Springer Proceedings in Mathematics & Statistics

Emmanuel Franck
Jürgen Fuhrmann
Victor Michel-Dansac
Laurent Navoret *Editors*

Finite Volumes for Complex Applications X—Volume 1, Elliptic and Parabolic Problems

FVCA10, Strasbourg, France, October 30, 2023–November 03, 2023, Invited Contributions

Springer

Springer Proceedings in Mathematics & Statistics

Volume 432

This book series features volumes composed of selected contributions from workshops and conferences in all areas of current research in mathematics and statistics, including data science, operations 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.

Emmanuel Franck · Jürgen Fuhrmann ·
Victor Michel-Dansac · Laurent Navoret
Editors

Finite Volumes for Complex Applications X—Volume 1, Elliptic and Parabolic Problems

FVCA10, Strasbourg, France, October 30, 2023–November 03, 2023, Invited Contributions

Springer

Editors
Emmanuel Franck
Université de Strasbourg, CNRS, Inria
IRMA
Strasbourg, France

Victor Michel-Dansac
Université de Strasbourg, CNRS, Inria
IRMA
Strasbourg, France

Jürgen Fuhrmann
Weierstrass Institute for Applied Analysis
and Stochastics, Numerical Mathematics
and Scientific Computing
Berlin, Germany

Laurent Navoret
Université de Strasbourg, CNRS, Inria
IRMA
Strasbourg, France

ISSN 2194-1009 ISSN 2194-1017 (electronic)
Springer Proceedings in Mathematics & Statistics
ISBN 978-3-031-40863-2 ISBN 978-3-031-40864-9 (eBook)
https://doi.org/10.1007/978-3-031-40864-9

Mathematics Subject Classification: 65-06, 65Mxx, 65Nxx, 74Sxx, 76Mxx, 80Mxx, 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

Organizing committee

Joubine Aghili, Université de Strasbourg, IRMA UMR CNRS 7501, Inria, Strasbourg, France
Clémentine Courtès, Université de Strasbourg, IRMA UMR CNRS 7501, Inria, Strasbourg, France
Emmanuel Franck, Université de Strasbourg, CNRS, Inria, IRMA, Strasbourg, France
J*ürgen Fuhrmann*, Weierstrass Institute for Applied Analysis and Stochastics, Numerical Mathematics and Scientific Computing, Berlin, Deutschland
Philippe Helluy, Université de Strasbourg, IRMA UMR CNRS 7501, Inria, Strasbourg, France
Victor Michel-Dansac, Université de Strasbourg, CNRS, Inria, IRMA, Strasbourg, France
Laurent Navoret, Université de Strasbourg, IRMA UMR CNRS 7501, Inria, Strasbourg, France
Andrea Thomann, Université de Strasbourg, CNRS, Inria, IRMA, Strasbourg, France

Scientific Committee

Marianne Bessemoulin-Chatard, LMJL, Université de Nantes, France
Franck Boyer, IMT, Université Toulouse 3—Paul Sabatier, France
Konstantin Brenner, Université Côte d'Azur, Inria, CNRS, LJAD, Nice, France
Donna Calhoun, Department of Mathematics, Boise State University, USA
Claire Chainais-Hillairet, Université de Lille, France
Pietro Marco Congedo, Inria, Centre de Mathématiques Appliquées, École Polytechnique, IPP, Palaiseau, France

Andreas Dedner, Warwick Mathematics Institute, University of Warwick, United Kingdom

Jerome Droniou, School of Mathematics, Monash University, Melbourne, Australia

Michael Dumbser, University of Trento, Department of Civil, Environmental and Mechanical Engineering, Trento, Italy

Isabelle Faille, IFP Energies nouvelles, Rueil-Malmaison, France

Jiří Fürst, Czech Technical University in Prague, Faculty of Mechanical Engineering, Department of Technical Mathematics, Prague, Czech Republic

Thierry Gallouët, I2M UMR 7373, Aix-Marseille Université, CNRS, Ecole Centrale de Marseille, Marseille, France

Raphaèle Herbin, 3I2M UMR 7373, Aix-Marseille Université, CNRS, Ecole Centrale de Marseille, Marseille, France

Jan S. Hesthaven, Chair of Computational Mathematics and Simulation Science, École Polytechnique Fédérale de Lausanne, Switzerland

Jean-Marc Hérard, EDF R&D, Chatou, France

Volker John, Weierstrass Institute for Applied Analysis and Stochastics (WIAS), Berlin, Germany

Birane Kane, NORCE Norwegian Research Centre AS, Bergen, Norway

Eirik Keilegavlen, Center for Modeling of Coupled Subsurface Dynamics, Department of Mathematics, University of Bergen, Norway

Robert Klöfkorn, Lund University, Sweden

Konstantin Lipnikov, Applied Mathematics and Plasma Physics Group, Theoretical Division, Los Alamos National Laboratory, Los Alamos, USA

Pierre-Henri Maire, CEA Cesta, Le Barp, France

Sandra May, Uppsala University, Department of Information Technology, Sweden

Gabriella Puppo, Dipartimento di Matematica, La Sapienza Università di Roma, Italy

Jing-Mei Qiu, Department of Mathematics, University of Delaware, Newark, USA

Adrian Florin Radu, Center for Modeling of Coupled Subsurface Dynamics, University of Bergen, Norway

Preface

The finite volume method is a spatial discretization technique for partial differential equations based on the physical principle of local conservation. It has been successfully used in many applications, including fluid dynamics, magnetohydrodynamics, nuclear physics, or plasma physics. Motivated by their large applicability to real-world problems, finite volumes have been the purpose of an intensive research effort in the last decades, yielding significant progress in the design, the numerical analysis, or the practical implementation of the methods.

Research on finite volumes remains very active, as problems are becoming more and more complex. Among the current challenges addressed by the scientific community, let us mention, for instance, the design of robust (with respect to the mesh and/or physical parameters) numerical methods, high-order methods, methods preserving structural properties (positivity or dissipation of a prescribed quantity), or methods enhanced by machine learning approaches.

Previous conferences in this series have been held in Rouen (1996), Duisburg (1999), Porquerolles (2002), Marrakech (2005), Aussois (2008), Prague (2011), and Berlin (2014), Lille (2017) and Bergen (2020, online).

The present volumes contain the invited and contributed papers presented as posters or talks at the 10th International Symposium on Finite Volumes for Complex Applications held at the University of Strasbourg, from October 30 to November 3, 2023.

The first volume contains the invited contributions, as well as contributed papers focusing on finite volume schemes for elliptic and parabolic problems. Topics of the contributed papers include structure-preserving schemes, convergence proofs, and error estimates.

The second volume is focused on finite volume methods for hyperbolic and related problems, such as methods compatible with the low Mach number limit or able to exactly preserve steady solutions, the development and analysis of high order methods, or the discretization of kinetic equations.

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

Finally, we warmly thank the “Cellule Congrès” of the University of Strasbourg, as well as the organizing committee for helping make this conference a great success.

Strasbourg, France	Emmanuel Franck
Berlin, Germany	Jürgen Fuhrmann
Strasbourg, France	Victor Michel-Dansac
Strasbourg, France	Laurent Navoret

Contents

Invited Papers

Elliptic and Parabolic Problems: Contributed Papers

Invited Papers

A Personal Discussion on Conservation, and How to Formulate It

Rémi Abgrall

Abstract Since the celebrated theorem of Lax and Wendroff, we know a necessary condition that any numerical scheme for hyperbolic problem should satisfy: it should be written in flux form. A variant can also be formulated for the entropy. Even though some schemes, as for example those using continuous finite element, do not formally cast into this framework, it is a very convenient one. In this paper, we revisit this, introduce a different notion of local conservation which contains the previous one in one space dimension, and explore its consequences. This gives a more flexible framework that allows to get, systematically, entropy stable schemes, entropy dissipative ones, or accommodate more constraints. In particular, we can show that continuous finite element method can be rewritten in the finite volume framework, and all the quantities involved are explicitly computable. We end by presenting the only counter example we are aware of, i.e a scheme that seems not to be rewritten as a finite volume scheme.

Keywords Local conservation · Lax Wendroff theorem · Adding constraints · Flux · Entropy dissipation

In this paper, we will consider approximations of the hyperbolic problem

$$\frac{\partial \mathbf{u}}{\partial t} + \operatorname{div} \mathbf{f}(\mathbf{u}) = 0 \tag{1a}$$

which, for simplicity, we will assume to be defined on $\mathbb{R}^d$ ($d = 1, 2, 3$), and $\mathbf{u}$ is defined on $\mathbb{R}^d \times [0, T[$, $T > 0$, with values in $\Omega \subset \mathbb{R}^p$, Ω open. The flux $\mathbf{f} = (\mathbf{f}_1, \dots, \mathbf{f}_d)$ is such that each $\mathbf{f}_j$ are defined on Ω with values in $\mathbb{R}^p$. The flux $\mathbf{f}$ is assumed to be C^1 on Ω^d, and the hyperbolicity means that for any

R. Abgrall (✉)
Institute für Mathematik, Zürich Universität, CH 8057 Zürich, Switzerland
e-mail: remi.abgrall@math.uzh.ch
URL: https://www.math.uzh.ch/people?key1=8882

E. Franck et al. (eds.), *Finite Volumes for Complex Applications X—Volume 1, Elliptic and Parabolic Problems*, Springer Proceedings in Mathematics & Statistics 432,
https://doi.org/10.1007/978-3-031-40864-9_1

$\mathbf{n}=(n_1,\dots,n_d)\in\mathbb{R}^d$, the matrix $\nabla\mathbf{f}\cdot\mathbf{n}:=\sum_{j=1}^d \dfrac{\partial \mathbf{f}_j}{\partial\mathbf{u}}n_j$ is diagonalisable in $\mathbb{R}$. The PDE (1a) is supplemented with initial condition

$$\mathbf{u}(\mathbf{x},0)=\mathbf{u}_0(\mathbf{x})$$

for some initial condition with values in Ω. We will no do any theoretical consideration about this problem, and refer to [15] for more details.

