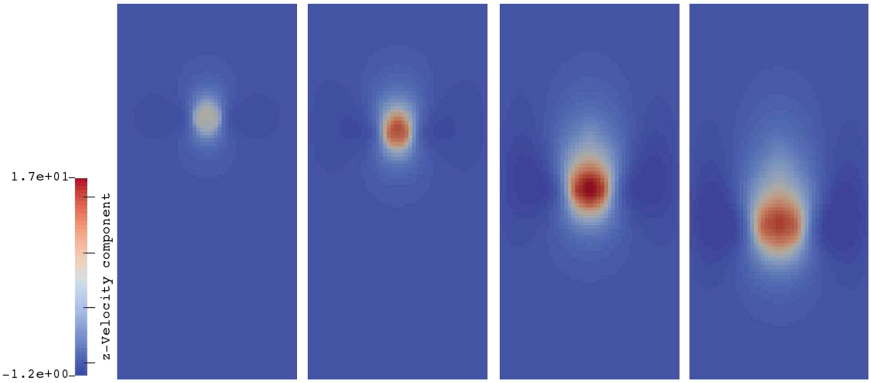


Fig. 2 Velocity component in the z-direction for the upwind scheme at $t = 0.01, 0.02, 0.05, 0.07$

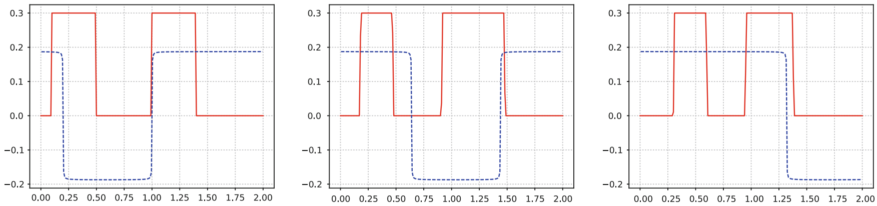


Fig. 3 State of the density (red line) at times $t = 0.0, 0.22, 0.56$. The figure also illustrates the velocity (dashed blue line) to highlight the change of sign of the velocity

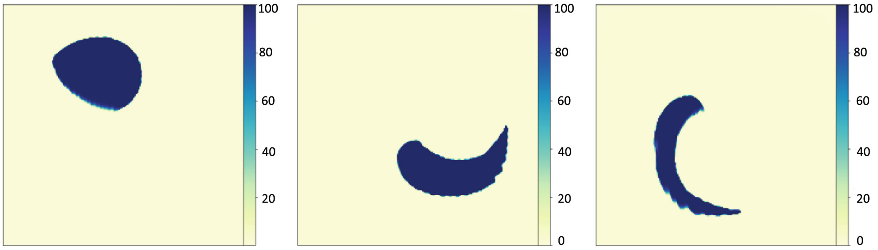


Fig. 4 State of the droplet and density values (colorbar) at times $t = 0.025, 0.25, 0.375$. For $t = 0$ we take $\rho(t, \mathbf{x}) = 100$ inside the droplet and zero everywhere else

2-dimensional domain $[0, \pi]^2$. Let $h = 0.005$ and $\delta t = 0.0025$. The droplet at $t = 0.0$ is defined as a ball with radius 0.3 and position $(\pi/3, \pi/3)$. While no rigid constraint is imposed on the droplet, we observe little diffusion and the conservation of density bounds over time (Fig. 4).

References

1. Angot, P., Bruneau, C.H., Fabrie, P.: A penalization method to take into account obstacles in incompressible viscous flows. Numer. Math. **81**, 497–520 (1999)
2. Bokanowski, O., Zidani, H.: Anti-dissipative schemes for advection and application to Hamilton–Jacobi–Bellmann equations. J. Sci. Comput. **30**, 1–33 (2007)
3. Després, B., Lagoutière, F.: Contact discontinuity capturing schemes for linear advection and compressible gas dynamics. J. Sci. Comput. **16**(4), 479–524 (2001)
4. Février̀e, C., Laminie, J., Poullet, P., Angot, P.: On the penalty-projection method for the Navier–Stokes equations with the MAC mesh. J. Comput. Appl. Math. **226**, 228–245 (2009)
5. Gallouët, T., Gastaldo, L., Herbin, R., Latché, J.-C.: An unconditionally stable pressure correction scheme for the compressible barotropic Navier–Stokes equations. ESAIM: M2AN **42**(2), 303–331 (2008)
6. Grapsas, D., Herbin, R., Kheriji, W., Latché, J.C.: An unconditionally stable staggered pressure correction scheme for the compressible Navier–Stokes equations. SMAI J. Comput. Math. **2**, 51–97 (2016)
7. Guermond, J., Minev, P., Shen, J.: An overview of projection methods for incompressible flows. Comput. Methods Appl. Mech. Eng. **195**(44), 6011–6045 (2006)

8. Guermond, J.L., Salgado, A.: A splitting method for incompressible flows with variable density based on a pressure Poisson equation. J. Comput. Phys. **228**(8), 2834–2846 (2009)
9. Herbin, R., Latché, J.C.: Kinetic energy control in the MAC discretization of compressible Navier–Stokes equations. Int. J. Fin. Vol **7**(2) (2010)
10. Vincent, S., de Motta, J.C.B., Sarthou, A., Estivalezes, J.L., Simonin, O., Climent, E.: A Lagrangian VOF tensorial penalty method for the DNS of resolved particle-laden flows. J. Comput. Phys. **256**, 582–614 (2014)

Data Assimilation for Ocean Drift Trajectories Using Massive Ensembles and GPUs

Håvard H. Holm, Martin L. Sætra, and André R. Brodtkorb

Abstract In this work, we perform fully nonlinear data assimilation of ocean drift trajectories using multiple GPUs. We use an ensemble of up to 10,000 members and the sequential importance resampling algorithm to assimilate observations of drift trajectories into the underlying shallow-water simulation model. Our results show an improved drift trajectory forecast using data assimilation for a complex and realistic simulation scenario, and the implementation exhibits good weak and strong scaling.

Keywords Particle filters · Finite-volume methods · Shallow-water simulations

MSC (2010) 62M99 · 60G35 · 35L65 · 65M08 · 76B15 · 68N19

1 Introduction

We present a proof-of-concept framework for performing fully nonlinear data assimilation of ocean drift trajectories into a shallow-water model. Forecasting drift trajectories in the ocean is an integral part of offshore preparedness services, and the forecasts are used in, e.g., search and rescue operations, oil spill tracking, and operations involving large floating structures [1]. Our approach is to use massive ensembles of simplified ocean models and assimilate observations using a particle filter based on the sequential importance resampling algorithm [2]. We first generate a massive

H. H. Holm (✉)
Mathematics and Cybernetics, SINTEF Digital, Oslo, Norway

Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim, Norway
e-mail: havard.heitlo.holm@sintef.no

M. L. Sætra · A. R. Brodtkorb
Norwegian Meteorological Institute, Oslo, Norway

Department of Computer Science, Oslo Metropolitan University, Oslo, Norway
e-mail: martinls@met.no
e-mail: andreb@met.no

R. Klöfkorn et al. (eds.), *Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples*, Springer Proceedings in Mathematics & Statistics 323,
https://doi.org/10.1007/978-3-030-43651-3_68

ensemble of perturbed ocean states and simulate each ensemble member forward in time until we have an observation. Next, we use the particle filter to discard ensemble members that match poorly with the observation, and then reinitialize the discarded members based on the simulated states that have a good match. We continue the simulation until the next available observation and repeat the process.

Particle filters as used here are embarrassingly parallel and require synchronization only when resampling individual ensemble members. We therefore use MPI to distribute ensemble members to different nodes, and each member is simulated forward in time using a modern explicit finite-volume scheme. The scheme is implemented on the GPU in a massively data-parallel fashion and includes all the complex source terms required for oceanographic simulations of real-world domains [3].

Our experiments show significantly increased forecast skill compared to both deterministic and Monte Carlo simulations, and we are able to run experiments with 10,000 ensemble members on the Nvidia DGX-2 server, which has 16 GPUs [4]. The implementation exhibits good weak and strong scaling and is possible to extend with more complex perturbation methods with minimal effort.

2 Data Assimilation of Ocean Drift Observations

Sequential importance resampling is an example of a particle filter for fully nonlinear data assimilation (see the recent review paper by Vetra-Carvalho et al. [5] on ensemble-based data assimilation techniques). A benefit of the algorithm is that, contrary to e.g., the ensemble Kalman filter [6], it does not manipulate variables of individual simulations that match well with available observations. This means that the resulting ensemble contains ocean states that are consistent with respect to the physics of the model. Furthermore, particle filters do not make any assumptions on linearity in the physical model or Gaussian probability distributions. However, the algorithm requires a large number of ensemble members as the probability that an individual ensemble member matches an observation is small. In fact, the required number of members increases exponentially with the number of observations [2, 7]. This means that it is most suitable for nonlinear problems with few observations, which is the typical situation for our target application area.

General ocean circulation models, such as ROMS [8], conserves mass, three dimensional momentum, salinity, and temperature, and operational setups typically require large computational resources to run even a single simulation. Herein, we investigate using simplified ocean models through the two-dimensional shallow-water equations, which were used operationally in the early days of computational oceanography [9]. These simplified ocean models are suitable for short-term forecasts in which the ocean can be modeled as a barotropic fluid, and they can be efficiently simulated using GPUs. A further challenge is that even the operational models often have limited forecast abilities due to uncertain initial conditions, model parameters and forcing. By using a simplified model instead of the full three dimensional equations, we can afford to run a much larger ensemble of perturbed ocean states that can give us a more detailed description of the uncertainties in ocean forecasts that can be used as a complement to the current operational methods [10].

We have developed a GPU-based simulation framework that uses operational ocean forecasts for initial and boundary conditions, bathymetry, and forcing [3]. The framework is an extension of the high-resolution, central-upwind finite-volume scheme proposed by Chertock et al. [11], which is well-balanced with respect to steady states in which the Coriolis force balances a non-zero momentum and water surface displacement (the geostrophic balance). The scheme uses H as the water depth, η as the deviation from mean sea level, and hu and hv as the momentum along the abscissa and ordinate, respectively. We can perturb this ocean state using the approach in [12], in which we first generate a smooth random field, $\Delta\eta$, for each ensemble member, representing deviations of the ocean surface elevation. We continue by computing the momentum required to balance this perturbation, namely

$$\Delta hu_{j,k} = -\frac{gH_{j,k}}{f_{j,k}} \frac{\Delta\eta_{j,k+1} - \Delta\eta_{j,k-1}}{2\Delta y}, \qquad \Delta hv_{j,k} = \frac{gH_{j,k}}{f_{j,k}} \frac{\Delta\eta_{j+1,k} - \Delta\eta_{j-1,k}}{2\Delta x}, \tag{1}$$

and finally add these perturbations to the state variables.

Using the perturbations from (1), we generate an ensemble of ocean states and use the ensemble $\psi^n = \{\psi_0^n, \psi_1^n, ..., \psi_N^n\}$ at time t_n as an approximation to the probability density function (pdf) $p(\psi^n)$ of our ocean state. If we have an observation y^n of part of the state (e.g., one drift trajectory), we can improve the probabilistic forecast by using the conditional pdf $p(\psi^n|y^n)$. Using Bayes theorem, $p(\psi^n|y^n) = p(y^n|\psi^n)p(\psi^n)/p(y^n)$, we can write $p(\psi^n|y^n)$ as a weighted ensemble[1]:

$$p(\psi^n|y^n) \propto \sum_{i=1}^{N} \frac{p(y^n|\psi_i^n)}{\sum_{j=1}^{N} p(y^n|\psi_j^n)} \delta(\psi^n - \psi_i^n) = \sum_{i=1}^{N} w_i^n \delta(\psi^n - \psi_i^n), \tag{2}$$

in which δ is the Dirac's delta function. The weights, w_i^n, reflect how well ensemble member i matches the observation, and members with very low weights have a negligible contribution to $p(\psi^n|y^n)$. Sequential importance resampling therefore discards members with low weights and duplicates members with high weights to maintain a higher sample density in the high-probability areas.

A challenge with sequential importance resampling is the so-called curse of dimensionality, as we are operating in a very high-dimensional space. The particle filter is prone to ensemble collapse, in which the ensemble quickly reduces into only a very few significant states and thereby only has marginally better predictive skill than a purely deterministic simulation [2, 7]. This means that we need a much larger number of ensemble members compared to the number of observations. A major benefit, however, is that the particle filter makes no changes to the states of

[1] We have ignored the marginal probability as it only serves as a normalization constant.

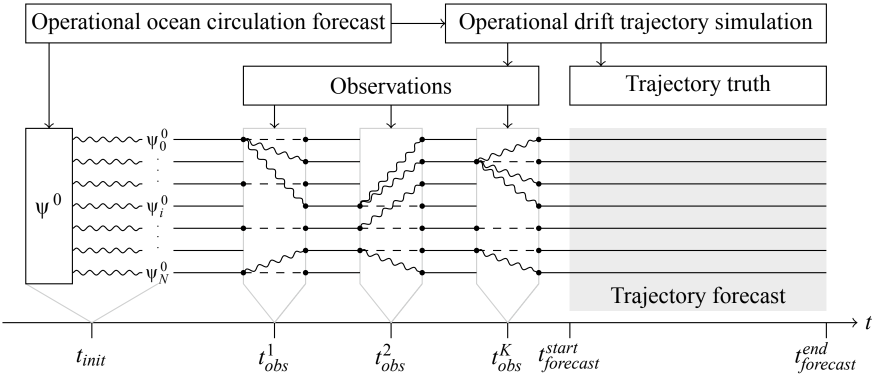


Fig. 1 Algorithmic overview of the simulation, data assimilation and drift trajectory forecast. Straight lines are deterministic simulation, wiggly lines are perturbations, and dashed lines show ensemble members that are kept during the resampling phase

individual ensemble members during the data assimilation phase. This means that the perturbation strategy is not limited by the data assimilation method.

Figure 1 gives an overview of the ensemble prediction system used in this paper. We use the sea-surface elevation and vertically integrated ocean currents from the operational ocean forecast provided by the ROMS-based NorKyst-800 model system [13] as initial and boundary conditions, and give an independent perturbation to each ensemble member to represent the uncertainty in the initial condition. Furthermore, we also use the same bathymetry and wind forcing as NorKyst-800. The true drift trajectories are generated by OpenDrift [14] using the vertically integrated ocean currents from the hourly NorKyst-800 data, which means that the underlying physical model for the simulated truth is significantly different from our much simpler shallow-water model. From these drift trajectories, we estimate the underlying direction and velocity of the ocean currents and use this as an observation in the particle filter. We assume that the observations contain a Gaussian error with standard deviation σ and that they are independent from each other. The weight of each ensemble member is then computed as

$$w_i = \alpha \cdot \exp\left(-0.5\left(\frac{\|\text{obs} - \text{sim}\|}{\sigma}\right)^2\right), \tag{3}$$

in which we use the Euclidean norm to compute the distance between the observed and simulated momentum. Furthermore, α is the normalization constant such that $\sum_{i=1}^{N_e} w_i = 1$. There are several strategies for choosing which ensemble members to discard and duplicate (see [2] and references therein), and we use the residual resampling scheme [15]. Between observation times, each ensemble member runs independently and deterministically, which means that we need to perturb the duplicated ensemble states during the resampling stage.

3 Results

We test our ensemble prediction system using a domain along the coast of Northern Norway. The domain consists of 315 × 630 grid cells with 800 m horizontal resolution, which is the same horizontal resolution as the operational ocean circulation model that we use for initial and boundary conditions. We run 48 h of data assimilation and use the final observed positions for the drifters as initial positions for 24 h trajectory forecasts. To avoid that the ensemble collapses during resampling, we need to balance the number of drifters we observe with the ensemble size. Here we run experiments with 1000 and 10,000 ensemble members and limit ourselves to four drifters. All experiments are run on an Nvidia DGX-2 server, equipped with 16 Tesla V100 GPUs and two CPUs, each with 24 cores.

Ensemble forecasts of drift trajectories We run three different experiments to illustrate the effect of data assimilation on forecasting of drift trajectories. The first is a Monte-Carlo experiment, meaning that we do not assimilate any observations during the first 48 h. The second and third experiments are with data assimilation, and we use observations to run a particle filter every 30 min and 5 min, respectively. Figure 2 shows the domain and the drift trajectories used as the truth.

Figure 3 shows the forecasted drift trajectories for the four drifters in each of the three ensemble experiments with 1000 members, compared to a single deterministic forecast in green (dashed) and the truth in red (dash-dotted) . The results show variable impact from data assimilation between the drifters. We see most positive effect for drifters three and four, and marginal improvement for drifter two, whereas the forecast for drifter one seems to be worse with data assimilation. The initial forecast for the first hour for drifter one, however, is significantly improved by the data assimilation, but our model is unable to capture the downward turn shortly into the forecast. This makes

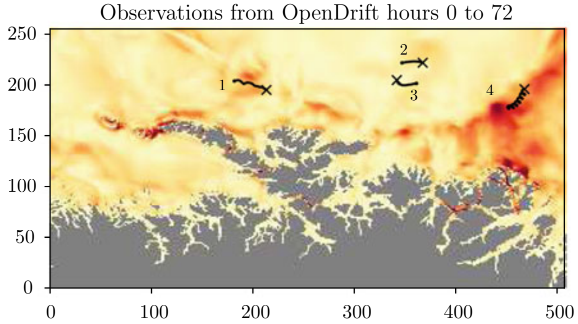


Fig. 2 Drift trajectories of four drifters over a three day period shown in the computational domain used in all our experiments. Red and yellow colors indicates strong and weak currents, respectively. Dots mark start positions and crosses end positions, and the values along the axes are in km

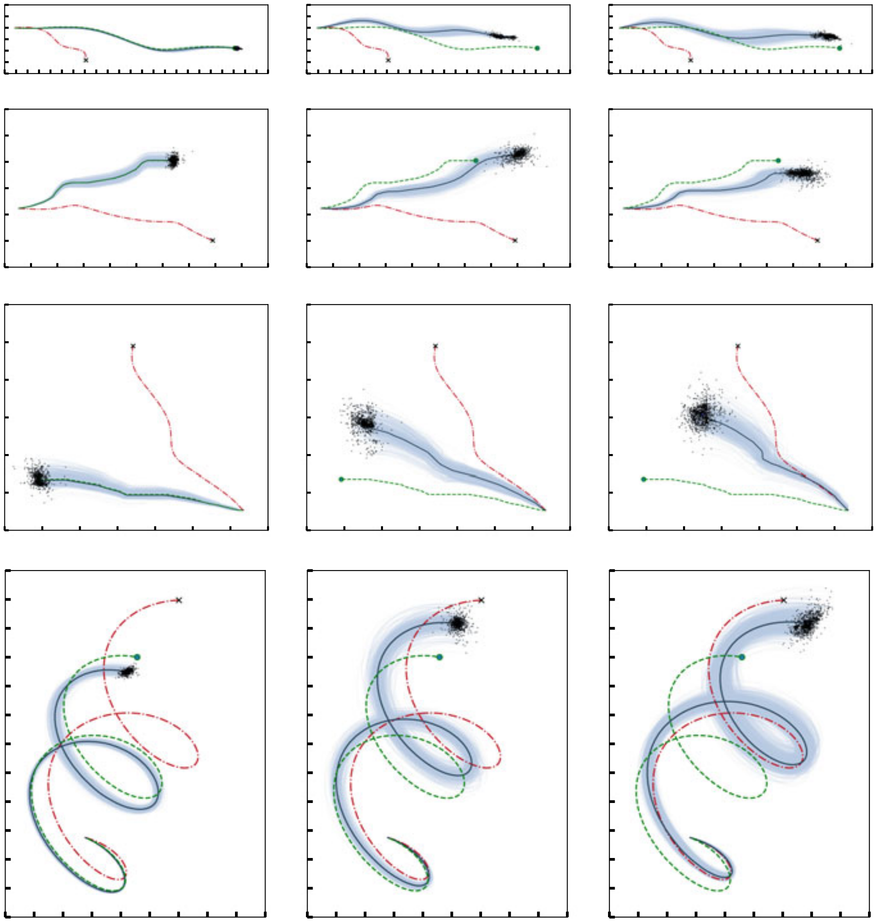


Fig. 3 Ensemble forecasts of drift trajectories with 1000 members in light blue, with the red (dash-dotted) line representing the truth, the green (dashed) line as the deterministic forecast, and the dark blue line as the ensemble mean. The four drifters are shown in separate rows, with the columns representing the three different experiments. From left to right: Monte Carlo without data assimilation, assimilation of observations every 30 min, and assimilation of observations every 5 min. The distance between the markers along the axis are one km

the ensemble perform worse in the long run when compared to the deterministic forecast. The results for drifters three and four show improvements when using observations in intervals of five minutes compared to 30 min. Finally, Fig. 4 shows how increasing the ensemble size to 10,000 members significantly improves the forecast for drifter 2. Increasing the ensemble size increases the sampling of the pdf, and thereby also the chance that the true state is better represented by the ensemble.

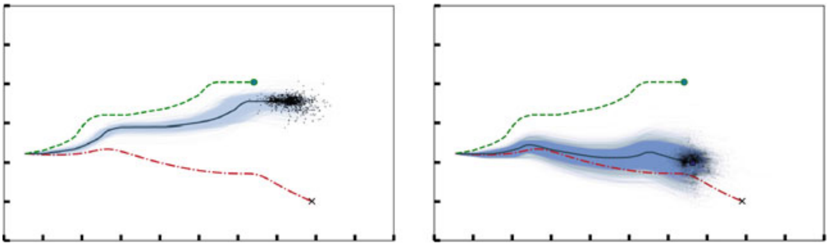


Fig. 4 Ensemble forcasting with 1000 ensemble members (left) and 10,000 members (right). When using 10,000 ensemble members, the forecast is significantly improved for drifter 2, while the effect is smaller for the other three drifters

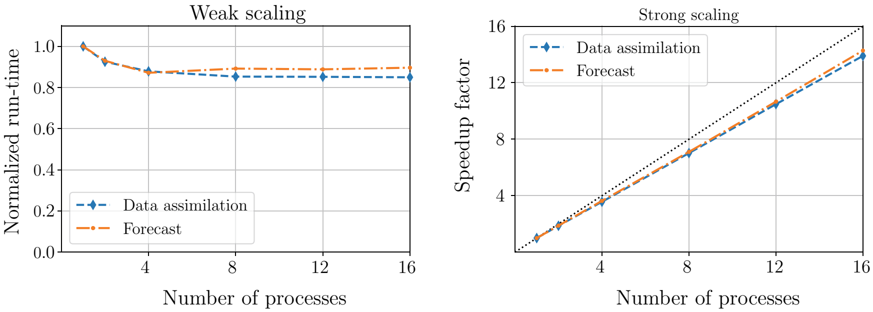


Fig. 5 The graphs show weak and strong scaling, using 1–16 GPUs on an Nvidia DGX-2 server. The top dotted line illustrates perfect strong scaling

Weak and strong performance scaling We evaluate the ensemble-level parallel performance by measuring the time spent in the data assimilation and forecasting parts of the code, running one hour of data assimilation with observations every five minutes and one hour of drift forecast. Note that the forecast contains no communication and should therefore show close to perfect scaling, whereas the data assimilation includes serial resampling and communication. For the weak scaling experiment, we fix the per-process ensemble size at 20 members and increase the number of processes from one to 16. In the strong scaling experiment, the global ensemble size is fixed at 960 members and we vary the number of processes on which they are distributed. Each process utilizes one GPU. The results are shown in Fig. 5, with both experiments showing a 14× speedup by using 16 GPUs for the forecast. The speedup for data assimilation is nearly as good as the forecast, which means that the data assimilation does a good job preserving the parallel performance.

4 Discussion and Summary

We have presented a framework for fully nonlinear data assimilation of ocean drift trajectories into a shallow-water model. The framework is implemented using the GPU for the shallow-water simulation and MPI for distribution of, and communication between, ensemble members. Our experiments show significantly increased forecast skill for simulations with data assimilation, and we are able to run over 10,000 ensemble members on the Nvidia DGX-2 with 16 GPUs. The results indicate, as expected, that data assimilation with 10,000 members based on 5 min sampling of the observed drifter positions yields a better forecast than 5 min sampling with 1000 members.

The results presented herein show that the data assimilation increases the forecast skill for three of four drifters. The forecast skill for drifter one, however, appears to be unaffected by the data assimilation, and we believe this is caused by a local predominant baroclinic ocean dynamic, which is not captured by our current simplified model. It will be an important future development to see what criteria are significant for the data assimilation to be most effective, and perhaps include a multi-layer or reduced-gravity model which can represent such dynamics better.

The simple perturbation strategy presented in this paper adds a smooth perturbation to the sea-surface level and computes the momentum required to keep the perturbation in geostrophic balance. An important extension will be to conduct more experiments with more sophisticated perturbation methods and stochastic placement of the ocean eddies, as well as perturbation of the tidal wave phase.

Acknowledgements This work is supported by the Research Council of Norway (RCN) through grant number 250935 (GPU Ocean). The computations in this paper were performed on equipment provided by the Experimental Infrastructure for Exploration of Exascale Computing (eX^3), which is financially supported by the RCN under contract 270053. The source code for the methods and experiments described in this paper is available under an GNU free and open source license released under https://doi.org/10.5281/zenodo.3591850.

References

1. Christensen, K., Breivik, Ø., Dagestad, K.-F., Röhrs, J., Ward, B.: Short-term predictions of oceanic drift. Oceanography **31**(3), 59–67 (2018)
2. van Leeuwen, P.: Particle filtering in geophysical systems. Mon. Weather Rev. **137**(12), 4089–4114 (2009)
3. Brodtkorb, A., Holm, H.: Real-world oceanographic simulations on the GPU using a two-dimensional finite volume scheme. preprint: arXiv:1912.02457 (in review) (2019)
4. NVIDIA. DGX-2/2H System user guide. Technical report (2019)
5. Vetra-Carvalho, S., van Leeuwen, P., Nerger, L., Barth, A., Altaf, M., Brasseur, P., Kirchgessner, P., Beckers, J.-M.: State-of-the-art stochastic data assimilation methods for high-dimensional non-Gaussian problems. Tellus A: Dyn. Meteorol. Oceanogr. **70**(1), 1–43 (2018)
6. Evensen, G.: Data Assimilation: The Ensemble Kalman Filter. Springer, Berlin Heidelberg (2006)

7. Snyder, C., Bengtsson, T., Bickel, P., Anderson, J.: Obstacles to high-dimensional particle filtering. Mon Weather Rev **136**(12), 4629–4640 (2008)
8. Shchepetkin, A., McWilliams, J.: The regional oceanic modeling system (ROMS): a split-explicit, free-surface, topography-following-coordinate oceanic model. Ocean Model. **9**(4), 347–404 (2005)
9. Martinsen, E., Gjevik, B., Røed, L.: A numerical model for long barotropic waves and storm surges along the western coast of Norway. J. Phys. Oceanogr. **9**, 1126–1138 (1979)
10. Røed, L.: Documentation of simple ocean models for use in ensemble predictions. Theory. Technical report, Norwegian Meteorological Institute, Part I (2012)
11. Chertock, A., Dudzinski, M., Kurganov, A., Lukácová-Medvidová, M.: Well-balanced schemes for the shallow water equations with Coriolis forces. Numerische Mathematik (2017)
12. Holm, H., Sætra, M., van Leeuwen, P.: Massively parallel implicit equal-weights particle filter for ocean drift trajectory forecasting. Journal of Computational Physics: X **6**(2590–0552), 100053 (2020). https://doi.org/10.1016/j.jcpx.2020.100053
13. Albretsen, J., Sperrevik, A., Staalstrøm, A., Sandvik, A., Vikebø, F.: NorKyst-800 report no. 1: User manual and technical descriptions. Technical Report 2, Institute of Marine Research (2011)
14. Dagestad, K.F., Röhrs, J., Breivik, Ø., Ådlandsvik, B.: OpenDrift v1.0: a generic framework for trajectory modelling. Geosci. Model Dev. **11**(4), 1405–1420) (2018)
15. Liu, J., Chen, R.: Sequential Monte Carlo methods for dynamic systems. J. Am. Stat. Assoc. **93**, 1032–1044 (1998)

Application of an Unstructured Finite Volume Method to the Shallow Water Equations with Porosity for Urban Flood Modelling

Abdelhafid Moumna, Imad Kissami, Imad Elmahi, and Fayssal Benkhaldoun

Abstract We present a finite volume model for the simulation of floods in urban areas. The model consists of the two-dimensional shallow water equations with variable horizontal porosity which is introduced in order to reflect the effects of obstructions. An extra porosity source term appears in the momentum equations. The main advantage of this model is the significant reduction of the computational cost while preserving an acceptable level of accuracy. The finite volume method uses a modified Roe's scheme involving the sign of the Jacobian matrix in the system for the discretization of gradient fluxes. The performance of the numerical model is demonstrated by comparing the results obtained using the proposed method to laboratory experiments for a flow problem over an array of obstacles.

Keywords Shallow water with porosity · Urban flood modelling · Finite volume method · SRNH scheme · Unstructured grids

MSC (2010) 65M08 · 65N08 · 35Q30

A. Moumna
FS, Département de Physique, Université Abdelmalek Essaadi,
BP. 2121, M'Hannech II, 93030 Tetouan, Morocco
e-mail: abdelhafid.moumna2018@gmail.com

I. Kissami (✉)
Mohammed VI Polytechnic University, CSEHS, Lot. 660,
43150 Ben Guerir, Morocco
e-mail: imad.kissami@um6p.ma

I. Elmahi
UMP, ENSAO - UM6P CSEHS Complexe universitaire,
B.P 669, 60000 Oujda, Morocco
e-mail: i.elmahi@ump.ac.ma

F. Benkhaldoun
LAGA, Université Paris 13, 99 Av J.B. Clement, 93430 Villetaneuse, France
e-mail: fayssal@math.univ-paris13.fr

R. Klöfkorn et al. (eds.), *Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples*, Springer Proceedings in Mathematics & Statistics 323,
https://doi.org/10.1007/978-3-030-43651-3_69

1 Introduction

Mathematical modelling of shallow water flows is based on the formulation and solution of the appropriate equations of continuity and momentum. In general, hydrodynamical flows represent a three-dimensional turbulent Newtonian flow in complicated geometrical domains. The cost of incorporating three-dimensional data in natural water courses is often excessively high. Computational efforts needed to simulate three-dimensional turbulent flows can also be significant. In view of such considerations, many researchers have tended to use rational approximations in order to develop two-dimensional hydrodynamical models for water flows. Indeed, under the influence of gravity, many free-surface water flows can be modeled by the shallow water equations with the assumption that the vertical scale is much smaller than any typical horizontal scale. The shallow water equations in depth-averaged form have been successfully applied to many engineering problems and their application fields include a wide spectrum of phenomena other than water waves.

The recent interest for flood simulation involving urbanized areas has also drawn the attention to the possible use of modified shallow water models with porosity for large scale flood simulations involving urbanized areas. Here, the porosity accounts for the reduction in storage and in the exchange sections due to the presence of buildings and other structures in the flood plain. The concept of porosity leads to a modification of the propagation equations (flux and source terms), an additional source term appears in the momentum equations. The porosity coefficient can be calculated by the ratio of the area available for the flow over the total area. Although similar in their structure to the source terms induced by the topographic gradient, the source term induced by the porosity also requires specific treatment of both momentum and continuity equations. The modified shallow water equations with porosity were first introduced in a simplified form by Defina et al. [4] and later modified by Hervouet et al. [5]. In this sense, it should be mentioned the work by Soares-Frazao et al. [8], who have proposed a finite volume solver for the two-dimensional shallow water equations with porosity based on the HLL scheme.

In the current study, a finite volume method is proposed for the numerical simulation of transient flows involving porosity variations. The method consists of a predictor stage where the numerical fluxes are constructed and a corrector stage to recover the conservation equations. The sign matrix of the Jacobian matrix is used in the reconstruction of the numerical fluxes. The method has been investigated in [2] for solving the canonical shallow water models without accounting for porosity variation. The current study presents an extension of this method to transient flows involving porosity variation in the water flows.

2 Shallow Water Equations in Porous Media

The shallow water equations with porosity are obtained by depth-averaging the three-dimensional incompressible Navier-Stokes equations under the assumptions that the vertical component of acceleration has a negligible effect on the water pressure (i.e., the pressure is hydrostatic). These equations can be written in conservation form as

$$\begin{aligned}
&\frac{\partial}{\partial t}(\phi h) + \frac{\partial}{\partial x}(\phi h u) + \frac{\partial}{\partial y}(\phi h v) = 0 \qquad (1)\\
&\frac{\partial}{\partial t}(\phi h u) + \frac{\partial}{\partial x}(\phi h u^2 + \frac{1}{2}\phi g h^2) + \frac{\partial}{\partial y}(\phi h u v) = -g\phi h\frac{\partial Z}{\partial x} + \frac{g}{2}h^2\frac{\partial \phi}{\partial x} - \frac{\tau_{f,x}}{\rho} - \frac{\tau_{d,x}}{\rho}\\
&\frac{\partial}{\partial t}(\phi h v) + \frac{\partial}{\partial x}(\phi h u v) + \frac{\partial}{\partial y}(\phi h v^2 + \frac{1}{2}\phi g h^2) = -g\phi h\frac{\partial Z}{\partial y} + \frac{g}{2}h^2\frac{\partial \phi}{\partial y} - \frac{\tau_{f,y}}{\rho} - \frac{\tau_{d,y}}{\rho}
\end{aligned}$$

where t is the time variable, $(x, y)^T$ the space coordinates, $(u, v)^T$ the depth-averaged water velocity, h the water depth, Z the bottom topography, g the gravitational acceleration and ϕ the porosity. The value of ϕ lies between 0 and 1, $\phi = 1$ means no solid structures in the control volume, and $\phi = 0$ means no water in the control volume. Note also that for a constant porosity ϕ, Eq. (1) reduce to the conventional shallow water equations widely investigated in computational hydraulics.

The variables Z and ϕ can be involved in the system to have a homogeneous system by adding the following two equations $\frac{\partial Z}{\partial t} = 0$ and $\frac{\partial \phi}{\partial t} = 0$, expressing the fact that the bottom Z and the porosity ϕ depend only on the space variables.

$\tau_{f,x}$ and $\tau_{f,y}$ are the components of the bed friction stress, they are defined by

$$\frac{\tau_{f,x}}{\rho} = \phi g h\frac{\eta^2|U|}{h^{4/3}}u, \qquad \frac{\tau_{f,y}}{\rho} = \phi g h\frac{\eta^2|U|}{h^{4/3}}v \qquad (2)$$

where η represents the Manning friction coefficient and ρ the water density.

$\tau_{d,x}$ and $\tau_{d,y}$ represent the drag force components that water exerts on obstructions when it is in motion, it is a parallel and opposite force to the flow. This additional drag source term can be expressed, as [8], by

$$\frac{\tau_{d,x}}{\rho} = \frac{1}{2}\frac{NhL_x}{A}C_x|U|u, \qquad \frac{\tau_{d,y}}{\rho} = \frac{1}{2}\frac{NhL_y}{A}C_y|U|v \qquad (3)$$

where C_x, C_y are the drag coefficients following the two horizontal directions x and y respectively, $|U|$ is the averaged velocity modulus of the water, L_x, L_y are the obstructions projection lengths inside the urban region in the x and y directions, and A is the area of a region in which there are N obstructions, each of them with a horizontal surface S. In this situation the surface porosity is given by $\phi = 1 - NS/A$.

For simplicity in the presentation, Eq. (1) are reformulated in a compact vector form as

$$\frac{\partial \mathbf{W}}{\partial t}+\frac{\partial \mathbf{F}(\mathbf{W})}{\partial x}+\frac{\partial \mathbf{G}(\mathbf{W})}{\partial y}=\mathbf{S}_1(\mathbf{W})+\mathbf{S}_2(\mathbf{W}) \qquad (x,y)\in \Omega \tag{4}$$

where $\mathbf{W}$ is the vector of variables representing conserved quantities, $\mathbf{S}_1(\mathbf{W})$ and $\mathbf{S}_2(\mathbf{W})$ define the source terms in the system (1), $\mathbf{F}(\mathbf{W})$ and $\mathbf{G}(\mathbf{W})$ are the advection flow functions.

$$\mathbf{W}=\begin{pmatrix}\phi h\\ \phi hu\\ \phi hv\\ Z\\ \phi\end{pmatrix},\quad \mathbf{F}(\mathbf{W})=\begin{pmatrix}\phi hu\\ \frac{1}{2}\phi g h^2+\phi hu^2\\ \phi huv\\ 0\\ 0\end{pmatrix},\quad \mathbf{G}(\mathbf{W})=\begin{pmatrix}\phi hv\\ \phi huv\\ \frac{1}{2}\phi g h^2+\phi hv^2\\ 0\\ 0\end{pmatrix}.$$

$$\mathbf{S}_1(\mathbf{W})=\begin{pmatrix}0\\ -g\phi h\frac{\partial Z}{\partial x}+\frac{g}{2}h^2\frac{\partial \phi}{\partial x}\\ -g\phi h\frac{\partial Z}{\partial y}+\frac{g}{2}h^2\frac{\partial \phi}{\partial y}\\ 0\\ 0\end{pmatrix},\qquad \mathbf{S}_2(\mathbf{W})=\begin{pmatrix}0\\ -\frac{\tau_{f,x}}{\rho}-\frac{\tau_{d,x}}{\rho}\\ -\frac{\tau_{f,y}}{\rho}-\frac{\tau_{d,y}}{\rho}\\ 0\\ 0\end{pmatrix}.$$

3 Application of the SRNH Scheme

3.1 Finite Volume Discretization

The finite volume method is based on two main steps: the discretization of the computational domain into a finite number of control volumes, and the integration of the system of equations on each control volume.

To discretize the space domain $\bar{\Omega}=\Omega\cup\Gamma$ we use conforming triangular elements T_i such as $\bar{\Omega}=\cup_{i=1}^{Ne}T_i$, with Ne is the total number of elements. Each triangle represents a control volume and the variables are located at the geometric centres of the cells. Hence, a finite volume discretization of (4) yields

$$\frac{\partial \mathbf{W}_i}{\partial t}+\frac{1}{|T_i|}\sum_{j\in N(i)}\int_{\Gamma_{ij}}\mathscr{F}(\mathbf{W};\mathbf{n})\,d\sigma=\frac{1}{|T_i|}\int_{T_i}\mathbf{S_1}(\mathbf{W})\,dV+\frac{1}{|T_i|}\int_{T_i}\mathbf{S_2}(\mathbf{W})\,dV \tag{5}$$

where $|T_i|$ is the area of the element T_i and $N(i)$ the set of neighboring cells of T_i, $\mathbf{W}_i^n$ an average value of the solution $\mathbf{W}$ in T_i and $\mathscr{F}(\mathbf{W};\mathbf{n})$ is the physical flux function.