In order to integrate (1), several choices need to be done: the type of mesh, the type of functional approximation, how to discretise the divergence term, and finally the time stepping strategy. Often, one considers a tessellation of the computational domain by polygons. A common choice is to choose an approximation space V_h which is a subset of $L^2(\mathbb{R}^d)$, then a variational approximation allows to approximate the divergence term, and finally the method of lines is used for time discretisation. Often, the choice $V_h\subset L^2(\mathbb{R}^d)$ hides a common belief that global continuity is not a good idea for (1) which allows discontinuous solutions.

In addition, the system (1) is complemented by a differential inequality about a state function, the entropy η, a convex function of $\mathbf{u}\in\Omega$ (and hence Ω is assumed to be convex from now on). There exists also $\mathbf{g}=(\mathbf{g}_1,\dots,\mathbf{g}_d)$, C^1 such that for any $j=1,\dots,d$

$$\nabla_{\mathbf{u}}\eta\frac{\partial \mathbf{f}_i}{\partial\mathbf{u}}=\frac{\partial \mathbf{g}_i}{\partial\mathbf{u}}.$$

From this, we see that

$$\frac{\partial\eta}{\partial t}+\text{ div }\mathbf{g}=0.$$

For non smooth solution, we impose

$$\frac{\partial\eta}{\partial t}+\text{ div }\mathbf{g}\leq 0 \tag{2}$$

which is true in the sense of distribution.

The existence of (2) is motivated by the canonical example: the Euler equations. Here, the conserved variables are $\mathbf{u}=(\rho,\rho\mathbf{v},E)^T$ where ρ is the density, $\mathbf{v}$ is the fluid velocity, and $E=e+\frac{1}{2}\rho\mathbf{v}^2$ is the total energy which is the sum of the internal energy e and the kinetic energy $\frac{1}{2}\rho\mathbf{v}^2$. In the simplest case, one can write the internal energy as a function of the density and the pressure p, $e=e(\rho,p)$, and the entropy is also a function of these variables, $\eta=\eta(\rho,p)$. We can navigate from one thermodynamic set of two variables, for example $\{\rho,p\}$ to $\{e,\eta\}$ and vice versa. We do enter into more thermodynamic consideration, see eg. [18]. The conserved variable satisfy (1) with the flux defined by

$$\mathbf{f}=\begin{pmatrix}\rho\mathbf{v}\\ \rho\mathbf{v}\otimes\mathbf{v}+p\mathrm{Id}_{d\times d}\\ (E+p)\mathbf{v}\end{pmatrix} \tag{3}$$

The entropy flux is $\mathbf{g} = \mathbf{v}\eta$ and in the simplest thermodynamics case, namely the case of a calorically perfect gas, $p = (\gamma - 1)e$ where γ is a constant and $\eta = \rho\big(\log p - \gamma \log \rho) - \eta_0$ where η_0 is a reference. The system (1) with the flux (3) is hyperbolic in the domain

$$\Omega = \{\mathbf{u} \text{ such that } \rho > 0 \text{ and } e > 0\},$$

and the entropy is a convex function of $\mathbf{u}$ if $\gamma > 1$.

The simplest example of non linear problem is the Burgers equation (which can be obtained from the Euler equation in the case $\gamma = 3$). It is written (in the inviscid case) as

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = 0 \text{ or } \frac{\partial u}{\partial t} + \frac{1}{2}\frac{\partial u^2}{\partial x} = 0. \tag{4}$$

Taking $u(x, 0) = \sin(\pi x)$, $x \in [-1, 1]$, it is easy to see with the method of characteristics that after $t^\star = \frac{1}{\pi}$ the solution cannot be smooth, so that the relation (4) on the left has no meaning. The same would held for the non conservative form of the Euler equation, for example

$$\frac{\partial}{\partial t}\begin{pmatrix}\rho\\ \mathbf{v}\\ p\end{pmatrix} + \begin{pmatrix}\text{div }\big(\rho\mathbf{v}\big)\\ \big(\mathbf{v}\cdot\nabla\big)\mathbf{v} + \frac{\nabla p}{\rho}\\ \mathbf{v}\cdot\nabla p + \rho c^2 \text{div }\mathbf{v}\end{pmatrix} = 0 \tag{5}$$

though these relations are more interesting for practitioners (since we have a direct access to the pressure and the velocity).

Motivated by this, P. Lax has formalized what was already known in the engineering community by taking volumes and looking at what is getting into and out of them. It is well known that a smooth function is a solution of (1a) if and only if for any smooth test function $\varphi \in C_0^1(\mathbb{R}^d \times \mathbb{R}^+)$ with compact support, we have

$$\int_{\mathbb{R}^d\times\mathbb{R}^+} \big(\frac{\partial \varphi}{\partial t}\mathbf{u} + \nabla\varphi \cdot \mathbf{f}(\mathbf{u})\big)\, d\mathbf{x}\, dt + \int_{\mathbb{R}^d} \varphi(\mathbf{x}, 0)u_0(\mathbf{x})\, d\mathbf{x} = 0 \tag{6}$$

From this we can define the notion of weak solution. Similarly, we have the weak form of the entropy inequality: taking any positive test function, we get

$$\int_{\mathbb{R}^d\times\mathbb{R}^+} \big(\frac{\partial \varphi}{\partial t}\eta(\mathbf{u}) + \nabla\varphi \cdot \mathbf{g}(\mathbf{u})\big)\, d\mathbf{x}\, dt + \int_{\mathbb{R}^d} \varphi(\mathbf{x}, 0)\eta_0(\mathbf{x})\, d\mathbf{x} \geq 0 \tag{7}$$

This notion of weak solution is the guiding line of numerical discretisation. The celebrated Lax-Wendroff theorem is “simply” a way to mimic this notion, and it provides a generic from of numerical schemes that allows to guaranty, under natural conditions, the convergence to weak entropy solutions. For example, in one dimension,

- being given a regular mesh $\{x_j\}_{j\in\mathbb{Z}}$, and defining control volumes $C_j = (x_{j-1/2}, x_{j+1/2})$ with $x_{j+1/2} = \frac{x_j+x_{j+1}}{2}$,
- given an initialization[1]
$$u_j^0 \approx \frac{1}{|C_j|}\int_{C_j} u_0(\mathbf{x})\, d\mathbf{x}$$
- A numerical scheme of the form
$$\mathbf{u}_j^{n+1} = \mathbf{u}_j^n - \frac{\Delta t}{|C_j|}\big(\hat{\mathbf{f}}_{j+1/2} - \hat{\mathbf{f}}_{j-1/2}\big)$$
where $\hat{\mathbf{f}}_{k+1/2} = \hat{\mathbf{f}}(u_{k-p}, \dots, \mathbf{u}_k \dots, \mathbf{u}_{k+p})$ is a Lipschitz continuous function depending $2p+1$ arguments centered around $\mathbf{u}_k$.

Then we know that if the numerical flux is consistant : $\hat{\mathbf{f}}(u, \dots, u) = \mathbf{f}(u)$, if the the sequence $\{\mathbf{u}_j^n\}$ is bounded under $\Delta t/|C_j| \leq C$ and such that one subsequence converges to some $\mathbf{v}$ in L^2, then $\mathbf{v}$ is a weak solution. The same applies for entropy solutions.

Since about 60 years, all the research has turn around this result and variants. Is this the end of the story? Certainly not. Several natural questions arise:

- What happens when a scheme has no longer a flux form? Concerning this question, a partial answer was given by Hou and Le Floch in [19] where they show that a scheme written in incremental form
$$\mathbf{u}_i^{n+1} = \mathbf{u}_i^n - C_{i-1/2}^n(\mathbf{u}_i^n - \mathbf{u}_{i-1}^n) + D_{i+1/2}^n(\mathbf{u}_{i+1}^n - \mathbf{u}_i^n),$$
under suitable positivity constraints on the coefficients $C_{l+1/2}$ and $D_{l+1/2}$, a CFL type condition, that a subsequence converges to a function that is a weak solution of
$$\frac{\partial u}{\partial t} + \frac{\partial f}{\partial x} = \mu$$
where μ is Borel measure. The measure μ is conjectured to sit on the discontinuities of the solution. Not much more is said, in particular, μ could possibly vanish.
- It is know in any text book that a non linear change of variable is in general not permitted. The canonical example is again the Burgers equation. When we set $v = u^3$, the irregular solutions will satisfy
$$\frac{\partial v}{\partial t} + \frac{3}{4}\frac{\partial v^3}{\partial x} = 0.$$

[1] If the set S is discrete, $|S|$ is its cardinal. If S is part of a domain, it is its measure with respect to the lebesgue measure, i.e. its length/area/volume.

but the discontinuities will not travel at the same speeds, so that the shocks are different. However, it would be very interesting to have the possibility to change variable, think, for example, of the Euler equations in conservative and primitive variables.

The purpose of this paper is to explain some ways to overcome these two obstacles. The format of this paper is as follows. We first recall the classical setting of finite volume schemes and the notion of numerical flux, as well as the classical Lax-Wendroff theorem. Then we show a rewriting of the local conservation condition for these schemes, and describe several classical numerical schemes that satisfy this condition. We show that this condition is then equivalent to the existence of numerical flux, the only difference is that in general these flux are not standard.

Then using this condition, we show several extensions using a non conservative form of a conservative problem, how to modify a scheme to satisfy one or more additional conservative constraints (such an entropy inequality), the use of stagerred grids. We conclude by show an example of scheme that does not seem to be cast in the same framework, though can be shown leading to proper weak solutions of the problem.

1 Classical Conservation Versus RD

In this section, we rephrase in part the content of [6]. We first start from a standard finite volume scheme, and rewrite it in an equivalent form. In the one dimensional case, we simply say that $\hat{\mathbf{f}}_{j+1/2} - \hat{\mathbf{f}}_{j-1/2} = \hat{\mathbf{f}}_{j+1/2} - \mathbf{f}(u_j) + \mathbf{f}(u_j) - \hat{\mathbf{f}}_{j-1/2} = \Phi_i^{[x_i,x_{i+1}]} + \Phi_i^{[x_{i-1},x_i]}$ with

$$\Phi_i^{[x_i,x_{i+1}]} = \hat{\mathbf{f}}_{j+1/2} - \mathbf{f}(u_j), \quad \Phi_i^{[x_{i-1},x_i]} = \mathbf{f}(u_j) - \hat{\mathbf{f}}_{j-1/2}$$

so that the standard finite volume can be equivalently rewritten as

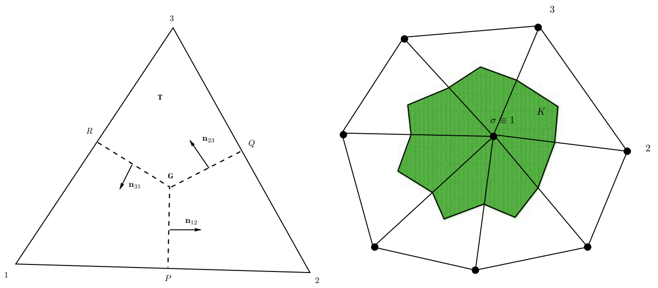
$$\mathbf{u}_i^{n+1} = \mathbf{u}_i^n - \frac{\Delta t}{|C_i|}\big(\Phi_i^{[x_i,x_{i+1}]} + \Phi_i^{[x_{i-1},x_i]}\big). \tag{8}$$

In addition, we note[2]

$$\Phi_i^{[x_i,x_{i+1}]} + \Phi_{i+1}^{[x_i,x_{i+1}]} = \mathbf{f}(\mathbf{u}_{i+1}) - \mathbf{f}(\mathbf{u}_i).$$

This can be extended to any kind of finite volume scheme. Instead of going into the full generality, let us take an example to show the principle. For this $\mathbb{R}^d$ is covered

[2] This is the essence of Roe's 1981 paper: setting $\Phi_i^{[x_i,x_{i+1}]} = a^-_{i+1/2}(\mathbf{u}_{i+1} - \mathbf{u}_i)$ and $\Phi_{i+1}^{[x_i,x_{i+1}]} = a^+_{j+1/2}(\mathbf{u}_{i+1} - \mathbf{u}_i)$, we see that the method of characteristics (8) is conservative if and only if $\mathbf{f}(\mathbf{u}_{i+1}) - \mathbf{f}(\mathbf{u}_i) = a_{i+1/2}(\mathbf{u}_{i+1} - \mathbf{u}_i)$.



(a) (b)

Fig. 1 Notations for the finite volume schemes. On the left: definition of the control volume for the degree of freedom σ. The vertex σ plays the role of the vertex 1 on the left picture for the triangle K. The control volume C_σ associated to $\sigma = 1$ is green on the right and corresponds to $1PGR$ on the left. The vectors $\mathbf{n}_{ij}$ are normal to the internal edges scaled by the corresponding edge length

by non overlapping simplex denoted by K. The vertices of the mesh are denoted by $\{\sigma_j\}$. For any σ, we consider a control volume obtained by joining the centroids $\mathbf{x}_K$ of the simplex sharing σ and the mid points of the edges coming out of σ, see Fig. 1 for the notations.

Again, we specialize ourselves to the case of triangular elements, but *exactly the same arguments* can be given for more general elements, provided a conformal approximation space can be constructed. This is the case for triangle elements, and we can take $k = 1$.

Since the boundary of C_σ is a closed polygon, the scaled outward normals $\mathbf{n}_\gamma$ to ∂C_σ sum up to 0:

$$\sum_{\gamma \subset \partial C_\sigma} \mathbf{n}_\gamma = 0$$

where γ is any of the segment included in ∂C_σ, such as PG on Fig. 1. The finite volume scheme writes

$$|C_\sigma|\mathbf{u}_\sigma^{n+1} = |C_\sigma|\mathbf{u}_\sigma^n - \Delta t \sum_{\gamma \subset \partial C_\sigma} \hat{\mathbf{f}}_{\mathbf{n}_\gamma}(\mathbf{u}_\sigma, \mathbf{u}_\gamma^-) \tag{9}$$

where $\hat{\mathbf{f}}$ is a consistant numerical flux, and $\mathbf{u}_\gamma^-$ is the argument on the other side of γ. It can be evaluated via the MUSCL method, for example. Looking at the spatial increment, we see that

$$\begin{aligned}\sum_{\gamma \subset \partial C_\sigma} \hat{\mathbf{f}}_{\mathbf{n}_\gamma}(\mathbf{u}_\sigma, \mathbf{u}^-) &= \sum_{\gamma \subset \partial C_\sigma} \hat{\mathbf{f}}_{\mathbf{n}_\gamma}(\mathbf{u}_\sigma, \mathbf{u}^-) - \Bigg(\sum_{\gamma \subset \partial C_\sigma} \mathbf{n}_\gamma\Bigg) \cdot \mathbf{f}(\mathbf{u}_\sigma)\\ &= \sum_{K, \sigma \in K} \sum_{\gamma \subset \partial C_\sigma \cap K} \Big(\hat{\mathbf{f}}_{\mathbf{n}_\gamma}(\mathbf{u}_\sigma, \mathbf{u}^-) - \mathbf{f}(\mathbf{u}_\sigma) \cdot \mathbf{n}_\gamma\Big)\end{aligned}$$

To make things explicit, in K, the internal boundaries are PG, QG and RG, and those around $\sigma \equiv 1$ are PG and RG. We set

$$\begin{aligned}\Phi_\sigma^K(\mathbf{u}^h) &= \sum_{\gamma\subset\partial C_\sigma\cap K} \big(\hat{\mathbf{f}}_{\mathbf{n}_\gamma}(\mathbf{u}_\sigma,\mathbf{u}^-) - \mathbf{f}(\mathbf{u}_\sigma)\cdot\mathbf{n}_\gamma\big)\\ &= \sum_{\gamma\subset\partial(C_\sigma\cap K)} \hat{\mathbf{f}}_{\mathbf{n}_\gamma}(\mathbf{u}_\sigma,\mathbf{u}^-).\end{aligned} \tag{10}$$

The last relation uses the consistency of the flux and the fact that $C_\sigma \cap K$ is a closed polygon. The quantity $\Phi_\sigma^K(\mathbf{u}^h)$ is the normal flux on $C_\sigma \cap K$. If now we sum up these three quantities, we get:

$$\begin{aligned}\sum_{\sigma\in K}\Phi_\sigma^K(\mathbf{u}_h) &= \Big(\hat{\mathbf{f}}_{\mathbf{n}_{12}}(\mathbf{u}_1,\mathbf{u}_2) - \hat{\mathbf{f}}_{\mathbf{n}_{13}}(\mathbf{u}_1,\mathbf{u}_3) - \mathbf{f}(\mathbf{u}_1)\cdot\mathbf{n}_{12} + \mathbf{f}(\mathbf{u}_1)\cdot\mathbf{n}_{31}\Big)\\ &\quad + \Big(\hat{\mathbf{f}}_{\mathbf{n}_{23}}(\mathbf{u}_2,\mathbf{u}_3) - \hat{\mathbf{f}}_{\mathbf{n}_{12}}(\mathbf{u}_2,\mathbf{u}_1) + \mathbf{f}(\mathbf{u}_2)\cdot\mathbf{n}_{12} - \mathbf{f}(\mathbf{u}_2)\cdot\mathbf{n}_{23}\Big)\\ &\quad + \Big(-\hat{\mathbf{f}}_{\mathbf{n}_{23}}(\mathbf{u}_3,\mathbf{u}_2) + \hat{\mathbf{f}}_{\mathbf{n}_{31}}(\mathbf{u}_3,\mathbf{u}_1) - \mathbf{f}(\mathbf{u}_3)\cdot\mathbf{n}_{23} + \mathbf{f}(\mathbf{u}_3)\cdot\mathbf{n}_{31}\Big)\\ &= \mathbf{f}(\mathbf{u}_1)\cdot\big(\mathbf{n}_{12}-\mathbf{n}_{31}\big) + \mathbf{f}(\mathbf{u}_2)\cdot\big(-\mathbf{n}_{23}+\mathbf{n}_{31}\big) + \mathbf{f}(\mathbf{u}_3)\cdot\big(\mathbf{n}_{31}-\mathbf{n}_{23}\big)\\ &= \mathbf{f}(\mathbf{u}_1)\cdot\frac{\mathbf{n}_1}{2} + \mathbf{f}(\mathbf{u}_2)\cdot\frac{\mathbf{n}_2}{2} + \mathbf{f}(\mathbf{u}_3)\cdot\frac{\mathbf{n}_3}{2}\end{aligned}$$

where $\mathbf{n}_j$ is the scaled inward normal of the edge opposite to vertex σ_j, i.e. twice the gradient of the $\mathbb{P}^1$ basis function φ_{σ_j} associated to this degree of freedom. Thus, we can reinterpret the sum as the boundary integral of the Lagrange interpolant of the flux. The finite volume scheme is then a residual distribution scheme with residual defined by (10) and a total residual defined by

$$\Phi^K := \int_{\partial K} \mathbf{f}^h\cdot\mathbf{n}, \qquad \mathbf{f}^h = \sum_{\sigma\in K}\mathbf{f}(\mathbf{u}_\sigma)\varphi_\sigma. \tag{11}$$

Form now on, we will assume that the domain can be split into polygons K (above it was simplex) which vertex will be the σs. From these polygons, we can construct of control volumes noted C_σ. The other situation is the converse: from a family of control volumes, we can construct polygons K with vertices σ: the difference is between vertex centered or volume centered schemes in the Finite volume vocabulary. We will focus on schemes that can be written in the following form:

$$|C_\sigma|\mathbf{u}_\sigma^{n+1} = |C_\sigma|\mathbf{u}_\sigma^n - \Delta t \sum_{K,\sigma\in K}\Phi_\sigma^K(\mathbf{u}^n) \tag{12a}$$

with

$$\sum_{\sigma\in K} \Phi_\sigma^K(\mathbf{u}^n) = \int_{\partial K} \hat{\mathbf{f}}_{\mathbf{n}}\, d\gamma. \tag{12b}$$

In (12a), $\Phi_\sigma^K(\mathbf{u}^n)$ is a function that depends on a finite number values of the $\mathbf{u}^n_{\sigma'}$, it is assumed to be Lipschitz continuous. In (12b) we assume that for any edge/face that is the intersection of two polygons, $\gamma = K \cap K'$,

$$\int_{\gamma\subset K} \hat{\mathbf{f}}_{\mathbf{n}}\, d\gamma + \int_{\gamma\subset K'} \hat{\mathbf{f}}_{\mathbf{n}}\, d\gamma = 0.$$

In other words, the flux are the same, up to the sign.