Using these notations, the semi-discrete equations (5) become

$$\frac{\partial \mathbf{W}_i}{\partial t} = -\frac{1}{|T_i|} \sum_{j \in N(i)} \mathscr{F}\left(\mathbf{W}_{ij}; \mathbf{n}_{ij}\right) \left|\Gamma_{ij}\right| + \mathbf{S}_{1i} + \mathbf{S}_{2i}. \tag{6}$$

The spatial discretization of the system (4) is complete once a reconstruction is chosen for the numerical fluxes and source terms in (6).

3.2 Discretization of the Gradient Fluxes

In this section, we extend and apply the Non Homogeneous Riemann Solver (SRNH) to the discretization of the shallow water system with porosity. Recall that the SRNH scheme is dedicated to the approximation of the hydrodynamic part of the system and the source terms coming from the bed variations and drag forces. It consists in two stages, a predictor stage in which the state variables on each interface of the mesh are evaluated by solving a Riemann problem projected along the normal and the tangential using an upwind scheme, and a corrector stage in which the time incrementation is performed by calculating the physical flux at the average states obtained in the predictor stage.

Let us then consider the advection part of equations (4) with the source terms of bed variations and drag forces

$$\frac{\partial \mathbf{W}}{\partial t} + \frac{\partial \mathbf{F}(\mathbf{W})}{\partial x} + \frac{\partial \mathbf{G}(\mathbf{W})}{\partial y} = \mathbf{S}_1(\mathbf{W}) \tag{7}$$

Using the expressions of the normal velocity $u_\eta = un_x + vn_y$ and tangential velocity $u_\tau = -un_y + vn_x$, we can reformulate the projected equations associated with (7) as

$$\begin{aligned}
\frac{\partial (\phi h)}{\partial t} + \frac{\partial \left(\phi h u_\eta\right)}{\partial \eta} &= 0,\\
\frac{\partial \left(\phi h u_\eta\right)}{\partial t} + \frac{\partial}{\partial \eta}\left(\phi h u_\eta^2 + \frac{1}{2} g \phi h^2\right) &= -g\phi h \frac{\partial Z}{\partial \eta} + \frac{1}{2} g h^2 \frac{\partial \phi}{\partial \eta},\\
\frac{\partial (\phi h u_\tau)}{\partial t} + \frac{\partial}{\partial \eta}\left(\phi h u_\eta u_\tau\right) &= 0,\\
\frac{\partial Z}{\partial t} = 0, \quad \frac{\partial \phi}{\partial t} &= 0.
\end{aligned} \tag{8}$$

Notice that the last two equations in (8) have been included in the system to formulate the projected system in an advection form without source terms. Thus, an equivalent system of (8) can also be rewritten in a vector form as

$$\frac{\partial \mathbf{U}}{\partial t} + \mathbf{A}_\eta(\mathbf{U})\frac{\partial \mathbf{U}}{\partial \eta} = \mathbf{0} \tag{9}$$

where

$$\mathbf{U} = \begin{pmatrix} \phi h \\ \phi h u_\eta \\ \phi h u_\tau \\ Z \\ \phi \end{pmatrix}, \qquad \mathbf{A}_\eta(\mathbf{U}) = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ gh - u_\eta^2 & 2u_\eta & 0 & \phi gh & -gh^2 \\ -u_\eta u_\tau & u_\tau & u_\eta & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Step 1: **Predictor Stage**

The predictor stage of the (SRNH) scheme consists in using the projected system to compute the average states $\mathbf{U}_{ij}^n$ on each interface Γ_{ij} between two control volumes T_i and T_j. It is formulated using an upwinding as

$$\mathbf{U}_{ij}^n = \frac{1}{2}\left(\mathbf{U}_i^n + \mathbf{U}_j^n\right) - \frac{1}{2}\,\mathrm{sgn}\Big[\mathbf{A}_\eta\left(\overline{\mathbf{U}}\right)\Big]\left(\mathbf{U}_j^n - \mathbf{U}_i^n\right). \tag{10}$$

In (10), the sign matrix of the Jacobian is defined by

$$\mathrm{sgn}\Big[\mathbf{A}_\eta\left(\overline{\mathbf{U}}\right)\Big] = \mathscr{R}(\overline{\mathbf{U}})\,\mathrm{sgn}\Big[\Lambda(\overline{\mathbf{U}})\Big]\mathscr{R}^{-1}(\overline{\mathbf{U}}),$$

with Λ is the diagonal matrix of eigenvalues, and $\mathscr{R}$ is the right eigenvector matrix. This sign matrix must be evaluated in the averaged state of Roe $\overline{\mathbf{U}}$ given in [1].

Step 2: **Corrector Stage**

The predictor stage (10) makes it possible to determine the projected convective states $\mathbf{U}_{ij}^n$ on each interface Γ_{ij}. The non-projected conservative states $\mathbf{W}_{ij}^n$ are then reconstructed using the transformations $u = u_\eta n_x - u_\tau n_y$ and $v = u_\eta n_y + u_\tau n_x$.

The incremental step is then written using the physical flux evaluated on these conservative states. With a first order Euler method for time integration, the corrector stage writes

$$\mathbf{W}_i^{n+1} = \mathbf{W}_i^n - \frac{\Delta t}{|T_i|}\sum_{j\in N(i)} \mathscr{F}(\mathbf{W}_{ij}^n; n_{ij})|\Gamma_{ij}| + \Delta t \mathbf{S}_{1i}^n. \tag{11}$$

The treatment of the source terms in the shallow water equations presents a challenge in many numerical methods, compare [2, 3, 7] among others. In our solver, the source term approximation $\mathbf{S}_{1i}^n$ in the corrector stage is reconstructed in such a way to ensure a well balanced scheme preserving positivity of the water depth. For the details on the reconstruction, the reader can refer to [1].

4 Numerical Results and Examples

To assess the performance of our method and the model, we present numerical simulations of floods due to dam-break in a channel containing a simplified city. We also compare the computed results to those obtained using the conventional shallow water equations and experimental data [6]. The simulations have been performed using a time stepsize Δt that is adjusted at each step according to the stability condition

$$\Delta t = Cr \min_{\Gamma_{ij}} \left(\frac{|T_i| + |T_j|}{2\,|\Gamma_{ij}|\,\max_p |(\lambda^p)_{ij}|} \right),$$

where Γ_{ij} is the edge between two cells T_i and T_j, $(\lambda^p)_{ij}$ denotes the pth eigenvalue evaluated on Γ_{ij} and Cr the Courant number taken here equal to 0.8.

This test case was proposed by the team of the Catholic University of Louvain to simulate the risks of floods. Inside the channel, a simplified city is arranged with a set of buildings distributed in staggered rows (see the left Fig. 1). The channel has a length $L = 36$ m and a width $l = 3.6$ m. A section narrowing and a door are arranged in order to simulate dam-break.

In the area containing the buildings, the porosity coefficient ϕ is computed by

$$\phi = 1 - \frac{S}{S_t} = 1 - \frac{22 * 0.3^2}{1.9^2} = 0.45.$$

For the rest of the channel, ϕ takes the value 1. The right Fig. 1 shows a 2D distribution of the porosity in the channel. The break is located at the gate of the channel which is kept open and a rate $Q = 0.09\,\text{m}^3/\text{s}$ is imposed there. The Manning's coefficient is taken equal to $0.01\,\text{s/m}^{1/3}$. The walls are considered solid, while Neumann conditions are imposed at the exit of the channel.

We compare two models with free surface: the first is Classical Saint-Venant model, noted STVC, which considers that buildings are represented by solid boundaries. These buildings are therefore not meshed but taken into consideration in the boundary conditions. The second model is Saint-Venant with porosity, noted STVP. This model as already indicated considers the entire channel in the generation of the mesh, including the urban areas. The areas of the simplified city (buildings) are taken into consideration by defining in these zones a porosity $\phi \neq 1$ (here $\phi = 0.45$)

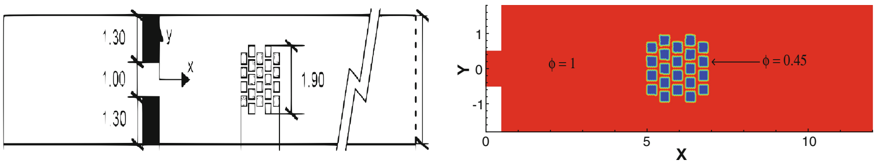


Fig. 1 Experimental device (left) and 2D distribution of the porosity in the channel (right)

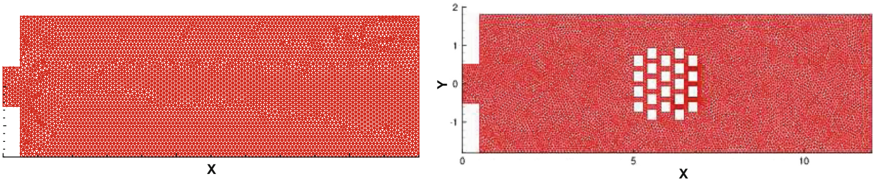


Fig. 2 Mesh used by the STVP (left) and STVC (right) models

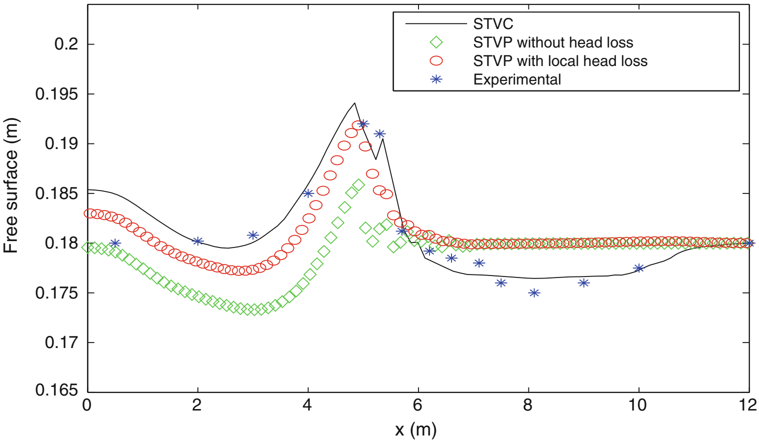


Fig. 3 Profiles of the free surface at physical time $t = 400\,\text{s}$ along the axis $y = 0\,\text{m}$

in these zones. Figure 2 shows the meshes used for both STVP and STVC models, containing respectively 6082 and 16232 elements.

In the Fig. 3, we present the free surface at the physical time $t = 400\,\text{s}$ along the axis $y = 0\,\text{m}$. A comparison is made between the classical Saint-Venant model (STVC), the Saint-Venant model with porosity (STVP) with and without loss of charge, and experimental data. It is clear from the figure that the STVC model gives consistent results with the experimental one.

Figure 3 also shows that although the simulation of the STVP model generally finds the behavior of the flow, the profile of the free surface obtained by neglecting the head losses ($S_x = S_y = 0$) is quite different from the experimental one, which shows that the porosity is not sufficient to represent the influence of the urban area.

The interest of taking into account the head loss terms is also clearly illustrated in the figure. Taking ($S_x = 4$, $S_y = 2$), we see that the profile of the free surface is significantly improved.

Figure 4 shows the velocity fields obtained by the two models STVC and STVP. It is clear that the STVC simulation gives results very close to the experimental measurements, so it can be used as a reference for the validation of the STVP model. In the left Fig. 4, one observes the good behavior of the flow at the crossroads. The extension of the very low velocity zone, located downstream the city, is also

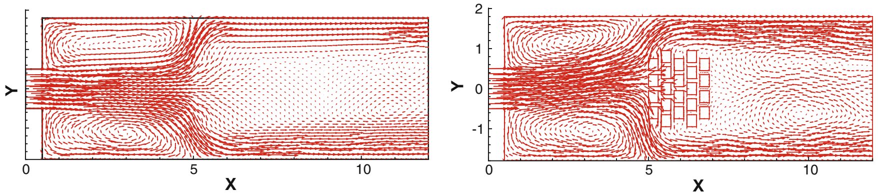


Fig. 4 Velocity fields obtained by the STVP (left) and the STVC (right)

well simulated, but the orientation of the velocity vectors is a little different in the two simulations. On a large scale, the two simulated velocity fields are very close. The main characteristics of the flow can be assumed to be well reproduced by the simulation of the STVP model.

5 Conclusions

A study of flood risks in channels with urbanized areas was considered. As an alternative to the classical Saint-Venant model, we proposed a model of Saint-Venant with porosity. This new model does not require the consideration of urban areas in the mesh generation, but rather a variable porosity is introduced. The test case considered as well as the comparisons with the experimental data, showed the performance of the STVP model and the developed finite volume solver. These results also showed that the porosity alone cannot adequately represent the flow in the crossroads and that the head loss terms are necessary in the model.

References

1. Benkhaldoun, F., Elmahi, I., Moumna, A., Seaid, M.: A non-homogeneous riemann solver for shallow water equations in porous media. J. Appl. Anal. **95**, 2181–2202 (2016)
2. Benkhaldoun, F., Elmahi, I., Seaid, M.: A new finite volume method for flux-gradient and source-term balancing in shallow water equations. Comp. Meth. Appl. Mech. Eng. **199**, 49–52 (2010)
3. Bermudez, A., Vazquez, M.E.: Upwind methods for hyperbolic conservation laws with source terms. Comput. Fluids **23**, 1049–1071 (1994)
4. Defina, A., D'Alpaos, L., Mattichio, B.: A new set of equations for very shallow water and partially dry area suitable to 2d numerical domains. In: Proceedings Specialty Conference Modelling of Flood Propagation over Initially Dry Areas. Milano, Italy, 29 June, 1 July 1994
5. Hervouet, J., Samie, R., Moreau, B.: Modelling urban areas in dam-break flood-wave numerical simulations. In: Proceedings of the International Seminar and Workshop on Rescue Actions Based on Dam-break Flow Analysis. Seinâjoki, Finland, 16 October 2000
6. Lhomme, J., Soares-Frazao, S., Guinot, V., Zech, Y.: Large-scale urban floods modelling and two-dimensional shallow water models with porosity. In: 7th International Conference on Hydroinformatics. Nice, France, 16 October 2006

7. Noelle, S., Pankratz, N., Puppo, G., Natvig, J.: Well-balanced finite volume schemes of arbitrary order of accuracy for shallow water flows. J. Comput. Phys. **213**, 4447–4499 (2006)
8. Soares-Frazao, S., Lhomme, J., Guinot, V., Zech, Y.: Two-dimensional shallow-water model with porosity for urban flood modeling. J. Hydraul. Res. **46**(1), 45–64 (2008)

Semi-implicit Two-Speed Well-Balanced Relaxation Scheme for Ripa Model

Emmanuel Franck and Laurent Navoret

Abstract In this paper, we propose a semi-implicit well-balanced scheme for the Ripa model based on a two-speed relaxation. The method both preserves equilibria and has an implicit step that reduces to the inversion of a constant Laplacian. Numerical simulations show that the scheme well capture low-Froude flows.

Keywords Finite volume · Relaxation method · Shallow-water model · Semi-implicit

MSC (2010) 65M08 · 65N08 · 35Q30

1 Introduction

To discretize two-scale hyperbolic problems with a good accuracy, several methods have been proposed and are generally based on semi-implicit schemes to prevent us from the fast scale stringent stability condition. In [1, 2], such a scheme have been proposed and it is based on very recent relaxation method and a dynamical splitting. This method allows to adapt the time discretization to the regime and the implicit step just reduces to the inversion of a discrete Laplacian with constant coefficient. In this paper, we adapt this numerical scheme to the Ripa model with topography, which describes shallow-water flows with horizontal temperature gradients. In the Ripa model, the fast scale dynamics are perturbative gravity waves around an equilibrium and the slow scale dynamics is the convection. Here equilibria are balance between the pressure gradient and the topography source term. In addition to the implicit treatment of the perturbative waves, the scheme has to preserve the equilbria so to

E. Franck (✉) · L. Navoret
INRIA Grand-Est and IRMA Strasbourg, 7 Rue René Descartes,
67000 Strasbourg, France
e-mail: emmanuel.franck@inria.fr

L. Navoret
e-mail: laurent.navoret@math.unistra.fr

R. Klöfkorn et al. (eds.), *Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples*, Springer Proceedings in Mathematics & Statistics 323,
https://doi.org/10.1007/978-3-030-43651-3_70

prevent the generation of spurious waves around them. This is the so-called well-balanced property.

We consider the one-dimensional Ripa model [3]:

$$\begin{cases} \partial_t h + \partial_x(hu) = 0, \\ \partial_t(hu) + \partial_x(hu^2 + p(h, \Theta)) = -gh\Theta\partial_x z, \\ \partial_t(h\Theta) + \partial_x(h\Theta u) = 0, \end{cases} \tag{1}$$

where $h(x, t)$ is the water height, $u(x, t)$ the velocity, $\Theta(x, t)$ the temperature and $z(x)$ the topography and the pressure law is given by: $p(h, \Theta) = g\Theta\frac{1}{2}h^2$, with g the gravity constant. This system is hyperbolic and has three characteristic speeds: $\Sigma = \{u - c, u, u + c\}$, with $c = \sqrt{gh\Theta}$. We introduce the Froude number $\mathrm{Fr} = u/c$, which describes the ratio between the two velocity scales. We are interested into perturbations of stationary solutions. These solutions are obtained by the balance law between pressure force and source term: $\partial_x p = -gh\Theta\partial_x z$. Hereafter we consider the following three families of equilibria, with zero velocity:

$$\begin{cases} u = 0, \\ \Theta = \text{cst}, \\ h + z = \text{cst}, \end{cases} \qquad \begin{cases} u = 0, \\ z = \text{cst}, \\ \Theta\frac{h^2}{2} = \text{cst}, \end{cases} \qquad \begin{cases} u = 0, \\ h = \text{cst}, \\ z + \frac{h}{2}\ln(\Theta) = \text{cst}. \end{cases} \tag{2}$$

The first aim is to write a scheme that preserves these steady states. Indeed, if the scheme does not preserve the steady states, spurious velocity and pressure modes would appear and would destroy the accuracy of the scheme for small velocity (low Froude regime). The second aim is to capture the dynamics near an equilibrium with an acceptable cost. For example, we consider the following perturbation $O(\mathrm{Fr})$ of the first steady equilibria with $\mathrm{Fr} \ll 1$. In that case, the perturbation has a small amplitude but moves with a large propagation speed of order $O(1/\mathrm{Fr})$. Therefore, implicit schemes are usually required to filter these small fast waves.

2 Two-Speed Relaxation System

To simplify the implicit treatment of the dynamics, we propose a relaxation model that linearizes the fast scale associated to the gravity waves. We introduce two additional unknows $\Pi(x, t)$ and $v(x, t)$ and consider the following extended hyperbolic system

$$\begin{cases} \partial_t h + \partial_x(hv) = 0, \\ \partial_t(hu) + \partial_x(huv + \Pi) = -gh\partial_x z, \\ \partial_t(h\Theta) + \partial_x(h\Theta v) = 0, \\ \partial_t \Pi + v\,\partial_x \Pi + h_m\lambda^2\partial_x v = \frac{1}{\varepsilon}\Big(p(h, \Theta) - \Pi\Big) \\ \partial_t v + v\,\partial_x v + \frac{1}{h_m}\partial_x \Pi = -\frac{h}{h_m}g\Theta\partial_x z + \frac{1}{\varepsilon}\Big(v - u\Big) \end{cases} \tag{3}$$

with $h_m > 0$ and $\lambda > 0$ are constant relaxation parameters and where $\varepsilon > 0$ is the relaxation parameter. This system is an approximation of (1) in the limit ε tends to zero. This result can be shown formally.