This framework is not adapted uniquely to finite volume. We can consider, using continuous finite elements, the SUPG scheme. Considering a mesh made of simplex, we take

$$V^h = \{\mathbf{v}^h \in C^0(\mathbb{R}^d), \text{ for any } K, \mathbf{v}^h_{|K} \in \mathbb{P}^r\}$$

and look for $\mathbf{u}^h \in \big(V^h\big)^p$ such that for any $\mathbf{v}^h \in V^h$ we have

$$\sum_{K\subset\mathbb{R}^d} \Big(-\int_K \nabla v^h \cdot \mathbf{f}(\mathbf{u}^h)\, d\mathbf{x} + \int_{\partial K} v^h \mathbf{f}(\mathbf{u}^h)\cdot \mathbf{n}\, d\gamma + h_K \int_K (\nabla_{\mathbf{u}}\mathbf{f}\cdot\nabla v^h)\tau\big(\nabla_{\mathbf{u}}\mathbf{f}\cdot\nabla\mathbf{u}^h\big)\, d\mathbf{x}\Big) = 0 \tag{13}$$

By continuity, of course the boundary term cancel, but we have written the scheme in this way to exhibit the residuals: if $\{\varphi_\sigma\}$ is the set of Lagrange basis functions,

$$\Phi_\sigma^K = -\int_K \nabla v^h \cdot \mathbf{f}(\mathbf{u}^h)\, d\mathbf{x} + \int_{\partial K} v^h \mathbf{f}(\mathbf{u}^h)\cdot \mathbf{n}\, d\gamma + h_K \int_K \big(\nabla_{\mathbf{u}}\mathbf{f}\cdot\nabla v^h\big)\tau\big(\nabla_{\mathbf{u}}\mathbf{f}\cdot\nabla\mathbf{u}^h\big)\, d\mathbf{x}, \tag{14}$$

and we get the conservation relation:

$$\sum_{\sigma\in K} \Phi_\sigma^K = \int_{\partial K} \mathbf{f}(\mathbf{u}^h)\cdot\mathbf{n}\, d\gamma \tag{15}$$

because $\sum\limits_{\sigma\in K} \big(\varphi_\sigma\big)_{|K} = 1.$

All the schemes we know, except maybe one that we will sketch at the end of this paper, can be rewritten in a distribution form: this is true from finite volume (or any order), to dG, via continuous finite element with different stabilisation mechanisms. Some details can be found in [6]. This can also apply to schemes adapted to the Lagrangian formalism, see [22] for example.

Let us examine now the converse: assuming a scheme of the form (12a) with (12b), can we identify "flux" so that the scheme (12a) can be rewritten equivalently in the form (9). First, what is a flux?

Definition 1 (*consistent flux*) A flux function $\hat{\mathbf{f}}$ is a function that depends on a normal $\mathbf{n}$ and a set of arguments $\{\mathbf{u}_1, \dots, \mathbf{u}_N\}$ such that:

1. $\hat{\mathbf{f}}$ is continuous with respect to its arguments,
2. $\hat{\mathbf{f}}(-\mathbf{n}; \mathbf{u}_1, \dots, \mathbf{u}_N) = -\hat{\mathbf{f}}(\mathbf{n}; \mathbf{u}_1, \dots, \mathbf{u}_N)$
3. It is consistent if for any $\mathbf{u}$ and $\mathbf{n}$, $\hat{\mathbf{f}}(-\mathbf{n}; \mathbf{u}, \dots, \mathbf{u}) = \mathbf{f}(\mathbf{u}) \cdot \mathbf{n}$

We will also use the notation $\hat{\mathbf{f}}_{\mathbf{n}}(\mathbf{u}_1, \dots, \mathbf{u}_N)$ or simply $\hat{\mathbf{f}}_{\mathbf{n}}$.

Any K appearing in the sum (12a), contains a set of degrees of freedom, say $\mathcal{S} = \{\sigma_1, \dots, \sigma_m\}$ from which we can construct a graph, or a triangulation, which vertex are the σ_is. An important property of this graph is that it is simply connected. Then the question is to find $\{\hat{\mathbf{f}}_{\sigma\sigma'}\}_{\sigma,\sigma' \in \mathcal{S}}$ such that

$$\Phi_\sigma = \sum_{\text{edges } [\sigma,\sigma']} \hat{\mathbf{f}}_{\sigma,\sigma'} + \hat{\mathbf{f}}_\sigma^b \tag{16a}$$

with

$$\hat{\mathbf{f}}_{\sigma,\sigma'} = -\hat{\mathbf{f}}_{\sigma',\sigma} \tag{16b}$$

and $\hat{\mathbf{f}}_\sigma^b$ is the 'part' of $\int_{\partial K} \hat{\mathbf{f}}_{\mathbf{n}}(\mathbf{u}^h, \mathbf{u}^{h,-})\, d\gamma$ associated to σ. The control volumes will be defined by their normals so that we get consistency. The normal will be defined later, as well as the control volumes.

Note that (16b) implies the conservation relation

$$\sum_{\sigma \in K} \Phi_\sigma = \sum_{\sigma \in K} \hat{\mathbf{f}}_\sigma^b. \tag{16c}$$

In short, we will take

$$\hat{\mathbf{f}}_\sigma^b = \oint_{\partial K} \varphi_\sigma\, \hat{\mathbf{f}}_{\mathbf{n}}(\mathbf{u}^h, \mathbf{u}^{h,-})\, d\gamma, \tag{16d}$$

but other examples can be considered provided the consistency (16c) relation holds true, see [6]. Any edge $[\sigma, \sigma']$ is either direct or, if not, $[\sigma', \sigma]$ is direct. Because of (16b), we only need to know $\hat{\mathbf{f}}_{\sigma,\sigma'}$ for direct edges. Thus we introduce the notation $\hat{\mathbf{f}}_{\{\sigma,\sigma'\}}$ for the flux assigned to the direct edge whose extremities are σ and σ'. We can rewrite (16a) as, for any $\sigma \in \mathcal{S}$,

$$\sum_{\sigma' \in \mathcal{S}} \varepsilon_{\sigma,\sigma'} \hat{\mathbf{f}}_{\{\sigma,\sigma'\}} = \Psi_\sigma := \Phi_\sigma - \hat{\mathbf{f}}_\sigma^b, \tag{17}$$

with

$$\varepsilon_{\sigma,\sigma'} = \begin{cases} 0 & \text{if } \sigma \text{ and } \sigma' \text{ are not on the same edge of } \mathcal{T}, \\ 1 & \text{if } [\sigma, \sigma'] \text{ is an edge and } \sigma \to \sigma' \text{ is direct}, \\ -1 & \text{if } [\sigma, \sigma'] \text{ is an edge and } \sigma' \to \sigma \text{ is direct}. \end{cases}$$

Hence the problem is to find a vector $\hat{\mathbf{f}} = (\hat{\mathbf{f}}_{\{\sigma,\sigma'\}})_{\{\sigma,\sigma'\}\text{ direct edges}}$ such that

$$A\hat{\mathbf{f}} = \Psi$$

where $\Psi = (\Psi_\sigma)_{\sigma\in\mathcal{S}}$ and $A_{\sigma\sigma'} = \varepsilon_{\sigma,\sigma'}$.

We have the following lemma [6]which shows the existence of a solution.

Lemma 1 *For any couple $\{\Phi_\sigma\}_{\sigma\in\mathcal{S}}$ and $\{\hat{\mathbf{f}}_\sigma^b\}_{\sigma\in\mathcal{S}}$ satisfying the condition (16c), there exists numerical flux functions $\hat{\mathbf{f}}_{\sigma,\sigma'}$ that satisfy (16). Recalling that the matrix of the Laplacian of the graph is $L = AA^T$, we have*

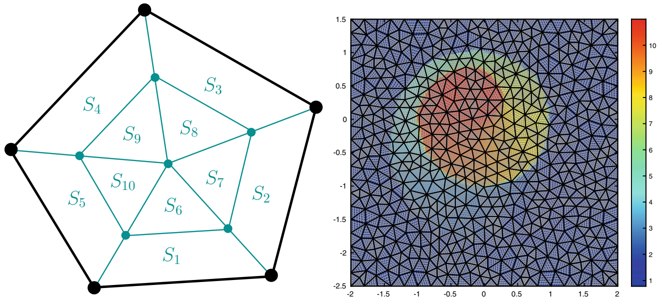
1. *The rank of L is $|\mathcal{S}| - 1$ and its image is $\big(span\{\mathbf{1}\}\big)^\perp$. We still denote the inverse of L on $\big(span\{\mathbf{1}\}\big)^\perp$ by L^{-1},*
2. *With the previous notations, a solution is*

$$\Big(\hat{\mathbf{f}}_{\{\sigma,\sigma'\}}\Big)_{\{\sigma,\sigma'\}\ direct\ edges} = A^T L^{-1}\big(\Psi_\sigma\big)_{\sigma\in\mathcal{S}}. \tag{18}$$

This result has been used to develop in [23] a numerical scheme of dG type, where, when a criteria explaining that some problem occur (negative density, or pressure, or creation of artificial oscillation), the elements are splitted into sub elements where a low order finite volume scheme is applied. An example of sub-cells and application if show in Fig. 2.

A convergence proof (in the statistical solution framework) of the schemes (12a) with (12b) can be found in [10].

The reinterpretation of schemes in term or residuals has other applications. For example in [3], we have considered a mixture of perfect gases. It is well known that because of the possible incompatibility between the numerical dissipations attached to each of the conserved variables, the pressure and the velocity may oscillate across a contact line. There are several solutions to cure that. One is to start from the Euler equation in primitive variables (as (5) but we need to consider two masses), but then



(a) (b)

Fig. 2 **a** Example of a subdivision, **b** Solution obtained with the method of [23] for the KPP problem. These figures have been generated by François Vilar, Université de Montpellier one has to do something very specific to take into account the shocks properly. For example, in [21], a standard finite volume method is used anywhere, but around the slip line which location is estimated by a level set. In [3], we start directly from the non conservative formulation where the variables are ρ, $\mathbf{v}$ and e the internal energy. If Δ represents the time increment, we notice that

$$\Delta E = \Delta e + \frac{\mathbf{v}^{n+1} + \mathbf{v}^n}{2} \Delta\big(\rho \mathbf{v}\big) - \frac{\mathbf{v}^{n+1}\mathbf{v}^n}{2} \Delta \rho. \tag{19}$$

Then, for a scheme of the form (12a), where $\mathbf{u} = (\rho, \rho\mathbf{v}, e)$, it is shown in [3] that if the residual on the energy (and only this one) is modified such that

$$\int_{\partial K} \mathbf{f}^E \cdot \mathbf{n}\, d\gamma = \sum_{\sigma \in K} \Phi_\sigma^e + \sum_{\sigma \in K} \frac{\mathbf{v}_\sigma^{n+1} + \mathbf{v}_\sigma^n}{2} \Phi_\sigma^{\rho\mathbf{v}} + \sum_{\sigma \in K} \frac{\mathbf{v}_\sigma^{n+1}\mathbf{v}_\sigma^n}{2} \Phi_\sigma^\rho, \tag{20}$$

then the scheme will be locally conservative. The modification is done by adding to the initial energy residuals Φ_σ^E the same quantify r^K for all the degrees of freedom in K such that (20) holds true. This formulation can be further refined so that one keeps the good behavior of the contact lines, see [3] for details and applications with very non linear equations of state.

2 Staggering

In another application, one considers an approximation of (5) where the velocity is globally continuous and the thermodynamic parameters (the density, the pressure, the internal energy) are only in L^2. This is an Eulerian version of [9, 13] which are inspired from [16] which is a finite element generalisation of the Wilkins scheme [24]. The Wilkins scheme is very popular in the Lagrange hydrodynamics community because use the specific internal energy as a variable though producing the correct weak solutions of the problem.

Here we summarise [7]. We start from a scheme having the from (12a). The velocity is approximated by a piecewise polynomial function or degree $r \geq 1$ that is globally continuous. For technical reasons, we assume that we use a Bernstein basis (each of the basis function is positive). The pressure, internal energy and the density are approximated by polynomials of degree $r - 1$, and again we use Bernstein basis functions. The thermodynamical degrees of freedoms are denoted by σ_T and the velocity degrees of freedom are σ_V. The density is in L^2, so we use a discontinuous Galerkin approximation, with a Riemann solver (that is used only for the density). Then, for each degree of freedom, we discretize the velocity equation in the form

$$\mathbf{v}_\sigma^{n+1} = \mathbf{v}_\sigma^n - \frac{\Delta t}{|C_\sigma|} \sum_{K, \sigma \in K} \Phi_\sigma^K.$$

Here, $|C_\sigma|$ is the mass of φ_σ which is positive. We take

$$\Phi_\sigma^{\mathbf{v},K} = \int_K \varphi_\sigma \Big((\mathbf{v} \cdot \nabla)\mathbf{v} + \frac{\nabla p}{\rho} \Big)\, d\mathbf{x}$$

computed by numerical quadrature, and

$$\Phi_\sigma^{e,K} = \int_K \varphi_\sigma \Big(\mathbf{v} \cdot \nabla e + (e+p)\,\mathrm{div}\,\mathbf{u} \Big)\, d\mathbf{x},$$

again by numerical quadrature. As such, there is no hope to obtain a good scheme.

In order to modify it, we start from an inspection of what would be needed to get a Lax-Wendroff like theorem. It turns out that for the update in time of the momentum, we have the relation:

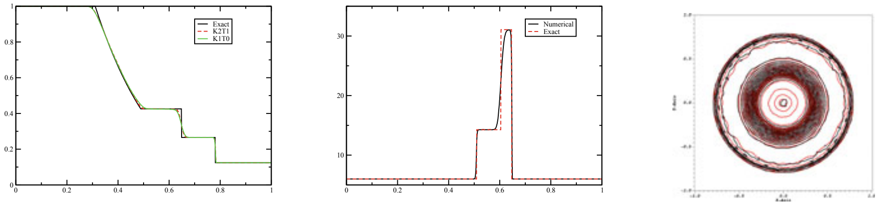
$$\rho^{n+1}\mathbf{v}^{n+1} - \rho^n \mathbf{v}^n = (\rho^{n+1} - \rho^n)\mathbf{v}^n + \rho^{n+1}(\mathbf{v}^{n1} - \mathbf{v}^n),$$

from which we can infer, after some calculations, that if the mass and velocity residual satisfy, in each element,

$$\int_{\partial K} \rho^n \mathbf{v}^n \cdot \mathbf{n} = \sum_{\sigma_V \in K} \theta_{\sigma_V}^{\mathbf{v}} \Phi_{\sigma_V}^{\mathbf{v}} + \sum_{\sigma_T \in K} \mu_{\sigma_T}^{\mathbf{v}} \Phi_{\sigma_T}^{\rho} \tag{21}$$

where the coefficients $\theta_{\sigma_V}^{\mathbf{v}}$ and μ_{σ_T} are explicitly computable, see [7]. In addition $\theta_{\sigma_V}^{\mathbf{v}} > 0$. Of course the initial scheme does not satisfy this constraint, but keeping the density at t_n and t_{n+1} unchanged, we can modify $\Phi_{\sigma_V}^{\mathbf{v}}$ by $\Phi_{\sigma_V}^{\mathbf{v}} + r^K$, so that (21) is true.

Similar algebraic manipulation can be done for the energy, and here we need to mimick (19), and this is always possible. We illustrate by some example taken from [7] in Fig. 3.



(a) (b) (c)

Fig. 3 **a** Density for the Sod test case. We observed the solution without correction (red) ans the numerical one with 1000 points, it superimpose the exact solution. **b** Exact and numerical solutions of the Collela and Woodwards test case for (**a**) density at time $T = 0.012$ with $CFL = 0.1$ computed on a mesh with 1000 points. **c** Comparison of the solutions of the 2D shock test case for the pressure obtained by a conservative scheme (red) and the staggered one (black) at time $T = 0.16$

3 Additional Conservation Laws

It is also possible to use this framework in order to take into account additional conservation laws, at least in the semi-discrete sense. This part is a short summary of [1] and then [4, 5, 11, 12].

Let us consider the entropy, for example. Multiplying (1a) by $\mathbf{v} := \nabla_{\mathbf{u}}\eta$ we obtain (2). Hence, considering a scheme of the form

$$|C_\sigma|\frac{d\mathbf{u}_\sigma}{dt} + \sum_{K,\sigma\in K} \Phi_\sigma^K = 0,$$

we would have

$$|C_\sigma|\frac{d\eta_\sigma}{dt} + \sum_{K,\sigma\in K} \mathbf{v}_\sigma \cdot \Phi_\sigma^K = 0.$$

This gives the idea of introducing the entropy residuals,

$$\Psi_\sigma^K = \mathbf{v}_\sigma \cdot \Phi_\sigma^K.$$

However, there is no reason why if

$$\sum_{\sigma\in K} \Phi_\sigma^K = \int_{\partial K} \hat{\mathbf{f}}_{\mathbf{n}}\, d\gamma$$

we would have $\sum_{\sigma\in K} \Psi_\sigma^K$ related to some boundary integral of the entropy flux.

The trick is similar as before, we introduce $\mathbf{r}_\sigma^K$ and consider $\widetilde{\Phi_\sigma^K} = \Phi_\sigma^K + \mathbf{r}_\sigma^K$ such that we sill have the conservation relation and have in addition

$$\sum_{\sigma\in K} \Psi_\sigma^K \geq \int_{\partial K} \hat{\mathbf{g}}_{\mathbf{n}}\, d\gamma$$

where $\hat{\mathbf{g}}_{\mathbf{n}}$ is some consistant numerical approximation of the entropy flux. The conservation requirement implies that

$$\sum_{\sigma\in K} \mathbf{r}_\sigma^K = 0.$$

This condition can be met if

$$\mathbf{r}_\sigma^K = \alpha_K\big(\mathbf{v}_\sigma - \overline{\mathbf{v}}), \quad \overline{\mathbf{v}} = \frac{1}{\#K}\sum_{\sigma\in K} \mathbf{v}_\sigma,$$

and we get

$$\alpha_K \sum_{\sigma\in K} \big(\mathbf{v}_\sigma - \overline{\mathbf{v}}\big)^2 \geq \int_{\partial K} \hat{\mathbf{g}}_{\mathbf{n}} \, d\gamma - \sum_{\sigma\in K} \mathbf{v}_\sigma \cdot \Phi_\sigma^K .$$

In [1] there is a discussion on the choice of $\hat{\mathbf{f}}$ such that

$$\int_{\partial K} \hat{\mathbf{g}}_{\mathbf{n}} \, d\gamma - \sum_{\sigma\in K} \mathbf{v}_\sigma \cdot \Phi_\sigma^K = O\big(\mathbf{v} - \overline{\mathbf{v}}\big)^2$$

to guaranty that α_K does not blow up. The discussion is certainly not finished.