Proposition 1 *As $\varepsilon \to 0$, the relaxation system (3) is consistent at first order in ε with*

$$\begin{cases} \partial_t h + \partial_x(hu) = \varepsilon\,\partial_x(\beta(\partial_x p + gh\partial_x z)), \\ \partial_t(hu) + \partial_x(hu^2 + \Theta\frac{g}{2}h^2) = -hg\Theta\partial_x z + \varepsilon\,\partial_x\,(u\beta(\partial_x p + gh\partial_x z)) + \varepsilon\partial_x(\gamma\partial_x u) \\ \partial_t(h\Theta) + \partial_x(h\Theta u) = \varepsilon\partial_x\,(\Theta\beta(\partial_x p + gh\partial_x z))\,, \end{cases} \tag{4}$$

with $\beta = \left(\frac{h}{h_m} - 1\right)$, $\gamma = \left(h_m\lambda^2 - hc^2\right)$.

The proof is based on a classical Chapman-Enskog expansion detailed for the Euler system case in [2]. Note that an equilibrium given by $\partial_x p = -gh\Theta\partial_x z$ and $u = 0$ is still a steady state of the first order approximation (4) of the relaxation system. It is not the case for all the relaxation models like the Jin-Xin relaxation used in [1]. This property is necessary, but not sufficient, to obtain a well-balanced scheme since we will discretize the relaxation system (3) and not the original one. In [2], the entropy stability of (4) is analysed. This computation adapted to our model gives the following stability conditions: $\beta \geq 0$, $\gamma \geq 0$.

3 Semi-implicit Scheme

The structure of the relaxation system (3) enables us to devise a semi-implicit scheme with a simple implicit part. This scheme is based on a splitting method between the different time-scale dynamics contained in (3). We split system (3) into the convection part (C), the gravity waves part (W) and the relaxation part (R):

$$(C) \quad \begin{cases} \partial_t h + \partial_x(hv) = 0, \\ \partial_t(hu) + \partial_x(huv + \mathcal{F}^2\Pi) = -\mathcal{F}^2\, gh\Theta\partial_x z, \\ \partial_t(h\Theta) + \partial_x(h\Theta v) = 0, \\ \partial_t\Pi + v\partial_x\Pi + h_m\lambda^2\partial_x v = 0 \\ \partial_t v + v\partial_x v + \frac{\mathcal{F}^2}{h_m}\partial_x\Pi = -\mathcal{F}^2\,\frac{h}{h_m}g\Theta\partial_x z \end{cases}$$

$$(W) \quad \begin{cases} \partial_t h = 0, \\ \partial_t(hu) + (1-\mathcal{F}^2)\,(\partial_x\Pi + hg\partial_x z) = 0, \\ \partial_t h\Theta = 0 \\ \partial_t\Pi + (1-\mathcal{F}^2)h_m\lambda^2\partial_x v = 0 \\ \partial_t v + (1-\mathcal{F}^2)\left(\frac{1}{h_m}\partial_x\Pi + \frac{h}{h_m}g\partial_x z\right) = 0 \end{cases}$$

$$(R) \quad \begin{cases} \partial_t\Pi = \dfrac{1}{\varepsilon}\,(p(h,\Theta) - \Pi)\,, \quad \partial_t v = \dfrac{1}{\varepsilon}\,(u - v)\,, \end{cases}$$

where $\mathcal{F} = \max\left(\mathcal{F}_{\min}, \min\left(\frac{u}{\sqrt{h\Theta g}}, 1\right)\right)$ is an estimation of the global Froude number. Introduced in [4], this dynamic splitting allows to adapt itself to the dynamics.

In practice, the relaxation step (R) is treated as a projection: $\Pi = p(h, \Theta)$ and $v = u$. The key point is the discretization of the waves part (W) with an implicit solver. First, we note that $\partial_t h = 0$ so that the topography source term can be treated explicitly. The construction of the implicit scheme is based on the following remark: the equations on v and Π form a linear independent system. We can thus discretize all the equations of (W) with an implicit Euler scheme and then we get an implicit elliptic problem on Π^{n+1} by plugging the expression of v^{n+1} in the Π equation. This elliptic problem is linear and has constant coefficient. After solving this problem, we obtain Π^{n+1} and then v^{n+1} and $(hu)^{n+1}$ can be computed with an explicit cost. For the spatial discretization, we use a classical finite volume scheme for the elliptic part and the central flux for the first derivative. The final algorithm writes:

- Step 1: solve

$$\left(\Pi_j^{n+1} - (1-\mathcal{F}^2)^2\Delta t^2\lambda^2 \frac{\Pi_{j+1}^{n+1} - 2\Pi_j^{n+1} + \Pi_{j-1}^{n+1}}{\Delta x^2}\right) =$$

$$\Pi_j^n - \Delta t(1-\mathcal{F}^2)\lambda^2 \frac{v_{j+1}^n - v_{j-1}^n}{2\Delta x}(1-\mathcal{F}^2)^2\Delta t^2\lambda^2 \frac{1}{\Delta x}\left(S_{j+\frac{1}{2}}^n - S_{j-\frac{1}{2}}^n\right),$$

  with

$$S_{j+\frac{1}{2}}^n = h_{j+\frac{1}{2}}^n \Theta_{j+\frac{1}{2}}^n \frac{z_{j+1} - z_j}{\Delta x}, \tag{5}$$

  where the quantities $h_{j+\frac{1}{2}}$ and $\Theta_{j+\frac{1}{2}}$ will be define below.
- Step 2: compute

$$v_j^{n+1} = v_j^n - (1-\mathcal{F}^2)\frac{\Delta t}{h_m}\frac{\Pi_{j+1}^{n+1} - \Pi_{j-1}^{n+1}}{2\Delta x} - (1-\mathcal{F}^2)\frac{\Delta t}{h_m}\frac{g}{2}\left(S_{j+\frac{1}{2}}^n - S_{j-\frac{1}{2}}^n\right),$$

$$(hu)_j^{n+1} = (hu)_j^n - \Delta t(1-\mathcal{F}^2)\frac{\Pi_{j+1}^{n+1} - \Pi_{j-1}^{n+1}}{2\Delta x} - \frac{g\Delta t}{2}(1-\mathcal{F}^2)\left(S_{j+\frac{1}{2}}^n - S_{j-\frac{1}{2}}^n\right).$$

4 Well-Balanced Fluxes

The convective part (C) is discretized with a first order explicit finite volume scheme. The numerical flux for the transport terms is constructed so that it preserves the steady states and this will provide the value of $h_{j+\frac{1}{2}}$ and $\Theta_{j+\frac{1}{2}}$. Following [5, 6], the idea consists in splitting the flux and then using a specific numerical fluxes for each part. Here, we propose to decompose the flux term in three parts:

$$\begin{pmatrix} \partial_x(hv), \\ \partial_x(huv+\mathcal{F}^2\Pi) \\ \partial_x(h\Theta v) \\ v\partial_x\Pi + h_m\lambda^2\partial_x v \\ v\partial_x v + \frac{\mathcal{F}^2}{h_m}\partial_x\Pi \end{pmatrix} = \partial_x \begin{pmatrix} hv \\ huv+\mathcal{F}^2\Pi \\ h\Theta v \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ 0 \\ v\partial_x\Pi \\ v\partial_x v \end{pmatrix} + \partial_x \begin{pmatrix} 0 \\ 0 \\ 0 \\ h_m\lambda_c v \\ \frac{1}{h_m}\mathcal{F}^2\Pi \end{pmatrix}$$
$$= \partial_x F_c + F_{nc}^* + \partial_x F_{\ell a}$$

For the non-conservative part F_{nc}^*, we use a non-conservative upwind scheme. For the linear acoustic part $F_{\ell a}$, we propose a modified acoustic flux that takes into account spatial variation due to the source term using the Jin-Levermore method [7]. The linear acoustic part writes:

$$\partial_t\Pi + h_m\lambda^2\partial_x v = 0, \quad \partial_t v + \mathcal{F}^2\frac{1}{h_m}\partial_x\Pi = 0,$$

and can be diagonalized as follows:

$$\begin{cases} \partial_t\left(\frac{h_m\lambda}{\mathcal{F}}v - \Pi\right) - \mathcal{F}\lambda\partial_x\left(\frac{h_m\lambda}{\mathcal{F}}v - \Pi\right) = 0, \\ \partial_t\left(\frac{h_m\lambda}{\mathcal{F}}v + \Pi\right) + \mathcal{F}\lambda\partial_x\left(\frac{h_m\lambda}{\mathcal{F}}v + \Pi\right) = 0. \end{cases}$$

To define the intermediate value of Π and v in the fluxes at node $x_{j+\frac{1}{2}}$, we thus consider the upwinded quantities:

$$\begin{cases} \left(\frac{h_m\lambda v}{\mathcal{F}} - \Pi\right)_{j+\frac{1}{2}} = \frac{h_m\lambda}{\mathcal{F}}v(x^+_{j+\frac{1}{2}}) - \Pi(x^+_{j+\frac{1}{2}}), \\ \left(\frac{h_m\lambda v}{\mathcal{F}} + \Pi\right)_{j+\frac{1}{2}} = \frac{h_m\lambda}{\mathcal{F}}v(x^-_{j+\frac{1}{2}}) + \Pi(x^-_{j+\frac{1}{2}}). \end{cases} \tag{6}$$

Following [7] and as already done for the Euler gravity system in [8], we precise these formula by considering the possible spatial variation of Π at equilibria:

$$\Pi(x_j) \approx \Pi(x^-_{j+\frac{1}{2}}) - \frac{\Delta x}{2}\partial_x\Pi(x_{j+\frac{1}{2}}) \approx \Pi(x^-_{j+\frac{1}{2}}) + \frac{\Delta x}{2}h(x_{j+\frac{1}{2}})\Theta(x_{j+\frac{1}{2}})g\partial_x z(x_{j+\frac{1}{2}}).$$

We define similarly $\Pi(x^+_{j+\frac{1}{2}})$. Then plugging these values in (6) and considering $v(x^+_{j+\frac{1}{2}}) = v(x_{j+1})$ and $v(x^-_{j+\frac{1}{2}}) = v(x_j)$, we obtain at the discrete level:

$$\begin{cases} \left(\frac{h_m\lambda v}{\mathcal{F}} - \Pi\right)_{j+\frac{1}{2}} = \frac{h_m\lambda}{\mathcal{F}}v_{j+1} - \Pi_{j+1} + \frac{\Delta x}{2}h_{j+\frac{1}{2}}\Theta_{j+\frac{1}{2}}g\frac{z_{j+1}-z_j}{\Delta x}, \\ \left(\frac{h_m\lambda v}{\mathcal{F}} + \Pi\right)_{j+\frac{1}{2}} = \frac{h_m\lambda}{\mathcal{F}}v_j + \Pi_j - \frac{\Delta x}{2}h_{j+\frac{1}{2}}\Theta_{j+\frac{1}{2}}g\frac{z_{j+1}-z_j}{\Delta x}. \end{cases} \tag{7}$$

Then we obtain the following intermediate values for the modified acoustic fluxes:

$$\begin{cases} v^*_{j+\frac{1}{2}} = \frac{1}{2}\left(v_{j+1}+v_j\right) - \frac{\mathcal{F}}{2h_m\lambda_c}\left(\Pi_{j+1} - \Pi_j + gh_{j+\frac{1}{2}}\Theta_{j+\frac{1}{2}}(z_{j+1}-z_j)\right), \\ \Pi^*_{j+\frac{1}{2}} = \frac{1}{2}\left(\Pi_{j+1}+\Pi_j\right) - \frac{h_m\lambda_c}{2\mathcal{F}}\left(v_{j+1}-v_j\right). \end{cases}$$

Finally, for the convection part, we **use an upwind scheme at the velocity** $v^*_{j+\frac{1}{2}}$ **for the quantities** h, hu, $h\Theta$ **and we use** $\Pi^*_{j+\frac{1}{2}}$ **for the pressure term**.

To ensure the well-balanced property, we consider the following discretization for the source terms [8]:

$$S_j = -g\frac{1}{2}\left(S_{j+\frac{1}{2}} + S_{j-\frac{1}{2}}\right). \tag{8}$$

where $S_{j+\frac{1}{2}}$ are defined in (5).

Proposition 2 *Considering that* $u_j = 0$, $v_j = 0$ *and* $\Pi_j = \frac{1}{2}g\Theta_j h_j^2$, *the schemes for the convective part (C) and for the wave part (W) are well-balanced for the three type of steady states:*

$$\begin{cases} \Theta_j = cst, \\ h_j + z_j = cst, \end{cases} \quad \begin{cases} z_j = cst, \\ \Theta_j(h_j)^2/2 = cst, \end{cases} \quad \begin{cases} h_j = cst, \\ z_j + h_j \ln(\Theta_j) = cst, \end{cases} \tag{9}$$

if we choose in the scheme

$$h_{j+\frac{1}{2}} = \frac{1}{2}(h_j + h_{j+1}), \quad \Theta_{j+\frac{1}{2}} = \begin{cases} \frac{\Theta_{j+1}-\Theta_j}{\ln(\Theta_{j+1})-\ln(\Theta_j)}, & \text{if } \Theta_{j+1} \neq \Theta_j, \\ \Theta_j, & \text{if } \Theta_{j+1} = \Theta_j. \end{cases}$$

Indeed, with this choice, then $v^*_{j+\frac{1}{2}} = 0$ for all the steady-states. Consequently, all the transport terms at velocity $v^*_{j+\frac{1}{2}}$ in (C) vanish. It remains to prove that the balance between the pressure term $(\Pi^*_{j+\frac{1}{2}} - \Pi^*_{j-\frac{1}{2}})/\Delta x = (\Pi^n_{j+1} - \Pi^n_{j-1})/(2\Delta x)$ and the source term. This follows from the choice (8). We do not detail the computation here. For the implicit scheme for (W), the preservation of these steady states results from this balance between discret gradient pressure and the source term and also from the centered discretization.

5 Numerical Results

First, we consider a classical test case for well-balanced schemes. The initial data is taken as a steady state and we compare the classical Rusanov scheme with the semi-implicit two-speed well-balanced scheme (SI two-speed WB) on a large time interval $[0, T_f]$. The CFL condition of our scheme is given by

$$\Delta t \leq \frac{\Delta x}{\max_x |u(t,x) + \mathcal{F}(t)\sqrt{h(t,x)\Theta(t,x)g}|}.$$

$$\Delta t \leq \frac{\Delta x}{\max_j \mid u_j + \mathcal{F}(t)\sqrt{h_j\Theta_j g} \mid}$$

The parameter $\mathcal{F}(t)$ depends on an important parameter $\mathcal{F}_{\min}$. If $u = 0$, there is still a small part of the gravity wave dynamics which is explicit and generates a CFL condition: $\Delta t \leq \Delta x/(\mathcal{F}_{\min}\sqrt{h\Theta g})$. Hence, decreasing this parameter $\mathcal{F}_{\min}$ increases the time step. We remark also that when $\mathcal{F}_{\min} = 1$, the scheme is explicit. The initial data are given by $u_0(x) = v_0(x) = 0$, $\Pi_0(x) = \frac{1}{2}\Theta_0(x)h_0(x)^2$ with the following topography $z(x)$, initial water height $h_0(x)$ and temperature $\Theta_0(x)$:

$$\begin{array}{llll} (ST1) & z(x) = 0.1 + G_{x_0,\sigma}(x), & h_0(x) = 8.0 - z(x), & \Theta_0(x) = 1, \\ (ST2) & z(x) = 1, & h_0(x) = 1.0 + 0.2G_{x_0,\sigma}(x), & \Theta_0(x) = \frac{1}{gh_0(x)^2}, \\ (ST3) & z(x) = x(1-x), & h_0(x) = 1, & \Theta_0(x) = 2e^{-x(1-x)}. \end{array}$$

with $G(x,\sigma) = \frac{1}{\sqrt{2\pi\sigma}}\exp(-\frac{(x-x_0)^2}{\sigma})$ and $\sigma = 0.06$. We consider $g = 1$, $N_c = 200$ cells and $T_f = 20$. The results are given in the table below.

The results show that the SI two-speed WB is exactly well-balanced in the explicit version ($\mathcal{F}_{\min} = 1$). In the semi-implicit case, it is more complicated. At the theoretical level the scheme is well-balanced, but the implicit part generates very small errors close to machine precision and these errors are slowly propagated. The scheme does not preserve the steady states exactly. However, we remark that with $\mathcal{F}_{\min} = 0.005$, the errors are very small (between $1.0E^{-11}$ and $1.0E^{-13}$) with a time step 200 larger than the one of the explicit schemes. In the following, we will show that the scheme well capture the flow around steady states and that these perturbative errors does not deteriorate the results contrary to classical non well-balanced schemes.

Δt/Error	Tests	Explicit Rusanov	SI two-speed WB ($\mathcal{F}_{\min} = 1$)	SI two-speed WB ($\mathcal{F}_{\min} = 0.1$)	SI two-speed WB ($\mathcal{F}_{\min} = 0.005$)
ST1	Error h	$1.5E^{-2}$	$1.5E^{-17}$	$1.5E^{-13}$	$3.6E^{-13}$
	Error u	$5.9E^{-3}$	$1.5E^{-15}$	$4.8E^{-11}$	$6.7E^{-13}$
	Error Θ	0.0	0.0	0.0	0.0
	Δt	$8.1E^{-4}$	$7.1E^{-4}$	$7.1E^{-3}$	$1.42E^{-1}$
ST2	Error h	$9.3E^{-2}$	0.0	$6.4E^{-11}$	$8.4E^{-12}$
	Error u	$7.3E^{-9}$	0.0	$8.7E^{-13}$	$1.3E^{-13}$
	Error Θ	0.13	$1.8E^{-17}$	$8.2E^{-11}$	$6.0E^{-12}$
	Δt	$2.5E^{-3}$	$2.3E^{-3}$	$2.3E^{-2}$	$4.7E^{-1}$
ST3	Error h	0.59	0.0	$7.1E^{-9}$	$1.38E^{-12}$
	Error u	0.65	$1.6E^{-15}$	$1.0E^{-9}$	$4.4E^{-14}$
	Error Θ	0.19	0.0	$9.4E^{-9}$	$1.4E^{-12}$
	Δt	$2.4E^{-3}$	$1.8E^{-3}$	$1.8E^{-2}$	0.49

Then we introduce a perturbation in the (ST3) test-case, $h_0(x) = 1 + 0.001\, G_{0.8,\sigma}(x)$, on the domain is $[0, 3]$. In Fig. 1, we compare the explicit Rusanov scheme with the SI two-speed WB scheme. We observe that the non-WB explicit Rusanov scheme is not able to capture the physical perturbation: the scheme creates a numerical perturbation larger that the physical one (in red), for 1200 cells, or with the same size (in blue) with 12000 cells. The SI two-speed WB scheme does not create numerical perturbations and capture correctly the physical perturbation with a coarser grid. The results are better with 600 cells and the SI two-speed WB scheme that with

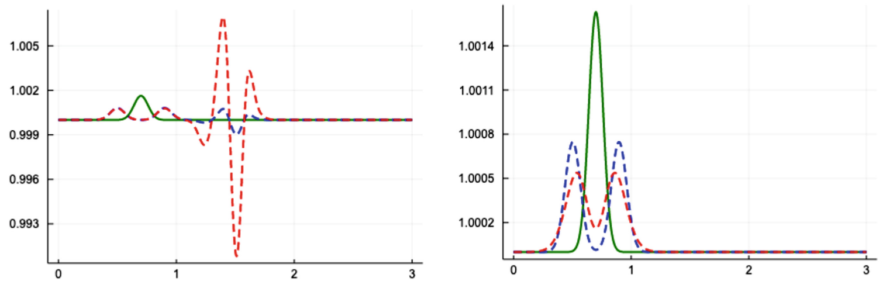


Fig. 1 Left: explicit Rusanov scheme; In green the initial data. In red the solution on a semi-coarse grid (1200 cells), in blue the solution on a fine grid (12,000 cells). Right: SI two-speed WB; in green the initial data. In red the solution on a coarse grid (600 cells), in blue the solution on a semi-coarse grid (4800 cells)

12,000 cells with the non-WB Rusanov scheme. Additionally, the results obtained with the SI two-speed WB scheme are given with a 10 times larger time step than the explicit one. We cannot increase too much the time step here since the implicit part of the schemes would create numerical diffusion on the gravity wave part.