In [12], it is shown how to extend this approach when we have several constraints: in that paper, we had discussed the case of the entropy and the kinetic energy preservation. Instead of a simple linear equation, one gets a system of size the number of constraints, and it can be solved by least square. Interestingly, the more degrees of freedom (i.e. the higher the formal order is), the more constraint one can a priori satisfy. We are not able to prove that if one starts with a stable scheme, the modified scheme will also be stable. In practice, we have never observed any instability.

In [4], using this technique, we have shown that a fully centered scheme can be made stable! There is no contradiction with the classical analysis that uses periodicity. Of course this cannot apply to periodic problems, but to problems with proper boundaries. The idea is, for a transport problem, to write a scheme. In [4], the examples were considering triangular type meshes, the unknown was approximated with $\mathbb{P}^k$ spatial approximation, and the scheme was a simple weak formulation with accurate enough quadrature formula. In order to have an energy bound, we modify the scheme, so that on each element, we have an exact energy production. This is done following the above ideas. When summing all the contributions, the interior of the domain does not produce energy, and all has to be controlled at the boundary. In fact, up to this energy summation argument that can be made possible here at the full discrete level, this very classical: it is well know that a dG method is L^2 stable if one has a dissipative boundary flux. Hence we are considering a dG method with one element (the whole domain), nobody has ever said that the approximation must be polynomial: it simply need to be accessible to some form of the divergence theorem. All is made in purpose here for that. This also applies to non linear problems, see [5].

This technique can be further used. In [11], a similar approach is used so that a local conservation equation for the kinetic momentum is also satisfied. In [8], it has been used to obtain a thermodynamically compatible scheme for a system that can simultaneously describe a fluid and a solid. There is no space here to describe the method, we refer to the publication. Let us only mention that the variables contains the entropy, not the total energy, and we nevertheless have a method that is locally conservative.

4 A Scheme that does not Fit in this Framework

In [2, 14], and following an idea of Roe [17], we consider a grid with points $x_i < x_{i+1}$. The conserved variable $\mathbf{u}$ is approximated by its average in each volume $[x_i, x_{i+1}]$, and we also consider the *point* values of some other set of variables denoted by $\mathbf{v}_i$ (for the point x_i, $\mathbf{v}$ can be $\mathbf{u}$, but also any transformation of $\mathbf{u}$ by some regular mapping Ψ) and it satisfies the evolution equation

$$\frac{\partial \mathbf{v}}{\partial t} + \underbrace{(\nabla_{\mathbf{u}}\psi)^{-1}\nabla_{\mathbf{u}}\mathbf{f}\nabla_{\mathbf{u}}\psi}_{J} \frac{\partial \mathbf{v}}{\partial x} = 0.$$

For fluid mechanics, $\mathbf{v}$ can be the primitive variables, for example. In each cell $[x_i, x_{i+1}]$ one has $\overline{\mathbf{u}}_{i+1/2}$ and the point values $\mathbf{v}_i = \Psi(\mathbf{u}_i)$, $\mathbf{v}_{i+1} = \Psi(\mathbf{u}_{i+1})$. From this one can construct a globally continuous reconstruction which is quadratic in each cell. It amounts to Simpson's formula:

$$\overline{\mathbf{u}}_{i+1/2} = \frac{1}{4}\big(\psi^{-1}(\mathbf{v}_i) + \psi^{-1}(\mathbf{v}_{i+1/2}) + \psi^{-1}(\mathbf{v}_{i+1})\big)$$

so that one can obtain a third order approximation of $\mathbf{v}_{i+1/2} \approx \mathbf{v}(\frac{x_i+x_{i+1}}{2})$.

The averaged values and point value are evolved by

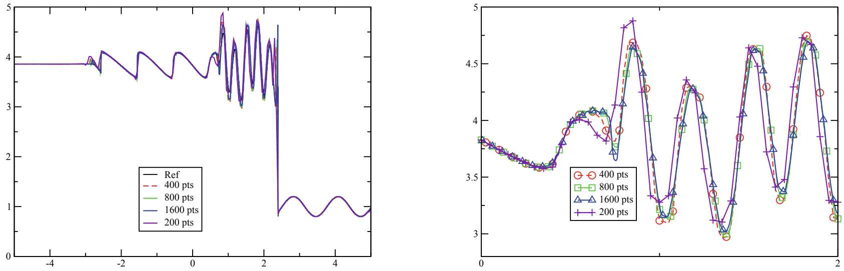
$$(x_{i+1} - x_i)\frac{d\overline{\mathbf{u}}_{i+1/2}}{dt} + \mathbf{f}(\mathbf{u}_{i+1}) - \mathbf{f}(\mathbf{u}_i) = 0, \qquad \frac{d\mathbf{v}_i}{dt} + \Phi_i^{[x_i,x_{i+1}]} + \Phi_i^{[x_{i-1},x_i]} = 0$$

where $\Phi_i^{[x_i,x_{i+1}]}$ (resp. $\Phi_i^{[x_{i-1},x_i]}$) is a consistant approximation of $J^-(\mathbf{u}_i)\frac{\partial \mathbf{v}}{\partial x}$ (resp. $J^+(\mathbf{u}_i)\frac{\partial \mathbf{v}}{\partial x}$). To define these approximations, we use the data $\mathbf{v}_l$ and $\mathbf{v}_{l+1/2}$ and the approximations contain in [20]. In [2] we show that under assumptions similar to the standard Lax Wendroff theorem, that this scheme will provide a sequence that will converge to a weak solution. It is unclear if this scheme can be written in flux form, simply because the volumes associated to the average already cover the whole computational domain.

One illustration, see Fig. 4 is given by the Shu Osher problem where the initial condition are:

$$(\rho, u, p) = \begin{cases} (3.857143, 2.629369, 10.3333333) & \text{if } x < -4 \\ (1 + 0.2\sin(5x), 0, 1) & \text{else} \end{cases}$$

on the domain $[-5, 5]$ until $T = 1.8$.



(a)

(b)

Fig. 4 **a** Solution of the Shu Osher problem, **b** zoom of the solution around the shock

Acknowledgements I take this opportunity to thanks my many collaborators and students: M. Ricchiuto (Inria), P.H. Maire (Cea), F. Vilar (Montpellier), R. Loubère (Bordeaux), Ph. Öffner (Mainz), W. Barsukow (Bordeaux), M. Dumbser (Trento), S. Busto (Vigò), S. Tokareva (Los Alamos), P. Baccigaluppi (Milano), L. Micalizzi (Zürich), A. Larat (Grenoble), M. Mezine (Numeca), M. Lukakova (Mainz). Without them nothing would have been possible.

References

1. Abgrall, R.: A general framework to construct schemes satisfying additional conservation relations. Application to entropy conservative and entropy dissipative schemes. J. Comput. Phys. **372**, 640–666 (2018)
2. Abgrall, R.: A combination of residual distribution and the active flux formulations or a new class of schemes that can combine several writings of the same hyperbolic problem: application to the 1d Euler equations. Commun. Appl. Math. Comput. **5**, 370–402 (2023)
3. Abgrall, R., Bacigaluppi, P., Tokareva, S.: A high-order nonconservative approach for hyperbolic equations in fluid dynamics. Comput. Fluids **169**, 10–22 (2018)
4. Abgrall, R., Nordström, J., Öffner, P., Tokareva, S.: Analysis of the SBP-SAT stabilization for finite element methods. I: Linear problems. J. Sci. Comput. **85**(2), 28 (2020). Id/No 43
5. Abgrall, R., Nordström, J., Öffner, P., Tokareva, S.: Analysis of the SBP-SAT stabilization for finite element methods part II: entropy stability. Commun. Appl. Math. Comput. (2021)
6. Abgrall, R.: Some remarks about conservation for residual distribution schemes. Comput. Methods Appl. Math. **18**(3), 327–351 (2018)
7. Abgrall, R.: Staggered residual distribution scheme for compressible flow (2021). in revision
8. Abgrall, R., Busto, S., Dumbser, M.: A simple and general framework for the construction of thermodynamically compatible schemes for computational fluid and solid mechanics. Appl. Math. Comput. **440**, 40 (2023). Id/No 127629
9. Abgrall, R., Lipnikov, K., Morgan, N., Tokareva, S.: Multidimensional staggered grid residual distribution scheme for Lagrangian hydrodynamics. SIAM J. Sci. Comput. **42**(1), a343–a370 (2020)
10. Abgrall, R., Luckacova-Medvid'ova, M., Oeffner, P.: On the convergence of residual distribution schemes for the compressible Euler equations via dissipative weak solutions. M3AS **33**(1), 139–173 (2023). arXiv:2207.11969
11. Abgrall, R., Mojarrad, F.N.: Conservative scheme compatible with some other conservation laws: conservation of the local angular momentum. Comput. Fluids **247**, 15 (2022). Id/No 105663