6 Conclusion

In this paper, we propose a semi-implicit relaxation scheme for the Ripa model, that is well-adapted to treat the low-Froude regime. Indeed, we are able to take very large time steps compared to the gravity waves time scale and the accuracy of the scheme is independent of the Froude number. The two-speed relaxation allows to have an implicit step with a constant linear Laplacian. Additionally the scheme is able to preserve non-trivial steady states with a very good accuracy, with small errors for very large time steps. However the scheme is not able to treat wet/dry transitions. The 2D extension has been performed for the Euler system in [2]. In the future, we propose to extend the method to flows around equilibria MHD system.

References

1. Coulette, D., Franck, E., Helluy, P., Ratnani, A., Sonnendrueker, E.: Implicit time schemes for compressible fluid models based on relaxation methods. Comput. Fluids **188**(30), 70–85 (2019)
2. Bouchut, F., Franck, E., Navoret, L.: A low cost semi-implicit low-Mach relaxation scheme for the full Euler equations, preprint 2019
3. Berthon, C., Desveaux, V., Klingenberg, C.: Well-balanced scheme to capture non explicit steady-states: Ripa model. Math. Comput. **85**(300) (2016)
4. Iampietro, D., Daube, F., Galon, P., Herard, J.M.: A Mach-sensitive implicit-explicit scheme adapted to compressible multi-scale flows. J. Comput. Appl. Math. **340**(1), 122–150 (2018)

5. Tiam Kapen, P., Ghislain, T.: A new flux splitting scheme based on Toro-Vazquez and HLL schemes for the Euler equations. J. Comput. Methods Phys. (2018)
6. Toro, E.F., Vaquez-Cendon, M.E.: Flux splitting schemes for the Euler equations. Comput. Fluids **70** (2012)
7. Jin, S., Levermore, D.: Numerical schemes for hyperbolic conservation laws with stiff relaxation terms. J. Comput. Phys. **126**, 449–467 (1996)
8. Franck, E., Mendoza, L.: Finite volume scheme with local high order discretization of the hydrostatic equilibrium for the Euler equations with external forces. J. Sci. Comput. **69**(1) (2016)

Kinetic Over-Relaxation Method for the Convection Equation with Fourier Solver

Romane Hélie, Philippe Helluy, Emmanuel Franck, and Laurent Navoret

Abstract In this paper, we apply the CFL-less kinetic over-relaxation scheme presented in Coulette et al. (Comput Fluids 190:485–502 [1]) to the convection equation in two space dimensions. The method is a succession of free-transport steps and collisions steps. The free transport steps are solved with Fourier discretization. The collision steps are solved with over-relaxation for achieving high order. The method reaches six-order accuracy when using palindromic composition method. We apply the method to the guiding-center model in plasma physics.

Keywords Relaxation · Composition · Fourier · Guiding center model

MSC (2010) 35L65 · 65M12 · 65T50 · 35Q83 · 82D10

1 Introduction

The kinetic over-relaxation method [1] is a time semi-discrete method based on the approximation of a non-linear convection equation by a set of linear transport equations with constant velocities. Very efficient, CFL-less, and accurate transport solvers like Fourier methods can be used. Moreover, the over-relaxation technic lead to second-order accuracy in time. Even higher order can be achieved by composition methods. In this paper, we apply these methods to the convection equation in two-dimension and we show that it is particularly appropriate to solve the guiding center

R. Hélie · P. Helluy · E. Franck · L. Navoret (✉)
INRIA Grand-Est and IRMA Strasbourg, 7 Rue René Descartes, 67000 Strasbourg, France
e-mail: laurent.navoret@math.unistra.fr

R. Hélie
e-mail: romane.helie@math.unistra.fr

P. Helluy
e-mail: philippe.helluy@unistra.fr

E. Franck
e-mail: emmanuel.franck@inria.fr

R. Klöfkorn et al. (eds.), *Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples*, Springer Proceedings in Mathematics & Statistics 323,
https://doi.org/10.1007/978-3-030-43651-3_71

model, where the convection velocity field is given by a solution to a Poisson equation. The guiding center model is a simplified model to describe the two-dimensional dynamics of the charge density in a Tokamak. The particles are confined in the toroidal room thanks to a large external magnetic field B. Among several dynamics, this magnetic field leads to the so-called $E \times B$ drift of the particles, where E is the self-induced electric field. This model is also equivalent to the 2d incompressible Euler equation in the vorticity formulation. The dynamics result in very fine scale structures and thus require very accurate solvers.

2 Kinetic Over-Relaxation Approximation of the Convection Equation

We consider the following convection equation:

$$\partial_t \rho(t, \mathbf{x}) + \nabla \cdot \big(\rho(t, \mathbf{x})\, \mathbf{a}(t, \mathbf{x})\big) = 0, \tag{1}$$

where $\mathbf{a}(t, \mathbf{x}) \in \mathbb{R}^d$ is the velocity field and $\rho(t, \mathbf{x}) \in \mathbb{R}$ is the convected density.

To solve this convection equation with non-constant velocity field, the relaxation method consists in approximating it with several transport equations at constant velocities. More precisely, we introduce a kinetic vector $\mathbf{f}(t, \mathbf{x}) = (f_1(t, \mathbf{x}), f_2(t, \mathbf{x}), \dots, f_N(t, \mathbf{x})) \in \mathbb{R}^N$, whose components are associated to different velocities $(\boldsymbol{\lambda}_1, \dots, \boldsymbol{\lambda}_N) \in (\mathbb{R}^d)^N$. To a given kinetic vector $\mathbf{f}(t, \mathbf{x})$, we associate a macroscopic density

$$\rho_{\mathbf{f}}(t, \mathbf{x}) = \sum_{i=1}^{N} f_i(t, \mathbf{x}).$$

The numerical scheme is devised such that $\rho_{\mathbf{f}}$ is an approximation of the solution ρ. To this end, for any given density $\rho \in \mathbb{R}$, we introduce the so-called equilibrium kinetic vector $\mathbf{f}^{\mathrm{eq}}_{[\mathbf{a},\rho]}$ that satisfies the following consistency relations:

$$\rho = \sum_{i=1}^{N} f^{\mathrm{eq}}_{[\mathbf{a},\rho],i}, \quad \rho\mathbf{a} = \sum_{i=1}^{N} \boldsymbol{\lambda}_i f^{\mathrm{eq}}_{[\mathbf{a},\rho],i}. \tag{2}$$

The scheme is based on a time discretization of the following equation:

$$\partial_t \mathbf{f} + \sum_{k=1}^{d} \boldsymbol{\Lambda}_k \partial_{x_k} \mathbf{f} = \frac{1}{\varepsilon} \left(\mathbf{f}^{\mathrm{eq}}_{[\mathbf{a},\rho_{\mathbf{f}}]} - \mathbf{f}\right),$$

where $\boldsymbol{\Lambda}_k = \mathrm{diag}((\boldsymbol{\lambda}_1)_k, \dots, (\boldsymbol{\lambda}_2)_k)$ are $N \times N$ diagonal matrices, for $k = 1, \dots, d$, and where $\varepsilon > 0$ is a small parameter that controls the distance to the equilibria set. In the time-discretization, the time-dependent relaxation operator in the r.h.s. is replaced by a projection onto the equilibria set or a symmetry with respect to the equilibria set or a combination of the two.

The time semi-discretization of the over-relaxation scheme writes as follows. We start from the equilibrium distribution associated with the initial data: $\mathbf{f}(0,\mathbf{x}) = \mathbf{f}^{\text{eq}}_{[\mathbf{a}(0,\mathbf{x}),\rho_0(\mathbf{x})]}$. Then, at each time step $\Delta t > 0$, starting from $\mathbf{f}(t,\mathbf{x})$, we compute $\mathbf{f}(t + \Delta t, \mathbf{x})$ in two steps:

1. (transport step) advect the several kinetic components f_i with their respective velocities $\boldsymbol{\lambda}_i \in \mathbb{R}^d$
$$f_i^*(t + \Delta t, \mathbf{x}) = f_i(t, \mathbf{x} - \Delta t \boldsymbol{\lambda}_i), \quad \forall i \in \{1, \dots, N\},$$
which is also denoted in compact form: $\mathbf{f}^*(t + \Delta t, .) = T(\Delta t)\mathbf{f}(t, .)$.
2. (over-relaxation step) compute $\rho_{\mathbf{f}^*(t+\Delta t,.)}$ and then perform the following relaxation
$$\mathbf{f}(t + \Delta t, .) = \mathbf{f}^*(t + \Delta t, \mathbf{x}) + \omega \left(\mathbf{f}^{\text{eq}}_{[\mathbf{a}(t+\Delta t,.),\rho_{\mathbf{f}^*}(t+\Delta t,.)]} - \mathbf{f}^*(t + \Delta t, \mathbf{x})\right),$$
with $\omega \in [1, 2]$ a given parameter, also denoted: $\mathbf{f}(t + \Delta t, .) = R_\omega \mathbf{f}^*(t + \Delta t, .)$. For $\omega = 1$, we obtain the projection onto the equilibria set and for $\omega = 2$, we get the symmetry w.r.t the equilibria set.

The combination of these two steps writes as follows:

$$\mathbf{f}(t + \Delta t, .) = M_1(\Delta t)\mathbf{f}(t, .), \quad \text{with } M_1(\Delta t) = \Big(R_\omega \circ T(\Delta t)\Big),$$

Then $\rho_{\mathbf{f}}$ is a first-order approximation of the solution ρ to (1) for $\omega < 2$ and a second-order approximation if $\omega = 2$. We refer the reader to [1, 2] as regards the corresponding equivalent equation. From this equivalent equation, we can infer the so-called sub-characteristic condition that ensures the dissipativity of the second-order term in the expansion.

As presented in [1], higher-order time discretization can be devised by considering the following second-order time-symmetric operator:

$$M_2(\Delta t) = \left(T\left(\frac{\Delta t}{4}\right) \circ R_2 \circ T\left(\frac{\Delta t}{2}\right) \circ R_2 \circ T\left(\frac{\Delta t}{4}\right)\right),$$

and then using a palindromic composition method

$$M_p(\Delta t) = M_2(s_0 \Delta t) \circ M_2(s_1 \Delta t) \circ \cdots \circ M_2(s_p \Delta t),$$

where $s_i = s_{p-i}$, for $i = 0, \dots, p$. We will consider the fourth-order Suzuki scheme ($p = 4$) and the sixth order Kahan-Li scheme ($p = 8$). We refer to [1] for the expression of the corresponding parameters.

This numerical scheme has the advantage to concentrate all the non-linear operators in a local step, while the transport step becomes fully linear. Therefore, CFL-less method can be employed to make these transport steps. A semi-Lagrangian scheme has been used in [2]. On non-Cartesian meshes, implicit Discontinuous Galerkin method with upwind fluxes can be used as proposed in [1]. Here, we consider a Fourier discretization of the transport equation, ensuring a spectral accuracy.

In the sequel, we will use the so-called [D2Q4] kinetic approximation ($N = 4$). It consists in introducing the four velocities directed along the Cartesian axes:

$$\boldsymbol{\lambda}_1 = \begin{bmatrix} \lambda \\ 0 \end{bmatrix}, \quad \boldsymbol{\lambda}_2 = \begin{bmatrix} 0 \\ \lambda \end{bmatrix}, \quad \boldsymbol{\lambda}_3 = \begin{bmatrix} -\lambda \\ 0 \end{bmatrix}, \quad \boldsymbol{\lambda}_4 = \begin{bmatrix} 0 \\ -\lambda \end{bmatrix},$$

with $\lambda > 0$ and then we define the kinetic equilibrium vector:

$$f^{\text{eq}}_{[\mathbf{a},\rho],i} = \frac{\rho}{4} + \frac{\rho(\mathbf{a} \cdot \boldsymbol{\lambda}_i)}{2\lambda^2}, \quad \forall i \in \{1, 2, 3, 4\}.$$

This is the only solution to consistency relations (2), which satisfies symmetries. The sub-characteristic condition writes in that case: $\lambda > \max_{[0,T]\times\Omega} ||\mathbf{a}(t, \mathbf{x})||$, where $[0, T] \times \Omega$ is the computational domain.

We will also consider the [D2Q5] kinetic approximation ($N = 5$), where a fifth central null velocity is added:

$$\boldsymbol{\lambda}_1 = \begin{bmatrix} \lambda \\ 0 \end{bmatrix}, \quad \boldsymbol{\lambda}_2 = \begin{bmatrix} 0 \\ \lambda \end{bmatrix}, \quad \boldsymbol{\lambda}_3 = \begin{bmatrix} -\lambda \\ 0 \end{bmatrix}, \quad \boldsymbol{\lambda}_4 = \begin{bmatrix} 0 \\ -\lambda \end{bmatrix}, \quad \boldsymbol{\lambda}_5 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

where $\lambda > 0$. The kinetic equilibrium vector has to satisfy consistency relations (2):

$$\rho = f^{\text{eq}}_{[\mathbf{a},\rho],1} + f^{\text{eq}}_{[\mathbf{a},\rho],2} + f^{\text{eq}}_{[\mathbf{a},\rho],3} + f^{\text{eq}}_{[\mathbf{a},\rho],4} + f^{\text{eq}}_{[\mathbf{a},\rho],5},$$
$$\rho a_1 = \lambda(f^{\text{eq}}_{[\mathbf{a},\rho],1} - f^{eq}_{[\mathbf{a},\rho],3}), \quad \rho a_2 = \lambda(f^{\text{eq}}_{[\mathbf{a},\rho],2} - f^{eq}_{[\mathbf{a},\rho],4}).$$

This system is underdetermined. As already proposed in [2] for the one-dimensional case, we consider the following decomposition based on a flux-splitting

$$f^{\text{eq}}_{[\mathbf{a},\rho],i} = \rho\,(\boldsymbol{\lambda}_i \cdot \mathbf{a})_+\,, \quad \forall i \in \{1, 2, 3, 4\}, \quad f^{\text{eq}}_{[\mathbf{a},\rho],5} = \rho - \sum_{i=1}^{4} f^{\text{eq}}_{[\mathbf{a},\rho],i}\,,$$

where for any $v \in \mathbb{R}$, $v_+ = \max\{v, 0\}$ stands for the positive part of v or can be approximated by a smooth version $v_+ = (v + H_r(v))/2$ where $H_r(v)$ are Halley's functions defined recursively by: $H_0(x) = 1$, $H_{r+1}(x) = H_r(x)(H_r(x)^2 + 3x^2)/(3H_r(x)^2 + x^2)$. The sub-characteristic condition is the same as for the [D2Q4] approximation. As explained in [2], this scheme is expected to be more precise and better captures unidirectional flows.

3 Numerical Results

In this section, we validate the numerical scheme on two test-cases: the rotation advection test-case and the Kelvin-Helmholtz test-case for the guiding-center model. In these two test-cases, the transport part $T(\Delta t)$ is discretized with a Fourier method.

3.1 Rotation Test-Case

We consider the convection equation (1) where the velocity field is given by $\mathbf{a}(\mathbf{x}) = \mathbf{x}^{\perp}$. This velocity field is divergence free: $\nabla \cdot \mathbf{a} = 0$. Therefore, the convection equation (1) is equivalent to the advection equation:

$$\partial_t \rho(t, \mathbf{x}) + \mathbf{a}(\mathbf{x}) \cdot \nabla \rho(t, \mathbf{x}) = 0,$$

and the exact solution is just the rotation of the initial density around the origin.

In the following, we consider the domain $\Omega = [-1, 1] \times [-1, 1]$ and the exact solution:

$$\rho(t, \mathbf{x}) = \frac{1}{2\pi} \exp\left(-\frac{\|R(t)(\mathbf{x} - \mathbf{x}_0)\|^2}{\sigma^2}\right).$$

where $\sigma = 0.1$ and $\mathbf{x}_0 = (0.5, 0)$ and $R(t)$ is the rotation matrix of angle t. We use $N_x = N_y = 200$ discretization points in each direction.

Figure 1 (left) shows that the [D2Q4] scheme is first order accurate with the relaxation parameter $\omega = 1.95$. Second order accuracy is achieved with $\omega = 2$ for both [D2Q4] and [D2Q5] using the M_1 operator. As expected, we also note that the [D2Q5] is more accurate than the [D2Q4] scheme.

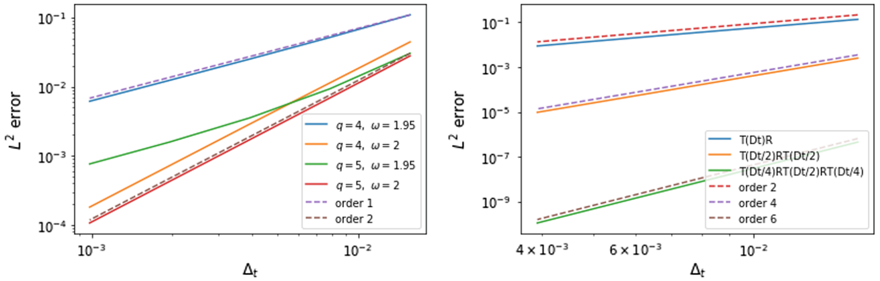


Fig. 1 L^2 error between the exact and the numerical solution obtained as function of the time step. Left: Comparison between $q = 4$ ([D2Q4]) and $q = 5$ ([$D2Q5$], $r = 4$) for different ω. Right: Comparison between different splitting operators when using the [$D2Q5$] method ($r = 4$) with a Kahan-Li palindromic composition and $\omega = 2$. Parameters: $\lambda = 2.1$, $N_x = N_y = 200$

In Fig. 1 (right), we observe that the M_2 operator is required to obtain the sixth-order accuracy of the Kahan-Li composition method. Using the M_1 operator leads to a second order operator and the Strang splitting $M_2^S(\Delta t) = \left(T\left(\frac{\Delta t}{2}\right) \circ R_2 \circ T\left(\frac{\Delta t}{2}\right)\right)$ to a fourth-order accuracy only.

As regards the computational time, we observe in Table 1 that considering M_1 is 1.26 times more efficient than considering M_2 when $\omega = 2$. Indeed, both methods are of order 2 and M_1 requires less transport steps. However, using M_2, we can use the Suzuki or the Kahan-Li composition methods that are respectively 22 and 34 times faster. For these comparisons, we use the $[D2Q5]$ method. The $[D2Q4]$ method seems just as fast even though it requires more transport steps. Although more accurate, the [D2Q5] is slowed down by the evaluation of the Halley functions.

3.2 Kelvin-Helmholtz Test-Case

We consider the guiding center model that describes the two-dimensional dynamics of electrons resulting from the $E \times B$ drift due to a large magnetic field. Their charge density is denoted $\rho(t, x) > 0$ and the guiding center model writes:

$$\partial_t \rho + E^{\perp} \cdot \nabla \rho = 0, \tag{3}$$

$$-\Delta \phi = \rho_0 - \rho, \qquad E = -\nabla \phi. \tag{4}$$

where $E(t, x) \in \mathbb{R}^d$ the electric field and $\phi(t, x) \in \mathbb{R}$ the electric potential. $\rho_0(t) > 0$ denotes the ion background charge density, which is supposed homogeneous. Actually, this model is equivalent to the 2d incompressible Euler equation in the vorticity formulation. Here we consider a square domain Ω with periodic boundary conditions and we thus assume that $\rho_0(t)$ equals the average of the density over the domain: $\rho_0(t) = \frac{1}{|\Omega|} \int_\Omega \rho(t, x) dx$.

Since $E = -\nabla \phi$, the advection vector field $E^{\perp}$ is divergence free. The transport equation is thus equivalent to the conservative convection equation

$$\partial_t \rho + \nabla \cdot \left(\rho E^{\perp}\right) = 0, \tag{5}$$

Unlike the previous advection equations presented so far, here the advection field depends on the density itself. Therefore, the over-relaxation scheme is slightly modified and writes:

1. (transport step) $f^*(t + \Delta t, .) = T(\Delta t)\mathbf{f}(t, .)$,
2. (Poisson step) compute $\rho_{\mathbf{f}^*(t+\Delta t}, .)$ and then find $\phi^*(t + \Delta t, .)$ by solving the Poisson equation and then $\mathbf{a}(t + \Delta t, .) = E^*(t + \Delta t, .)^{\perp}$.
3. (over-relaxation step) $f(t + \Delta t, .) = R_\omega f^*(t + \Delta t, .)$.

Note that both the transport step and the Poisson equation can be solved using a Fourier discretization in the square domain.

Table 1 Number of time steps and execution times needed to achieve an accuracy of 10^{-8} at time $T = \pi/2$ with $\lambda = 2.1$, $N_x = N_y = 200$

	Nb of time steps	Nb of transport steps	Error L^2	Execution time
$[D2Q5]$, $\omega = 2$, M_1	172,000	172,000	9.258×10^{-9}	4985.282
$[D2Q5]$, $\omega = 2$, M_2	82,000	246,000	9.975×10^{-9}	6298.363
$[D2Q5]$, $\omega = 2$, M_2, Suzuki	570	8550	9.516×10^{-9}	223.407
$[D2Q5]$, $\omega = 2$, M_2, Kahan-Li	190	5130	9.924×10^{-9}	145.426
$[D2Q4]$, $\omega = 2$, M_2, Kahan-Li	215	5805	9.627×10^{-9}	132.539

As already considered in [3–5], the Kelvin-Helmholtz instability test-case consists in considering the following initial condition:

$$\rho_{init}(x, y) = \sin x + \varepsilon\, \cos(ky),$$

in the domain $[0, 2\pi] \times [0, 2\pi/k]$, with periodic boundary conditions, and where $k \in \mathbb{R}$ is the perturbation wave number and $\varepsilon > 0$ is the perturbation amplitude. This is a perturbation of the stationary solution $\rho_0(x) = \sin x$ and $\phi_0(x) = -\sin x$. According to [4], there exists a critical wave number $k_s = 1$ such that an instability develops only for $k < k_s$. The instability rates are not known explicitly. However, we can compute them numerically.

We look for solutions of the form

$$\rho(x, y, t) = \rho_0(x) + \varepsilon\rho_1(x, y, t), \quad \phi_0(x, t) = \phi_0(x) + \varepsilon\, \phi_1(x, y, t)$$

where $\rho_1(x, y, t) = \tilde{\rho}_1(x) \exp(iky) \exp(-i\omega t)$, $\phi_1(x, y, t) = \tilde{\phi}_1(x) \exp(iky)$ $\exp(-i\omega t)$. Following [4], it can be proved that $\tilde{\phi}_1$ solves the generalized eigenvalue problem:

$$\phi_0' \left(\partial_{x^2}\tilde{\phi}_1 - k^2\tilde{\phi}_1\right) + \tilde{\phi}_1\rho_0' = -\omega/k\, \phi_0' \left(\partial_{x^2}\tilde{\phi}_1 - k^2\tilde{\phi}_1\right), \tag{6}$$

in which ω/k stands for the eigenvalue. As explained in [4], it can be proved that unstable solutions, corresponding to ω/k with positive imaginary part, exist if and only if $k < k_s = 1$. For k near k_s, a first order approximation of the instability rate can be computed: $\omega/k = 2(k_s - k)i$. Alternatively, we can also compute the instability rate by solving (6) numerically using a finite difference method. Introducing a space step $\Delta x = 1/N$ with $N \in \mathbb{N}$ and the corresponding spatial discretization of the interval $[0, 1]$, $x_i = i\Delta x$, we consider the approximate solution $\Phi_1 \in \mathbb{C}^N$, such that $(\Phi_1)_i \approx \tilde{\phi}_1(x_i)$ and which solves the following problem

$$C \left(D + (1 - k^2)\mathrm{Id}\right) \Phi_1 = \omega/k \left(D - k^2\mathrm{Id}\right) \Phi_1, \tag{7}$$

where $C = \mathrm{diag}(\cos(x_1), \dots, \cos(x_N))$ is diagonal matrix and D is discrete Laplacian matrix with periodic boundary conditions. Therefore, assembling $A = C\, (D+ (1 - k^2)\mathrm{Id})$ and $B = \left(D - k^2\mathrm{Id}\right)$, we just have to compute numerically the eigenvalues of the matrix $B^{-1}A$ and then keep the one with the largest imaginary part.

In Fig. 2 (left) is plotted the time evolution of the k-th Fourier mode of the potential. The instability rate fits perfectly with the expected one obtained solving (7). In the middle and right are plotted the contour lines of the density with the first-order scheme M_1 and the Kahan-Li composition methods. This illustrates the need to use high order scheme to capture the small structures.

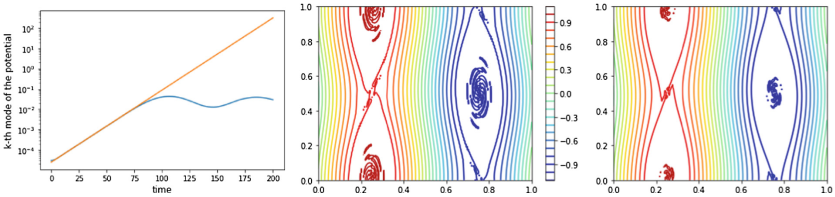


Fig. 2 (Kelvin-Helmholtz, $k = 0.95$, $\varepsilon = 10^{-4}$, $N_x = N_y = 200$, $\Delta t = 0.01$, [D2Q5], $r = 4$, $\lambda = 2.02$) Left: Time evolution of the k-th Fourier mode of the potential (in blue) and the straight line with slope $\mathrm{Im}(\omega) = 0.08185$ (in orange) with Kahan-Li, $\omega = 2$. Middle and left: Contour lines of the density at final time $T = 200$ with $\omega = 2$, Kahan-Li (middle) and $\omega = 1.95$, M_2 (right)

4 Conclusion

In this paper, we show that the kinetic over-relaxation method enables to devise numerical schemes for the convection equation based on Fourier discretization. The proposed method is optimally high-order accurate in space and can reach sixth order time accuracy with the Kahan-Li composition method. Unless high order schemes require more intermediate transport steps, the computational cost can be drastically decreased. Moreover, the method has been extended to the non-linear guiding center model. This is the first step before the extension to more complex advection equations like the gyro-kinetic equation in plasma physics.

Acknowledgements This work has been carried out within the framework of the EUROfusion Consortium and has received funding from the Euratom research and training programme 2014–2018 and 2019–2020 under grant agreement No 633053. The views and opinions expressed herein do not necessarily reflect those of the European Commission

References

1. Coulette, D., Franck, E., Helluy, P., Mehrenberger, M., Navoret, L.: High-order implicit palindromic discontinuous Galerkin method for kinetic-relaxation approximation. Comput. Fluids **190**, 485–502 (2019)
2. Courtès, C., Coulette, D., Franck, E., Navoret, L.: Vectorial kinetic relaxation model with central velocity. Application to implicit relaxations schemes. Commun. Comput. Phys. (2018)
3. Crouseilles, N., Mehrenberger, M., Sonnendrücker, E.: Conservative semi-Lagrangian schemes for Vlasov equations. J. Comput. Phys. **229**(6), 1927–1953 (2010)
4. Shoucri, M.M.: A two-level implicit scheme for the numerical solution of the linearized vorticity equation. Int. J. Numer. Meth. Eng. **17**(10), 1525–1538 (1981)
5. Sonnendrücker, E., Roche, J., Bertrand, P., Ghizzo, A.: The semi-Lagrangian method for the numerical resolution of the Vlasov equation. J. Comput. Phys. **149**(2), 201–220 (1999)

Cell-Centered Finite Volume Method for Regularized Mean Curvature Flow on Polyhedral Meshes

Jooyoung Hahn, Karol Mikula, Peter Frolkovič, Martin Balažovjech, and Branislav Basara

Abstract A cell-centered finite volume method is used to numerically solve a regularized mean curvature flow equation on polyhedral meshes. It is based on an over-relaxed correction method used previously for linear diffusion problems. An iterative nonlinear Crank-Nicolson method is proposed to obtain the second-order accuracy in time and space. The proposed algorithm is used for three-dimensional domains decomposed for parallel computing for two examples that numerically verify the second order accuracy on polyhedral meshes.

Keywords Regularized mean curvature flow · Polyhedral meshes · Over-relaxed correction method · Nonlinear Crank-Nicolson method

MSC (2010) 65M08 · 35G31 · 35K61

1 Introduction

A finite volume method to solve the level set formulation of regularized mean curvature flow [15] on a bounded Lipschitz continuous domain $\Omega \subset \mathbb{R}^3$ is presented:

$$\frac{\partial \phi}{\partial t} = |\nabla \phi|_\varepsilon \nabla \cdot \left(\frac{\nabla \phi}{|\nabla \phi|_\varepsilon} \right), \quad |\nabla \phi|_\varepsilon = (\varepsilon^2 + |\nabla \phi|^2)^{1/2}, \tag{1}$$

J. Hahn (✉) · B. Basara
AVL List GmbH, Hans-List-Platz 1, 8020 Graz, Austria
e-mail: jooyoung.hahn@avl.com

B. Basara
e-mail: branislav.basara@avl.com

K. Mikula · P. Frolkovič · M. Balažovjech
Department of Mathematics and Descriptive Geometry, Slovak University of Technology, Radlinskeho 11, 810 05 Bratislava, Slovakia
e-mail: karol.mikula@stuba.sk

P. Frolkovič
e-mail: peter.frolkovic@stuba.sk

R. Klöfkorn et al. (eds.), *Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples*, Springer Proceedings in Mathematics & Statistics 323,
https://doi.org/10.1007/978-3-030-43651-3_72

where the regularization parameter $\varepsilon > 0$ is used as a small constant [6]. The initial and Dirichlet boundary conditions are defined by

$$\begin{aligned} \phi(\mathbf{x}, 0) &= \phi_0(\mathbf{x}), \quad \mathbf{x} \in \Omega, \\ \phi(\mathbf{x}, t) &= \phi_b(\mathbf{x}, t), \quad (\mathbf{x}, t) \in \partial\Omega \times (0, T]. \end{aligned} \tag{2}$$

The level set form of mean curvature flow equation and its modifications are extensively used in numerical applications like the filtering or segmentation in image processing [11], the G-equation in combustion models of computational fluid dynamics [16], and the interface problems in material science; see more details in [8, 14, 17] and the references therein.

To solve (1) numerically and to develop related mathematical theories, several methods are used: the finite difference [13, 17], the finite element [3], and the finite volume methods [7, 11, 19]. In this paper, a method based on a cell-centered finite volume method is proposed in order to use the smallest number of unknowns on a polyhedron mesh. For a spatial discretization, one of practically used algorithms in computer-aided engineering to discretize an elliptic operator on polyhedron meshes, so-called over-relaxed correction method, is considered [4, 12], because a formal expansion of the right-hand side of (1) is a combination of a Laplacian and a nonlinear term of second order derivatives; see more details in [18]. For a temporal discretization, a nonlinear Crank-Nicolson method [1] is considered in order to have the second order accuracy with a time step size proportional to the space discretization step. In such a way, the proposed method can be conveniently combined with second order accurate methods for an advective or normal flow equation [9, 10], e.g., for the G-equation model [16]. We also use a deferred correction method [2] in order to achieve computational efficiency with a 1-ring face neighborhood structure on domains decomposed for parallel computing.

The paper is organized as follows. In Sect. 2.1, we derive the spatial discretization based on the over-relaxed correction method. In Sect. 2.2, the iterative nonlinear Crank-Nicolson method is proposed. In Sect. 3, the experimental order of convergence for two exact solutions on two computational domains is presented.

2 Cell-Centered Finite Volume Method

The computational domain $\Omega \subset \mathbb{R}^3$ is discretized by open non-overlapping polyhedral cells Ω_p and I is the set of cell indices. We indicate a set N_p as adjacent cell indices to Ω_p, where the cells Ω_q, $q \in N_p$ have a non-zero area intersection with Ω_p. An internal face, the result of such intersection, is denoted by $e_f \subset \partial\Omega_q \cap \partial\Omega_p$ and the set of all internal faces in a mesh is denoted by F. We similarly define a set B as the index set of all boundary faces $e_b \subset \partial\Omega_p \cap \partial\Omega$ for $p \in I$. The face indices of a cell Ω_p, $p \in I$ belong either to the set $F_p \subset F$ or to the set $B_p \subset B$. A numerical solution at time t is represented by unknowns $\phi_p \approx \phi(\mathbf{x}_p, t)$, where $\mathbf{x}_p$ is the center of cell Ω_p. The Dirichlet boundary condition in (2) is evaluated at the centers $\mathbf{x}_b$ of boundary faces e_b, i.e., $\phi_b = \phi_b(\mathbf{x}_b, t)$.

We integrate (1) on $\Omega_p, p \in I$,

$$\int_{\Omega_p} \frac{1}{|\nabla\phi|_\varepsilon} \frac{\partial\phi}{\partial t} = \int_{\Omega_p} \nabla \cdot (g\nabla\phi) = \sum_{f \in F_p \cup B_p} \int_{e_f} g\nabla\phi \cdot \mathbf{n}, \tag{3}$$

where $g = |\nabla\phi_p|_\varepsilon^{-1}$, and $\mathbf{n}$ is an outward normal vector. An approximation of gradient $\nabla\phi$ on a cell Ω_p, $\nabla\phi_p \approx \nabla\phi(\mathbf{x}_p, t)$, is computed by an inverse distance weighted least-squares minimization [5, 12]:

$$\nabla\phi_p \equiv L_p(\phi_p, \phi_b) = \underset{\mathbf{y}\in\mathbb{R}^3}{\operatorname{argmin}} \left(\sum_{q\in N_p\cup B_p} |\mathbf{d}_{pq}|^{-2} (\phi_p + \mathbf{y} \cdot \mathbf{d}_{pq} - \phi_q)^2 \right). \tag{4}$$

The notation $\mathbf{d}_{\alpha\beta} \equiv \mathbf{x}_\beta - \mathbf{x}_\alpha$ for directional vectors is used throughout the paper.

In Sect. 2.1, we present a spatial discretization of (3), including an approximation of normal flux $g\nabla\phi \cdot \mathbf{n}$. Afterwards, in Sect. 2.2, we discuss a temporal discretization.

2.1 Over-Relaxed Correction Method

In a derivation of spatial discretization, we follow mostly [4, 12]. We assume that all variables are continuous at the face centers $\mathbf{x}_f, f \in F$, and we denote their values by the subscript f. For an internal face e_f, the normal flux in (3) at time t is first approximated by

$$f \in F_p \Rightarrow \int_{e_f} g\nabla\phi \cdot \mathbf{n} \approx g_f \nabla\phi_f \cdot \mathbf{n}_{pf}, \tag{5}$$

where $\mathbf{n}_{pf}$ is the outward normal vector to the face such that $|\mathbf{n}_{pf}| = |e_f|$.

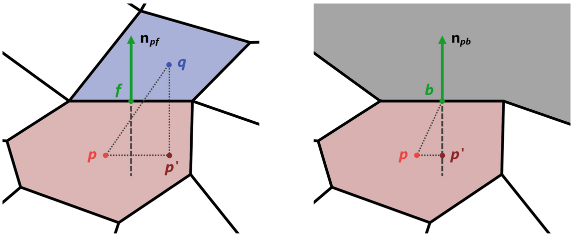
We use an orthogonal decomposition of the vector $\mathbf{d}_{pq}$ with respect to $\mathbf{n}_{pf}$ and $\mathbf{t}_f$, $\mathbf{n}_{pf} \perp \mathbf{t}_f$, written formally in the form

$$\mathbf{d}_{pq} = \frac{g_f}{c_f}\mathbf{n}_{pf} - \mathbf{t}_f, \tag{6}$$

where

$$c_f = g_f \frac{\mathbf{n}_{pf} \cdot \mathbf{n}_{pf}}{\mathbf{n}_{pf} \cdot \mathbf{d}_{pq}}, \quad \mathbf{t}_f = \left(\frac{\mathbf{n}_{pf}}{|\mathbf{n}_{pf}|} \cdot \mathbf{d}_{pq} \right) \frac{\mathbf{n}_{pf}}{|\mathbf{n}_{pf}|} - \mathbf{d}_{pq}.$$

Fig. 1 In **a**, the notation for an internal face $e_f, f \in F_p$ is shown, and $\mathbf{d}_{p'p} \perp \mathbf{n}_{pf}$. In **b**, the notation for a boundary face e_b, $b \in B_p$ is shown with the gray region being outside of the computational domain, and $\mathbf{d}_{p'p} \perp \mathbf{n}_{pb}$



(a) (b)

Note that $\mathbf{t}_f = \mathbf{d}_{p'p}$ in Fig. 1a. Rewriting (6) as $g_f \mathbf{n}_{pf} = c_f(\mathbf{d}_{pq} + \mathbf{t}_f)$, we can derive the approximation:

$$g_f \nabla\phi_f \cdot \mathbf{n}_{pf} \approx c_f \left(\phi_q - \phi_p + \nabla\phi_f \cdot \mathbf{t}_f\right). \tag{7}$$

The face gradient $\nabla\phi_f$ is approximated from gradients in the adjacent cells,

$$\nabla\phi_f = \omega_{qf}\nabla\phi_p + \omega_{pf}\nabla\phi_q, \quad \omega_{pf} + \omega_{qf} = 1, \quad \omega_{pf} = \frac{|\mathbf{d}_{pf}|}{|\mathbf{d}_{pf}| + |\mathbf{d}_{qf}|}.$$

Similarly, for a boundary face, the normal flux $g\nabla\phi \cdot \mathbf{n}$ in (3) is approximated by

$$b \in B_p \Rightarrow \int_{e_b} g\nabla\phi \cdot \mathbf{n} \approx g_p \nabla\phi_p \cdot \mathbf{n}_{pb}. \tag{8}$$

Using analogous orthogonal decomposition of $\mathbf{d}_{pb}$ in Fig. 1b, and $g_p\mathbf{n}_{pb} = c_b(\mathbf{d}_{pb} + \mathbf{t}_b)$, where $\mathbf{n}_{pb} \perp \mathbf{t}_b$, it gives us a discretization:

$$g_p \nabla\phi_p \cdot \mathbf{n}_{pb} \approx c_b(\phi_b - \phi_p + \nabla\phi_p \cdot \mathbf{t}_b), \tag{9}$$

where

$$c_b = g_p \frac{\mathbf{n}_{pb} \cdot \mathbf{n}_{pb}}{\mathbf{n}_{pb} \cdot \mathbf{d}_{pb}}, \quad \mathbf{t}_b = \left(\frac{\mathbf{n}_{pb}}{|\mathbf{n}_{pb}|} \cdot \mathbf{d}_{pb}\right)\frac{\mathbf{n}_{pb}}{|\mathbf{n}_{pb}|} - \mathbf{d}_{pb}.$$

Note that $\mathbf{t}_b = \mathbf{d}_{p'p}$ in Fig. 1b.