12. Abgrall, R., Öffner, P., Ranocha, H.: Reinterpretation and extension of entropy correction terms for residual distribution and discontinuous Galerkin schemes: application to structure preserving discretization. J. Comput. Phys. **453**, 24 (2022). Id/No 110955
13. Abgrall, R., Tokareva, S.: Staggered grid residual distribution scheme for Lagrangian hydrodynamics. SIAM J. Sci. Comput. **39**(5), a2317–a2344 (2017)
14. Barsukow, W., Abgrall, R.: Extensions of active flux to arbitrary order of accuracy. ESAIM: M2AN. in press. https://doi.org/10.1051/m2an/2023004
15. Dafermos, C.M.: Hyperbolic Conservation Laws in Continuum Physics, 4th edn. Grundlehren der Mathematischen Wissenschaften, vol. 325. Springer, Berlin (2016)
16. Dobrev, V.A., Kolev, T.V., Rieben, R.N.: High-order curvilinear finite element methods for Lagrangian hydrodynamics. SIAM J. Sci. Comput. **34**(5), b606–b641 (2012)
17. Eyman, T.A., Roe, P.L.: Active flux. In: 49th AIAA Aerospace Science Meeting (2011)
18. Godlewski, E., Raviart, P.-A.: Numerical Approximation of Hyperbolic Systems of Conservation Laws, 2nd edn. Applied Mathematical Sciences, vol. 118. Springer, New York, NY (2021)
19. Hou, T.Y., Le Floch, P.G.: Why nonconservative schemes converge to wrong solutions: error analysis. Math. Comput. **62**(206), 497–530 (1994)
20. Iserles, A.: Order stars and saturation theorem for first-order hyperbolics. IMA J. Numer. Anal. **2**, 49–61 (1982)
21. Karni, S.: Hybrid multifluid algorithms. SIAM J. Sci. Comput. **17**(5), 1019–1039 (1996)
22. Maire, P.-H.: A high-order cell-centered Lagrangian scheme for compressible fluid flows in two-dimensional cylindrical geometry. J. Comput. Phys. **228**(18), 6882–6915 (2009)
23. Vilar, F., Abgrall, R.: A Posteriori local subcell correction of high-order discontinuous galerkin scheme for conservation laws on two-dimensional unstructured grids (2022). submitted
24. Wilkins, M.L.: Methods in Computational Physics, vol. 3. Academic Press, New York (1964)

A High Order Semi-implicit Scheme for Ideal Magnetohydrodynamics

Claudius Birke, Walter Boscheri, and Christian Klingenberg

Abstract In this work we design a novel semi-implicit finite volume solver for the equations of ideal magnetohydrodynamics (MHD). The nonlinear convective terms as well as the time evolution of the magnetic field are discretized explicitly, while the terms related to the hydrodynamic pressure in the momentum and in the energy equation are solved implicitly, hence making the scheme particularly well suited for the simulation of low Mach number flows. An elliptic equation is then obtained for the pressure, and the associated system is linearized in time relying on a semi-implicit discretization of the kinetic energy and the enthalpy. High order of accuracy in time is achieved using implicit-explicit Runge-Kutta (IMEX-RK) methods, whereas an efficient CWENO reconstruction permits to gain high accuracy also in space. The solenoidal property of the magnetic field is respected at the discrete level relying on a high order constrained transport method, leading to a structure preserving scheme. The new scheme is conservative for mass, momentum and total energy, and both finite volume and central finite difference discretizations are adopted for the explicit and the implicit terms, respectively, hence introducing no numerical dissipation in the terms related to the pressure. We validate the new schemes against benchmarks for ideal MHD, showing the accuracy and the robustness of the novel methods even in the case of shock waves.

Keywords Semi-implicit · Divergence-free · Compressible low mach number flows · Finite volume schemes · Pressure-based method

W. Boscheri (✉)
Department of Mathematics and Computer Science, University of Ferrara, Ferrara, Italy
e-mail: walter.boscheri@unife.it

C. Birke · C. Klingenberg
Department of Mathematics, University of Würzburg, Würzburg, Germany

E. Franck et al. (eds.), *Finite Volumes for Complex Applications X—Volume 1, Elliptic and Parabolic Problems*, Springer Proceedings in Mathematics & Statistics 432,
https://doi.org/10.1007/978-3-031-40864-9_2

1 Introduction

Magnetized plasma flows are governed by the equations of magnetohydrodynamics (MHD), that describe the time evolution of electric conducting fluids embedded in a magnetic field. The simplest model is given by ideal MHD, where the fluid viscosity is neglected, which constitutes a system of nonlinear hyperbolic partial differential equations (PDE) involving conservation of mass, momentum and total energy coupled with Faraday law for the magnetic field.

The sonic Mach number, which is the ratio between the fluid velocity and the sound speed, describes the regime of the fluid under consideration. The low Mach regime typically arises in astrophysical phenomena such as the generation of magnetic fields in deep convective layers of stars. From the numerical viewpoint, compressible flows are typically discretized using explicit Godunov-type finite volume schemes since they are by construction conservative and thus allow the correct computation of shock waves. However, in the low Mach limit, the effect of numerical viscosity which is added to the numerical fluxes is proven to degrade the accuracy [2, 16]. Furthermore, in the incompressible regime the elliptic behavior of the pressure introduces a very severe restriction on the maximum admissible time step for low Mach number flows, making explicit schemes very ineffective. A possible remedy would be the adoption of fully implicit methods, which inevitably imply the solution of large nonlinear systems that are computationally very expensive and in which the convergence is numerically very difficult to control. Consequently, the MHD system in the low Mach limit has been widely investigated [13, 15, 17, 19, 20, 22, 26]. A successful idea consists in treating implicitly only one part of the system to be solved while keeping the remaining explicit. In this way, the implicit part is relatively simple to be inverted, whereas the nonlinear terms undergo an explicit discretization, making the resulting method capable of dealing with all Mach regimes. This idea has been originally conceived in the context of shallow water and incompressible flows [11, 12], where a semi-implicit time stepping technique has been used. Implicit-explicit (IMEX) schemes have been designed [1, 4–6, 23, 24] in order to deal with multi-scale phenomena, that are typically encountered in compressible fluids.

In this work we propose a novel pressure-based scheme for the solution of the ideal MHD equations. The time discretization is inspired by the class of semi-implicit IMEX schemes [3, 9], and here we treat implicitly the terms related to the pressure, hence not introducing any numerical dissipation and making the CFL stability condition independent from the acoustic wave speed. Differently from [17, 18], no nonlinear equations are used in our approach. Furthermore, the semi-implicit linearization is also used for the kinetic energy, contrarily to what has been proposed in [13].

2 Governing Equations

Let us consider a one-dimensional computational domain $\Omega \in \mathbb{R}$ defined by the spatial coordinate $x \in \Omega$, and let the time coordinate be denoted with $t \in \mathbb{R}_0^+$. The ideal equations of magnetohydrodynamics (MHD) in one space dimension constitute a hyperbolic system of the form

$$\frac{\partial \mathbf{q}}{\partial t} + \frac{\partial \mathbf{f}}{\partial x} = \mathbf{0}, \tag{1}$$

with the vector of state variables $\mathbf{q}$ and the fluxes $\mathbf{f}(\mathbf{q})$ that explicitly write

$$\mathbf{q} = \begin{pmatrix} \rho \\ \rho u \\ \rho v \\ \rho w \\ \rho E \\ B_x \\ B_y \\ B_z \end{pmatrix}, \qquad \mathbf{f}(\mathbf{q}) = \begin{pmatrix} \rho u \\ \rho u^2 + p + \frac{1}{8\pi}\mathbf{B}^2 - \frac{1}{4\pi}B_x^2 \\ \rho u v - \frac{1}{4\pi}B_x B_y \\ \rho u w - \frac{1}{4\pi}B_x B_z \\ u(\rho E + p + \frac{1}{8\pi}\mathbf{B}^2) - \frac{1}{4\pi}B_x(\mathbf{v}\cdot\mathbf{B}) \\ 0 \\ uB_y - vB_x \\ uB_z - wB_x \end{pmatrix}. \tag{2}$$

The fluid density and pressure are addressed with ρ and p, respectively, while $\mathbf{v} = (u, v, w)$ is the velocity field and the magnetic field is denoted with $\mathbf{B} = (B_x, B_y, B_z)$. The total energy ρE is obtained as the sum of three contributions, namely

$$\rho E = \rho e + \rho k + m, \qquad \rho e = \frac{p}{\gamma - 1}, \quad \rho k = \frac{1}{2}\rho \mathbf{v}^2, \quad m = \frac{1}{8\pi}\mathbf{B}^2, \tag{3}$$

where ρk is the kinetic energy and m is the magnetic energy. The internal energy ρe is computed relying on the ideal gas equation of state (EOS) with $\gamma = c_p/c_v$ denoting the ratio of specific heats at constant pressure and volume, respectively. By introducing the specific enthalpy $h = e + p/\rho$, one can reformulate the first part of the energy flux in (2) such that

$$u(\rho E + p + m) = u(\rho k + m) + h(\rho u). \tag{4}$$

The MHD system (2) is hyperbolic since the eigenvalues $\lambda_{i=\{1,\ldots,8\}}^{MHD}$ of the associated Jacobian matrix $\mathbf{A} = \partial \mathbf{f}/\partial \mathbf{q}$ with $B_x = const$ are

$$\lambda_{1,8}^{MHD} = u \pm c_f, \quad \lambda_{2,7}^{MHD} = u \pm c_a, \quad \lambda_{3,6}^{MHD} = u \pm c_s, \quad \lambda_4^{MHD} = u, \quad \lambda_5^{MHD} = 0, \tag{5}$$

with the wave speeds given by

$$\begin{aligned}
c_a &= \frac{B_x}{\sqrt{4\pi\rho}},\\
c_s^2 &= \frac{1}{2}\left(b^2 + c^2 - \sqrt{(b+c)^2 - 4c_a^2c^2}\right),\\
c_f^2 &= \frac{1}{2}\left(b^2 + c^2 + \sqrt{(b+c)^2 - 4c_a^2c^2}\right).
\end{aligned} \tag{6}$$

The Alfvén wave speed is c_a, the speeds of slow and fast magnetosonic waves are c_s and c_f, respectively, while $c^2 = \gamma p/\rho$ is the adiabatic sound speed that is computed from the ideal equation of state. Furthermore, we use the abbreviation $b^2 = \mathbf{B}^2/(4\pi\rho)$.