Substituting (5), (7), (8), and (9) in (3), and assuming a constant approximation of ϕ_p and $\nabla\phi_p$ on a cell Ω_p in the left hand side of (3), we have the final spatial discretization:

$$\frac{|\Omega_p|}{|\nabla\phi_p|_\varepsilon}\frac{d}{dt}\phi_p = \sum_{f \in F_p} c_f(\phi_q - \phi_p + \nabla\phi_f \cdot \mathbf{t}_f) + \sum_{b \in B_p} c_b(\phi_b - \phi_p + \nabla\phi_p \cdot \mathbf{t}_b). \tag{10}$$

2.2 Iterative Nonlinear Crank-Nicolson Method

Let us denote a time step as Δt, and $\phi_p^n \approx \phi(\mathbf{x}_p, n\Delta t), p \in I$, and $n \in \mathbb{N}$. The values given by the initial condition in (2) are denoted by $\phi^0 = (\phi_1^0, \ldots, \phi_{|I|}^0)^{\mathrm{T}}$. To compute ϕ^n, we use a nonlinear Crank-Nicolson method with a deferred correction method [2]. For $n \geq 1$ and $k \geq 1$, the method to solve (10) is presented:

$$\begin{aligned}\frac{|\Omega_p|}{\Delta t}\left(\phi_p^{n,k} - \phi_p^{n-1}\right) = {} & \frac{1}{2}\sum_{f\in F_p} \alpha_{pf}^{n,k-1}\left(\phi_q^{n,k} - \phi_p^{n,k} + \nabla\phi_f^{n,k-1}\cdot \mathbf{t}_f\right) \\ & + \frac{1}{2}\sum_{b\in B_p} \alpha_{pb}^{n,k-1}\left(\phi_b^{n} - \phi_p^{n,k} + \nabla\phi_p^{n,k-1}\cdot \mathbf{t}_b\right) \\ & + \frac{1}{2}\sum_{f\in F_p} \alpha_{pf}^{n-1}\left(\phi_q^{n-1} - \phi_p^{n-1} + \nabla\phi_f^{n-1}\cdot \mathbf{t}_f\right) \\ & + \frac{1}{2}\sum_{b\in B_p} \alpha_{pb}^{n-1}\left(\phi_b^{n-1} - \phi_p^{n-1} + \nabla\phi_p^{n-1}\cdot \mathbf{t}_b\right), \end{aligned} \tag{11}$$

where $\alpha_{pf}^{n,k-1} \equiv c_f^{n,k-1}|\nabla\phi_p^{n,k-1}|_\varepsilon$ and $\alpha_{pf}^{n-1} \equiv c_f^{n-1}|\nabla\phi_p^{n-1}|_\varepsilon$, for $f \in F_p \cup B_p$. Note that $\nabla\phi_p^{n,k-1} \equiv L_p(\phi_p^{n,k-1}, \phi_b^n)$ and $\nabla\phi_p^{n-1} \equiv L_p(\phi_p^{n-1}, \phi_b^{n-1})$. Moreover, for $k = 0$ the values are determined from the previous time step, e.g., $\alpha_{pf}^{n,0} = \alpha_{pf}^{n-1}$. For each n and k, one has to solve a system (11) of linear algebraic equations, where the elements of the matrix for the system can change with each n and k. Note that the original nonlinear Crank-Nicolson method [1] should use the terms $\nabla\phi_f^{n,k}\cdot\mathbf{t}_f$ and $\nabla\phi_b^{n,k}\cdot\mathbf{t}_b$ in (11) instead of $\nabla\phi_f^{n,k-1}\cdot\mathbf{t}_f$ and $\nabla\phi_b^{n,k-1}\cdot\mathbf{t}_b$, respectively. However, using the original form brings computational difficulties in general when practical industrial problems are solved because of a larger number of non-zero coefficients in the matrix and a larger communication cost for parallel computing. Therefore, the iterative deferred correction method is used in (11).

Rewriting (11) formally as a matrix equation $\mathbf{A}^{n,k-1}\phi^{n,k} = \mathbf{F}(\phi^{n,k-1})$, the k^{th} iteration is stopped at the smallest K_n such that a residual error is smaller than a chosen error bound η:

$$\frac{1}{|I|}\sum_{p\in I}\left|\left(\mathbf{A}^{n,K_n}\phi^{n,K_n} - \mathbf{F}(\phi^{n,K_n})\right)_p\right| < \eta. \tag{12}$$

Then, we define $\phi^n \equiv \phi^{n,K_n}$.

3 Numerical Experiments

Two exact solutions of the mean curvature flow equation are used in order to check the experimental order of convergence (*EOC*) of proposed algorithm (11). The numerical solutions are computed for two domains using polyhedral meshes generated by `AVL FIRE`™ in Fig. 2. For all examples in this paper, we use the threshold $\eta = 10^{-10}$ in (12) to stop the iteration. Moreover, we stop the iterations if $k > 100$ in (11). The *EOC* is computed by using an average discretization size,

$$h = \frac{1}{|I|}\sum_{p\in I} |\Omega_p|^{1/3}\,, \tag{13}$$

and four meshes for which h is decreasing. In Fig. 2, the polyhedral mesh in the cube domain Ω_1 is shown with $h = 1.90 \times 10^{-1}$ and we use the related finer meshes with the average discretization sizes $h = 9.52 \times 10^{-2}$, 4.76×10^{-2}, and 2.48×10^{-2}. The polyhedral mesh in the domain Ω_2 of more complex shape has $h = 6.64 \times 10^{-2}$ and the related finer meshes have $h = 4.17 \times 10^{-2}$, 2.27×10^{-2}, and 1.29×10^{-2}. Four types of the norms are used to compute the *EOC*. The errors E^2 and E^∞ are the $L^2((0,T) \times \Omega)$ and $L^\infty(0,T; L^2(\Omega))$ norms of the difference between the exact and the numerical solutions, respectively. The errors G^2 and G^∞ are the $L^2((0,T) \times \Omega)^3$ and $L^\infty(0,T; L^2(\Omega)^3)$ norms of the difference between the gradient of the exact and the numerical solutions, respectively.

The two exact solutions of (1) for $\varepsilon = 0$ on the domains in Fig. 2 are used:

$$\phi^i(\mathbf{x}, t) = \left(\frac{|\mathbf{x}|^2}{4} + t\right)^{i/2}\,, \tag{14}$$

where $(\mathbf{x}, t) \in \Omega_i \times [0, T]$, $i = 1, 2$, and $T = 0.16$. Note that the regularization parameter in (11) is chosen as $\varepsilon = h^2$. The functions ϕ_0 and ϕ_b in the initial and boundary conditions are obtained from the given exact solution.

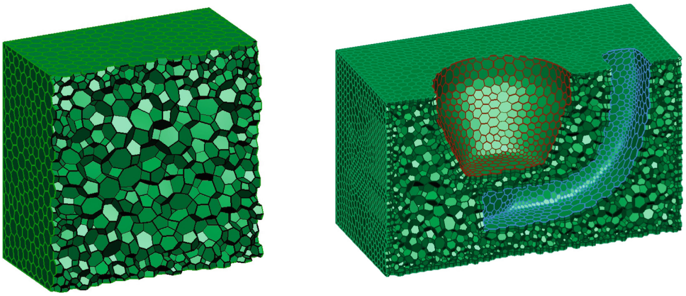


Fig. 2 A half cut view of polyhedral meshes in a cube domain $\Omega_1 = [-1.25, 1.25]^3 \subset \mathbb{R}^3$ (left) and in a domain Ω_2 of a complex shape (right)

Table 1 The *EOC* of numerical solution of (1) using the exact solution in (14) with $i = 1$ on Ω_1 (top) and Ω_2 (bottom) is presented by using the iterative nonlinear Crank-Nicolson method (11)

N	E^2	EOC	E^∞	EOC	G^2	EOC	G^∞	EOC
1	3.54×10^{-3}		9.41×10^{-3}		2.36×10^{-2}		6.60×10^{-2}	
2	7.39×10^{-4}	3.36	2.15×10^{-3}	3.17	9.58×10^{-3}	1.93	3.00×10^{-2}	1.69
3	2.08×10^{-4}	2.09	7.59×10^{-4}	1.72	4.46×10^{-3}	1.26	1.72×10^{-2}	0.92
4	4.70×10^{-5}	2.64	2.35×10^{-4}	2.08	1.86×10^{-3}	1.55	8.24×10^{-3}	1.30
N	E^2	EOC	E^∞	EOC	G^2	EOC	G^∞	EOC
1	7.03×10^{-4}		2.01×10^{-3}		1.03×10^{-2}		3.61×10^{-2}	
2	2.30×10^{-4}	2.40	8.58×10^{-4}	1.82	4.92×10^{-3}	1.59	2.21×10^{-2}	1.05
3	5.43×10^{-5}	2.38	3.49×10^{-4}	1.49	2.05×10^{-3}	1.45	1.22×10^{-2}	0.98
4	1.73×10^{-5}	2.03	1.44×10^{-4}	1.56	9.67×10^{-4}	1.33	7.19×10^{-3}	0.93

Table 2 The *EOC* of numerical solution of (14) with $i = 2$ on Ω_1 (top) and Ω_2 (bottom) is presented by using the iterative nonlinear Crank-Nicolson method (11)

N	E^2	EOC	E^∞	EOC	G^2	EOC	G^∞	EOC
1	5.02×10^{-3}		1.49×10^{-2}		4.79×10^{-2}		1.24×10^{-1}	
2	1.06×10^{-3}	3.33	2.81×10^{-3}	3.57	1.88×10^{-2}	2.01	4.99×10^{-2}	1.94
3	3.01×10^{-4}	2.08	8.94×10^{-4}	1.89	7.73×10^{-3}	1.46	2.26×10^{-2}	1.31
4	8.60×10^{-5}	2.22	2.74×10^{-4}	2.09	3.15×10^{-3}	1.59	9.85×10^{-3}	1.47
N	E^2	EOC	E^∞	EOC	G^2	EOC	G^∞	EOC
1	9.72×10^{-4}		2.58×10^{-3}		1.41×10^{-2}		3.92×10^{-2}	
2	2.91×10^{-4}	2.58	8.03×10^{-4}	2.51	6.68×10^{-3}	1.59	2.02×10^{-2}	1.42
3	6.08×10^{-5}	2.59	1.79×10^{-4}	2.48	2.46×10^{-3}	1.65	7.99×10^{-3}	1.53
4	2.10×10^{-5}	1.88	6.64×10^{-5}	1.76	1.15×10^{-3}	1.35	3.83×10^{-3}	1.30

In Tables 1 and 2, the EOC of numerical solutions of (14) with $i = 1$ and $i = 2$ are presented, respectively. We choose the time step $\Delta t = T/2^{N-1}$ for $N \in \{1, 2, 3, 4\}$, where $N = 1$ for the coarsest mesh and $N = 4$ for the finest mesh. The iterative nonlinear Crank-Nicolson method (11) shows $EOC \simeq 2$ in the error norms E^2 and E^∞ on Ω_1 and the EOC is larger than 1 in the error norms G^2 and G^∞. Note that in the case of domain Ω_2, the EOC is partially influenced by the nontrivial task of approximating the curved shape of $\partial\Omega_2$ with polyhedral meshes.

4 Conclusion

We present a cell-centered finite volume method for the regularized mean curvature flow equation, which is suitable on polyhedral meshes. The numerical experiments for the chosen examples indicate a convergence rate of around 2. Consequently, the proposed method can use the time step proportional to the average discretization size to obtain the second order accurate method in time and space.

Acknowledgements The work was supported by grants VEGA 1/0709/19 and 1/0436/20 and APVV-0522-15.

References

1. Balažovjech, M., Mikula, K.: A higher order scheme for a tangentially stabilized plane curve shortening flow with a driving force. SIAM J. Sci. Comp. **33**(5), 2277–2294 (2011)
2. Böhmer, K., Hemker, P.W., Stetter, H.J.: The defect correction approach. In: Defect Correction Methods, pp. 1–32. Springer (1984)
3. Deckelnick, K., Dziuk, G.: Error estimates for a semi-implicit fully discrete finite element scheme for the mean curvature flow of graphs. Interface Free Bound. **2**, 341–359 (2000)
4. Demirdžić, I.: On the discretization of the diffusion term in finite-volume continuum mechanics. Num. Heat Tr. B-Fund. **68**, 1–10 (2015)
5. Demirdžić, I., Muzaferija, S.: Numerical method for coupled fluid flow, heat transfer and stress analysis using unstructured moving meshes with cells of arbitrary topology. Comp. Meth. Appl. Mech. Eng. **125**, 235–255 (1995)
6. Evans, L.C., Spruck, J.: Motion of level sets by mean curvature. I. J. Differential Geom. **33**, 635–681 (1991)
7. Eymard, R., Handlovičová, A., Mikula, K.: Study of a finite volume scheme for the regularized mean curvature flow level set equation. IMA J. Numer. Anal. **31**, 813–846 (2011)
8. Gibou, F., Fedkiw, R., Osher, S.: A review of level-set methods and some recent applications. J. Comput. Phys. **353**, 82–109 (2018)
9. Hahn, J., Mikula, K., Frolkovič, P., Medl'a, M., Basara, B.: Iterative inflow-implicit outflow-explicit finite volume scheme for level-set equations on polyhedron meshes. Comp. Math. Appl. **77**, 1639–1654 (2019)
10. Hahn, J., Mikula, K., Frolkovič, P., Basara, B.: Inflow-based gradient finite volume method for a propagation in a normal direction in a polyhedron mesh. J. Sci. Comp. **72**, 442–465 (2017)
11. Handlovičová, A., Mikula, K., Sgallari, F.: Semi-implicit complementary volume scheme for solving level set like equations in image processing and curve evolution. Numeri. Math. **93**, 675–695 (2003)

12. Jasak, H.: Error analysis and estimation for the finite volume method with applications to fluid flows. Ph.D. thesis (1996)
13. Oberman, A.M.: A convergent monotone difference scheme for motion of level sets by mean curvature. Numer. Math. **99**, 365–379 (2004)
14. Osher, S., Fedkiw, R.: Level Set Methods and Dynamic Implicit Surfaces. Springer, Berlin (2000)
15. Osher, S., Sethian, J.A.: Fronts propagating with curvature dependent speed: algorithms based on Hamilton-Jacobi formulations. J. Comput. Phys. **79**, 12–49 (1988)
16. Peters, N.: Turbulent Combustion. Cambridge Monographs on Mechanics. Cambridge University Press (2000)
17. Sethian, J.A.: Level Set Methods and Fast Marching Methods, Evolving Interfaces in Computational Geometry, Fluid Mechanics, Computer Vision, and Materials Science. Cambridge University Press, New York (1999)
18. Smereka, P.: Semi-implicit level set methods for curvature and surface diffusion motion. J. Sci. Comput. **19**, 439–456 (2003)
19. Walkington, N.J.: Algorithms for computing motion by mean curvature. SIAM J. Numer. Anal. **33**, 2215–2238 (1996)

A Fully Eulerian Finite Volume Method for the Simulation of Fluid-Structure Interactions on AMR Enabled Quadtree Grids

Michel Bergmann, Antoine Fondanèche, and Angelo Iollo

Abstract We present a versatile fully Eulerian method for the simulation of fluid-structure interactions. The model equations are solved using a finite-volume scheme on a compact and possibly dynamic quadtree stencil. The structure geometry is followed using a level-set model and a distance function. A regularized Heaviside function that allows to discriminate between the fluid and the elastic phases is then defined with respect to the moving structure. The elastic deformation of the structure is described according to the backward characteristics which are in turn used to express the Cauchy stress tensor of a two-parameter Mooney-Rivlin material. The numerical model is validated with respect to the literature and an example of application is detailed.

Keywords Fluid-structure interaction · Finite volume · Eulerian model · Quadtree grids

MSC (2010) 74S10 · 74B20 · 74F10

1 Introduction

The simulation of Fluid-Structure Interactions (FSI) is of interest in a wide range of application fields, from engineering to medecine. For instance, the simulation of a flow around a wind turbine blade [1] or a blood flow in a thoracic aorta [2] is an

M. Bergmann · A. Fondanèche (✉) · A. Iollo
Equipe-projet Memphis, Inria Bordeaux-Sud Ouest, 33400 Talence, France
e-mail: antoine.fondaneche@inria.fr

M. Bergmann
e-mail: michel.bergmann@inria.fr

A. Iollo
e-mail: angelo.iollo@inria.fr

UMR 5251, Université de Bordeaux, IMB, 33400 Talence, France

R. Klöfkorn et al. (eds.), *Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples*, Springer Proceedings in Mathematics & Statistics 323,
https://doi.org/10.1007/978-3-030-43651-3_73

essential support for a quantitative understanding of complex phenomena. Among all the studies dealing with this problem, there are two categories of approaches to deal with the deforming structure.

The first approach is based on body-fitted or interface-tracking methods, such as Arbitrary Lagrangian-Eulerian (ALE) [3, 4] and Deforming Spatial-Domain/Space-Time (DSD/ST) [5, 6] methods. This kind of methods is interesting since efficient specialized techniques for solving both flow and structural sub-problems can be employed. However, the implementation of a fluid-structure coupling scheme requires a mesh adapted to the geometry and when the material has large deformations remeshing and partitioning is complex and computationally expensive. The second approach is based on fictitious domain methods, such as the immersed boundary (IB) methods introduced by Peskin [7, 8], or cut-cell methods [9]. This type of methods offers a good trade-off between accuracy and practicability of the simulation since they do not require remeshing.

Traditionally, fluid dynamics is represented through Eulerian approaches, while structural dynamics is modeled using Lagrangian methods. Here, we develop a fully Eulerian method for simulating the interaction between a viscous incompressible fluid and an hyperelastic Mooney-Rivlin material on quadtree grids. In the context of interface-capturing methods for the simulation of multiphase flows, the governing equations for the whole system are solved in a monolithic way, by using a single-continuum model for the whole domain. Using a regularized Heaviside function which depends on the level-set function, this diffuse-interface method guarantees the continuity of the solution at the interface. Rigid bodies are taken into account with the Brinkmann penalization method [10].

2 The Fully Eulerian FSI Model

A computational domain Ω is divided into three subdomains related to the different media such that $\Omega = \Omega_f \cup \Omega_s \cup \Omega_e$. We denote by f, e and s the subscripts refering to the fluid, rigid solid and elastic material respectively. As depicted in Fig. 1, the boundaries of the deformable and non-deformable bodies are called $\Gamma_s(t) = \partial\Omega_s(t)$ and $\Gamma_e(t) = \partial\Omega_e(t)$ respectively.

2.1 The Governing Equations

The interaction between an incompressible viscous fluid, a rigid non-deformable body and an hyperelastic structure is governed by the following system of PDEs:

$$\begin{cases} \rho_f\left(\frac{\partial \mathbf{u}_f}{\partial t} + (\mathbf{u}_f \cdot \nabla)\mathbf{u}_f\right) = -\nabla p + \nabla \cdot \boldsymbol{\sigma}_f(\mathbf{u}_f) & \text{in } \Omega_f(t), \\ \nabla \cdot \mathbf{u}_f = 0 & \text{in } \Omega_f(t), \end{cases} \tag{1a}$$

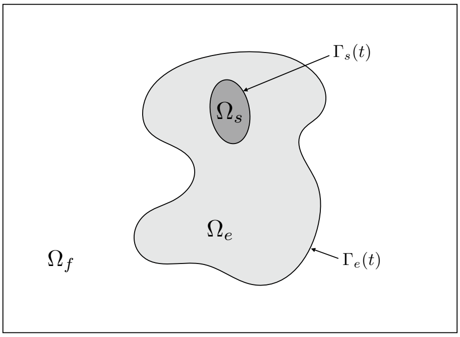


Fig. 1 Sketch of the FSI set-up

$$\rho_e\Big(\frac{\partial \mathbf{u}_e}{\partial t} + (\mathbf{u}_e \cdot \nabla)\mathbf{u}_e\Big) = -\nabla p + \nabla \cdot \boldsymbol{\sigma}_e \qquad \text{in } \Omega_e(t), \tag{1b}$$

$$\begin{cases} \mathbf{u}_f = \mathbf{u}_e & \text{on } \Gamma_e(t), \\ \boldsymbol{\sigma}_f \cdot \mathbf{n}_e = \boldsymbol{\sigma}_e \cdot \mathbf{n}_e & \text{on } \Gamma_e(t), \\ \mathbf{u}_e = \mathbf{u}_s & \text{on } \Gamma_s(t), \end{cases} \tag{1c}$$

where the density ρ, the velocity field $\mathbf{u}$, and the stress tensor $\boldsymbol{\sigma}$ are defined individually for each medium, and p is the pressure. The details concerning the stress tensors will be given in the following section. This system is composed of three subproblems (1a), (1b) and (1c) which are related to the fluid dynamics, the equation of motion for the elastic structure and the coupling conditions at the fluid/solid interfaces Γ_e and Γ_s. The quantity $\mathbf{u}_s$ refers to the imposed velocity of the rigid interface.

2.2 *The Monolithic Approach*

We develop an Eulerian method for simulating fluid-structure interactions which includes hyperelastic materials. In Lagrangian approaches, the hyperelastic constitutive law is defined from the deformation gradient tensor $\mathbf{F} = [\nabla \mathbf{X}]$ where $\mathbf{X}$ denotes the coordinates in the current domain $\Omega_e(t)$, with respect to the reference domain. Instead, in the present Eulerian representation, the deformation of the elastic material is described using backward characteristics $\mathbf{Y} : [0, T] \times \Omega_e(t) \longrightarrow \Omega_e(t = 0)$, being the inverse transformation of $\mathbf{X}$, i.e. $\mathbf{Y}(t, \mathbf{X}(t, \boldsymbol{\xi})) = \boldsymbol{\xi}$ and $\mathbf{X}(t, \mathbf{Y}(t, \mathbf{x})) = \mathbf{x}$, for all $\boldsymbol{\xi} \in \Omega_e(0)$ and $\mathbf{x} \in \Omega_e(t)$.

Instead of considering the system (1), we solve the whole fluid-structure system in a "monolithic" way, by using a single-continuum model for the entire domain Ω:

$$\rho\Big(\frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u} \cdot \nabla)\mathbf{u}\Big) = -\nabla p + \nabla \cdot \boldsymbol{\sigma}(\mathbf{u}, \boldsymbol{\psi}) + \frac{\chi_s}{\varepsilon}(\mathbf{u}_s - \mathbf{u}) \tag{2a}$$

$$\nabla \cdot \mathbf{u} = 0 \tag{2b}$$

$$\frac{\partial \boldsymbol{\psi}}{\partial t} + \nabla \cdot (\mathbf{u} \otimes \boldsymbol{\psi}) = 0 \tag{2c}$$

where the vector $\boldsymbol{\psi} = (\phi, \mathbf{Y})^T$ contains the quantities involved to describe the deformation of the elastic material, namely the level-set function ϕ being the Euclidian distance to the interface $\Gamma_e(t)$, and the backward characteristics $\mathbf{Y}$. These quantities are transported in time with velocity $\mathbf{u}$ in entire Ω according to Eq. (2c). The pressure p is defined in the whole domain, making no distinction between media.

The interface $\Gamma_e(t)$ is diffused on a small narrow band using a regularized Heaviside function $\widetilde{\chi}_e = \widetilde{\chi}_e(\phi)$ defined as a continuous function satisfying $\widetilde{\chi}_e = 1$ inside the elastic material, $\widetilde{\chi}_e = 0$ inside the fluid, and $0 < \widetilde{\chi}_e < 1$ in a small narrow band of the interface. This ensures that the solution is continuous accross $\Gamma_e(t)$ and the coupling constraints (1c) are hence properly satisfied. In that continuum formulation, the physical quantities (i.e. velocity, viscosity, and density) are defined in Ω as a mixture between fluid and elastic quantities as:

$$\begin{aligned}
\mathbf{u} &= (1 - \widetilde{\chi}_e)\mathbf{u}_f + \widetilde{\chi}_e \mathbf{u}_e \\
\mu &= (1 - \widetilde{\chi}_e)\mu_f + \widetilde{\chi}_e \mu_e \\
\rho &= (1 - \widetilde{\chi}_e)\rho_f + \widetilde{\chi}_e \rho_e
\end{aligned}$$

We consider a visco-hyperelastic model in which the solid deformation is described for two-parameter Mooney-Rivlin materials. The Cauchy stress tensor is expressed as:

$$\boldsymbol{\sigma}(\mathbf{u}, \boldsymbol{\psi}) = \mu(\nabla\mathbf{u} + \nabla\mathbf{u}^T) + \widetilde{\chi}_e(\phi)\boldsymbol{\sigma}_e(\mathbf{Y}) \tag{3}$$

The elastic stress tensor depends on the left Cauchy-Green deformation tensor $\mathbf{B} = \mathbf{F}\mathbf{F}^T = [\nabla\mathbf{Y}]^{-1}[\nabla\mathbf{Y}]^{-T}$ and its inverse:

$$\boldsymbol{\sigma}_e(\mathbf{Y}) = -(2c_1 I_1 - 2c_2 I_2)\mathbf{I} + 2c_1\mathbf{B} - 2c_2\mathbf{B}^{-1} \tag{4}$$

where I_1 and I_2 are the first and second invariants of $\mathbf{B}$ and $c_1, c_2 > 0$ are empirical constants of the material related to the shear modulus $G = 2(c_1 + c_2)$. This kind of model is particularly adapted to material which undergoes large deformations.

The rigid material is taken into account in (2a) using a penalization method [10], via a permeability parameter $\varepsilon \ll 1$. The characteristic function χ_s is 1 inside Ω_s and 0 elsewhere.

3 Discretization of the Governing Equations

3.1 Time Integration

The momentum (2a) and transport (2c) equations are solved independently, in a decoupled way. First, the backward characteristics $\mathbf{Y}$ and the level-set function ϕ are transported from time t^n to time t^{n+1} using a two-stage Runge-Kutta scheme.

These quantities are then used for the computation of the elastic stress tensor $\boldsymbol{\sigma}_e$ (4). Then, the fractional time step method introduced by Chorin [11] and Temam [12] is considered. We use a second-order Gear scheme as a time discretization of the prediction step. To guarantee the mass conservation, a projection/correction step is performed as in [13].

3.2 Finite Volume Discretizations

We perform a graded quadtree discretization of the whole computational domain Ω. Thanks to the library PABLO, as a part of Bitpit library,[1] we get access to an optimized tool for storing the data structure. This library allows an efficient Adaptative Mesh Refinement (AMR) to adapt the mesh dynamically to the solution, in order to preserve high accuracy during the whole simulation. For the domain decomposition, the number of communications between processors is limited to only one layer of ghost cells, which results in the development of compact numerical schemes. The finite-volume discretizations involved are:

- **the divergence operator**
  A second order quadrature formula is used for the approximation of the surface integrals. The divergence of a vector field $\mathbf{v}$ is then computed in a cell Ω_i as:

  $$(\nabla \cdot \mathbf{v})_i = \frac{1}{|\Omega_i|} \int_{\partial \Omega_i} \mathbf{v} \cdot \mathbf{n} \, ds = \frac{1}{|\Omega_i|} \sum_{f \subset \partial \Omega_i} \mathbf{v}_{fc} \cdot \mathbf{n}_f |f|$$

  where f denotes a face of the cell boundary $\partial \Omega_i$ and $\mathbf{n}_f$ is the unitary outward normal vector of f. $|\Omega_i|$ and $|f|$ denotes the area of cell Ω_i and the length of the face f respectively. The face-center quantity $\mathbf{v}_{fc}$ is interpolated thanks to Gaussian-type Radial Basis Functions (RBF) within the compact stencil composed of all cells surrounding the face.
- **the Laplacian operator**
  We use the diamond finite volume scheme proposed by [14, 15]. The face-center normal derivative is approximated as a linear combination involving the cells of the compact stencil of the face.
- **the convective/transport numerical flux**
  For any scalar function φ, the computation of $\nabla \cdot \mathbf{F}(\varphi)$ where $\mathbf{F}(\varphi) = U\varphi$ is inspired of the compact third order CWeno introduced by [16]. The conservative form is considered since the mass conservation is preserved after the projection step ($\nabla \cdot U = 0$). This corrected face-center velocity denoted by U_{fc} is then used for the calculation of the flux. A linear piecewise polynomial $\tilde{\varphi}$ is reconstructed using φ and its gradient. For any face $f = \partial \Omega_{in} \cap \partial \Omega_{out}$, we denote by $\mathbf{n}_f$ the normal vector pointing from Ω_{in} to Ω_{out} and $\mathbf{x}_{fc}$ the center of f. The quantity φ_{fc}

[1] https://optimad.github.io/bitpit.

is interpolated from both sides of the face, namely we have $\varphi^- := \varphi|_{\Omega_{in}}(\mathbf{x}_{fc})$ and $\varphi^+ := \varphi|_{\Omega_{out}}(\mathbf{x}_{fc})$. Finally, the monotone Rusanov numerical flux is employed, which has the form:

$$\mathcal{F}(\varphi^+, \varphi^-) = \frac{1}{2}U_{fc}(\varphi^+ + \varphi^-) - \frac{1}{2}|U_{fc}|(\varphi^+ - \varphi^-). \tag{5}$$

4 Results

4.1 A Solid Deformation in a Lid-Driven Cavity Flow

A validation of the model is performed on uniform cartesian grids. The FSI test case is based on the lid-driven cavity flow test. We perform a fully Eulerian simulation of a deformable solid immersed in a lid-driven cavity flow. In a cavity $\Omega = [0, 1]^2$, an elastic cylinder is immersed in a fluid of density $\rho_f = 1$ and viscosity $\mu_f = 0.01$. Initially, the centroid of the cylinder is $\mathbf{x}_c(t = 0) = (0.6, 0.5)^T$ and its diameter is 0.4. The densities and viscosities of the elastic structure and fluid are identical, i.e. $\rho_e = \rho_f$ and $\mu_e = \mu_f$. The elastic structure is a Neo-Hookean material ($c_2 = 0$) for which the shear modulus is set to $G = 0.1$. The results are compared with previous works (Sugiyama [17] and Deborde [18]) in Fig. 2.

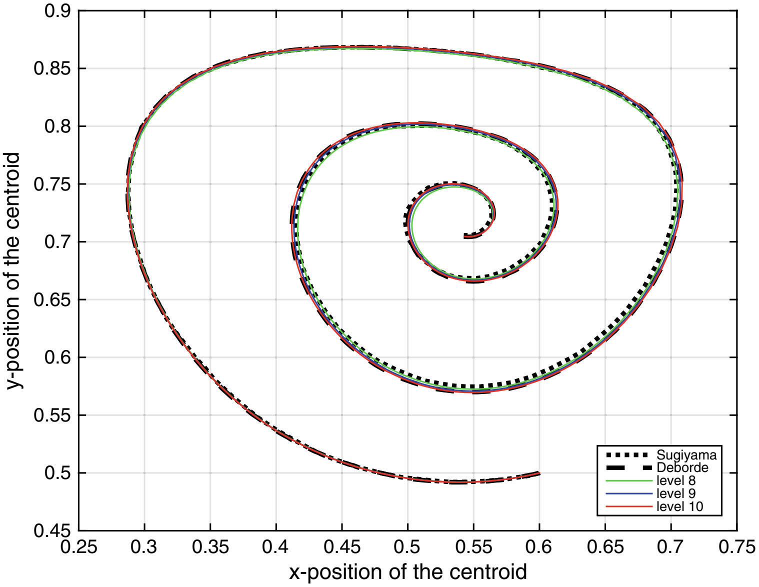


Fig. 2 Approximated position of the centroid over time in the lid-driven cavity flow test. The simulations are run for different levels of refinement

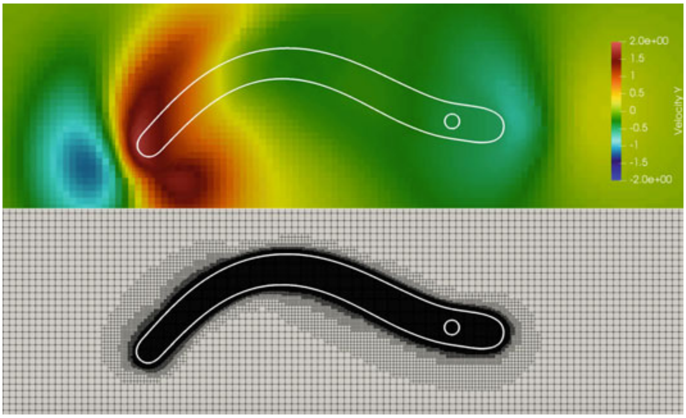


Fig. 3 Y-component of the velocity for the oscillating membrane in glycerin test after 3 periods of oscillations. The dynamic AMR mesh is shown in background

4.2 Hyperelastic Oscillating Membrane in Glycerin

Inside a 3 cm by 3 cm cavity, an hyperelastic membrane (rubber type, $G = 1\,\text{MPa}$) is immersed inside glycerin ($\rho_f = 1.26\,\text{g cm}^{-3}$, $\mu_f = 1.49\,\text{Pa s}$). The membrane is 1.95 cm long and the thickness varies between 1.2 and 1.7 mm. The membrane is actuated with gradual speed by an oscillating rigid cylindrical holder which is positioned on the right tip of the membrane. The simulations are performed on quadtree grids, using a frequent dynamic adaptation of the mesh according to the level-set function ϕ, see Fig. 3.

5 Conclusions and Prospects

In this work we proposed a finite-volume scheme for solving the single-continuum model (2). In Sect. 3.2, we introduced the numerical Rusanov (Local Lax-Friedrichs) flux to compute the convective/transport flux (see (5)). In this formulation, the stabilization is simply performed according to the normal face-center velocity U_{fc}, without considering the velocity of the waves which propagate inside the material. Hence, the numerical scheme is stable only for moderately stiff material (as in test 4.1) or for high viscosities (as in test 4.2) since wave and fluid velocities are similar. Ongoing work is carried out to develop a new scheme for which the flux is stabilized for stiff non-viscous materials.

References

1. Taymans, C.: Solving incompressible Navier-Stokes equations on Octree grids: towards application to wind turbine blade modelling. Doctoral dissertation, Bordeaux (2018)
2. Moireau, P., Xiao, N., Astorino, M., Figueroa, C.A., Chapelle, D., Taylor, C.A., Gerbeau, J.F.: External tissue support and fluid-structure simulation in blood flows. Biomech. Model. Mechanobiol. **11**(1–2), 1–18 (2012)
3. Nitikitpaiboon, C., Bathe, K.J.: An arbitrary Lagrangian-Eulerian velocity potential formulation for fluid-structure interaction. Comput. Struct. **47**(4–5), 871–891 (1993)
4. Turek, S., & Hron, J.: Proposal for numerical benchmarking of fluid-structure interaction between an elastic object and laminar incompressible flow. In: Fluid-Structure Interaction, pp. 371–385 (2006). Springer, Berlin, Heidelberg
5. Tezduyar, T.E., Behr, M., Mittal, S., Liou, J.: A new strategy for finite element computations involving moving boundaries and interfaces—the deforming-spatial-domain/space-time procedure: I. The concept and preliminary tests. Comput. Methods Appl. Mech. Eng. **94**(3), 339–351 (1992)
6. Tezduyar, T.E., Behr, M., Mittal, S., Liou, J.: A new strategy for finite element computations involving moving boundaries and interfaces—the deforming-spatial-domain/space-time procedure: II. Computation of free-surface flows, two-liquid flows, and flows with drifting cylinders. Comput. Methods Appl. Mech. Eng. **94**(3), 353–371 (1992)
7. Peskin, C.S.: Numerical analysis of blood flow in the heart. J. Comput. Phys. **25**(3), 220–252 (1977)
8. Peskin, C.S.: The fluid dynamics of heart valves: experimental, theoretical, and computational methods. Annu. Rev. Fluid Mech. **14**(1), 235–259 (1982)
9. Udaykumar, H.S., Shyy, W., Rao, M.M.: Elafint: a mixed Eulerian-Lagrangian method for fluid flows with complex and moving boundaries. Int. J. Numer. Methods Fluids **22**(8), 691–712 (1996)
10. Angot, P., Bruneau, C.H., Fabrie, P.: A penalization method to take into account obstacles in incompressible viscous flows. Numer. Math. **81**(4), 497–520 (1999)
11. Chorin, A.J.: Numerical solution of the Navier-Stokes equations. Math. Comput. **22**(104), 745–762 (1968)
12. Temam, R.: Sur l'approximation de la solution des équations de Navier-Stokes par la méthode des pas fractionnaires (II). Arch. Ration. Mech. Anal. **33**(5), 377–385 (1969)
13. Bergmann, M., Hovnanian, J., Iollo, A.: An accurate cartesian method for incompressible flows with moving boundaries. Commun. Comput. Phys. **15**(5), 1266–1290 (2014)
14. Coudière, Y., Vila, J.P., Villedieu, P.: Convergence rate of a finite volume scheme for a two dimensional convection-diffusion problem. ESAIM: Math. Model. Numer. Anal. **33**(3), 493–516 (1999)
15. Delcourte, S., Domelevo, K., Omnes, P.: Discrete Duality Finite Volume Method for Second Order Elliptic Problems. Hermes Science publishing, pp. 447–458 (2005)
16. Semplice, M., Coco, A., Russo, G.: Adaptive mesh refinement for hyperbolic systems based on third-order compact WENO reconstruction. J. Sci. Comput. **66**(2), 692–724 (2016)
17. Sugiyama, K., Ii, S., Takeuchi, S., Takagi, S., Matsumoto, Y.: A full Eulerian finite difference approach for solving fluid-structure coupling problems. J. Comput. Phys. **230**(3), 596–627 (2011)
18. Deborde, J.: Modélisation et simulation de l'interaction fluide-structure élastique: application à l'atténuation des vagues. Doctoral dissertation, Bordeaux (2017)

Author Index

R. Klöfkorn et al. (eds.), *Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples*, Springer Proceedings in Mathematics & Statistics 323,
https://doi.org/10.1007/978-3-030-43651-3