The hydrodynamic behavior of the fluid can be analyzed by considering the Mach number $M = u/c$. In the low Mach number limit, the sound speed becomes very high compared to the fluid velocity, hence the terms related to the pressure are dominant. Consequently, larger values of the fast and slow magnetosonic wave speeds are retrieved, and fully explicit numerical methods suffer from both an excessive amount of numerical viscosity, which is proportional to the eigenvalues, and a drastic reduction of the admissible time step Δt to ensure stability under a classical CFL condition of the type

$$\Delta t \leq \mathrm{CFL} \min_{\Omega} \frac{\max|\lambda^{MHD}|}{\Delta x}, \tag{7}$$

with Δx denoting the characteristic mesh spacing and the CFL ≤ 1 being the CFL number. Therefore, we propose to discretize implicitly the pressure gradient in the momentum equation and the enthalpy term in the energy equation, while keeping an explicit discretization for the nonlinear convective fluxes and the terms related to the magnetic field. To that aim, let the fluxes be split into a convective-type flux $\mathbf{f}^c(\mathbf{q})$ and a pressure-type flux $\mathbf{f}^p(\mathbf{q})$, that is

$$\mathbf{f}^c(\mathbf{q}) = \begin{pmatrix} \rho u \\ \rho u^2 + m - \frac{1}{4\pi}B_x^2 \\ \rho uv - \frac{1}{4\pi}B_xB_y \\ \rho uw - \frac{1}{4\pi}B_xB_z \\ u(\rho k + m) - \frac{1}{4\pi}B_x(\mathbf{v}\cdot\mathbf{B}) \\ 0 \\ uB_y - vB_x \\ uB_z - wB_x \end{pmatrix}, \qquad \mathbf{f}^p(\mathbf{q}) = \begin{pmatrix} 0 \\ p \\ 0 \\ 0 \\ h\rho u \\ 0 \\ 0 \\ 0 \end{pmatrix}. \tag{8}$$

We obtain two sub-systems with the following eigenvalues [17].

- Convective sub-system:

$$\frac{\partial \mathbf{q}}{\partial t} + \frac{\partial \mathbf{f}^c}{\partial x} = \mathbf{0}, \tag{9a}$$

$$\lambda^c_{1,8} = u \pm \sqrt{\frac{\mathbf{B}^2}{4\pi\rho}}, \quad \lambda^c_{2,7} = u \pm \frac{B_x}{\sqrt{4\pi\rho}}, \quad \lambda^c_{3,4,5,6} = 0. \tag{9b}$$

- Pressure sub-system:

$$\frac{\partial \mathbf{q}}{\partial t} + \frac{\partial \mathbf{f}^p}{\partial x} = \mathbf{0}, \tag{10a}$$

$$\lambda^p_1 = \frac{1}{2}\left(u - \sqrt{u^2 + 4c^2}\right), \quad \lambda^p_{2,3,4,5,6,7} = 0, \quad \lambda^p_8 = \frac{1}{2}\left(u + \sqrt{u^2 + 4c^2}\right). \tag{10b}$$

It is clear that, by taking the pressure sub-system implicitly, the maximum admissible time step of the scheme becomes

$$\Delta t \leq \text{CFL} \min_{\Omega} \frac{\max |\lambda^c|}{\Delta x}, \tag{11}$$

hence making the scheme particularly well suited for low Mach number flows ($M \ll 1$) where the pressure terms are dominant. On the other hand, for strongly convected flows with shocks, the convective eigenvalues lead the computation of the time step granting stability.

3 Numerical Method

The computational domain $\Omega = [x_L; x_R]$ is discretized using a total number of N_x equidistant cells of volume $\Delta x = (x_R - x_L)/N_x$. The cell centers are indicated with x_i and the cell interfaces are referred to with $x_{i+1/2}$. The time coordinate is bounded in the interval $t \in [0; t_f]$, and the final time t_f is reached performing a sequence of time steps $\Delta t = t^{n+1} - t^n$ that are computed according to the CFL stability condition (11).

3.1 *First Order Semi-discrete Scheme in Time*

Using the flux splitting (8), it is possible to design the following semi-discrete scheme for the explicit sub-system:

$$\rho^* = \rho^n - \Delta t \frac{\partial}{\partial x} (\rho u)^n , \tag{12a}$$

$$(\rho u)^* = (\rho u)^n - \Delta t \frac{\partial}{\partial x} \left(\rho u^2 + m - \frac{1}{4\pi} B_x^2 \right)^n , \tag{12b}$$

$$(\rho v)^* = (\rho v)^n - \Delta t \frac{\partial}{\partial x} \left(\rho u v - \frac{1}{4\pi} B_x B_y \right)^n , \tag{12c}$$

$$(\rho w)^* = (\rho w)^n - \Delta t \frac{\partial}{\partial x} \left(\rho u w - \frac{1}{4\pi} B_x B_z \right)^n , \tag{12d}$$

$$(\rho E)^* = (\rho E)^n - \Delta t \frac{\partial}{\partial x} \left(u(\rho k + m) - \frac{1}{4\pi} B_x (\mathbf{v} \cdot \mathbf{B}) \right)^n , \tag{12e}$$

$$B_x^* = 0, \tag{12f}$$

$$B_y^* = B_y^n - \Delta t \frac{\partial}{\partial x} \left(u B_y - v B_x \right)^n , \tag{12g}$$

$$B_z^* = B_z^n - \Delta t \frac{\partial}{\partial x} \left(u B_z - w B_x \right)^n . \tag{12h}$$

The above definitions are then employed to obtain a first order semi-implicit time discretization [3, 9, 10] of the MHD equations (2), which writes

$$\rho^{n+1} = \rho^*, \tag{13a}$$

$$(\rho u)^{n+1} = (\rho u)^* - \Delta t \frac{\partial}{\partial x} \left(p^{n+1} \right), \tag{13b}$$

$$(\rho v)^{n+1} = (\rho v)^*, \tag{13c}$$

$$(\rho w)^{n+1} = (\rho w)^*, \tag{13d}$$

$$(\rho e)^{n+1} + \frac{(\rho u)^{n+1} (\rho u)^n}{2\rho^{n+1}} + m^{n+1} = (\rho E)^* - \Delta t \frac{\partial}{\partial x} \left(h^n (\rho u)^{n+1} \right), \tag{13e}$$

$$B_x^{n+1} = B_x^*, \tag{13f}$$

$$B_y^{n+1} = B_y^*, \tag{13g}$$

$$B_z^{n+1} = B_z^*. \tag{13h}$$

To avoid nonlinear implicit terms, let us notice that the implicit flux in the energy equation has been discretized by taking the enthalpy explicitly, and the kinetic energy in the total energy definition splits into an explicit and an implicit contribution:

$$(\rho E)^{n+1} := (\rho e)^{n+1} + \frac{(\rho u)^{n+1} (\rho u)^n}{2\rho^{n+1}} + m^{n+1}, \tag{14}$$

following the approach presented in [9] for the hydrodynamics equations. Recall that the internal energy can be expressed in terms of the pressure relying on the ideal gas EOS (3), and that the new magnetic energy $m^{n+1} = (\mathbf{B}^{n+1})^2/(8\pi)$ can be explicitly computed because the fluxes of the magnetic field belong to the explicit sub-system

(9). Moreover, the time evolution of the density is also concerned with an explicit update, hence making ρ^{n+1} already known from (12a). Therefore, a preliminary discretization of the total energy equation is chosen by inserting the momentum equation (13b) into the energy equation (13e) leading to an elliptic equation for the pressure:

$$\frac{p^{n+1}}{\gamma-1} - \Delta t \frac{(\rho u)^n}{2\rho^{n+1}} \frac{\partial}{\partial x}\left(p^{n+1}\right) - \Delta t^2 \frac{\partial}{\partial x}\left(h^n \frac{\partial}{\partial x}\left(p^{n+1}\right)\right) = b^n, \tag{15}$$

with the known right-hand-side given by

$$b^n = (\rho E)^* - \frac{(\rho u)^n}{2\rho^{n+1}}(\rho u)^* - m^{n+1} - \Delta t \frac{\partial}{\partial x}\left(h^n (\rho u)^*\right). \tag{16}$$

The pressure equation (15) constitutes a linear system for the scalar unknown p^{n+1} that is solved using the iterative GMRES solver [25] up to a prescribed tolerance (we typically set tol $= 10^{-12}$). Differently from [17, 18], this approach does not need any fixed point method thanks to the semi-implicit splitting of the enthalpy flux and the kinetic energy in the energy equation. Once the new pressure is known, the new momentum $(\rho u)^{n+1}$ is updated with (13b), and then the new total energy is updated using the conservative formulation

$$(\rho E)^{n+1} = (\rho E)^* - \Delta t \frac{\partial}{\partial x}\left(h^n (\rho u)^{n+1}\right). \tag{17}$$

3.2 *First Order Discrete Spatial Operators*

The spatial operators are given by both finite volume and finite difference approximations, and they are introduced hereafter referring to the state vector $\mathbf{q}(x, t)$.

The convective sub-system (9) is discretized with a conservative Godunov-type finite volume method, that is

$$\mathbf{q}_i^* = \mathbf{q}_i^n - \frac{\Delta t}{\Delta x}\left(\mathbf{f}_{i+1/2}^c - \mathbf{f}_{i-1/2}^c\right). \tag{18}$$

We choose to use the simple Rusanov-type numerical flux $\mathbf{f}_{i+1/2}^c$ that is given by

$$\mathbf{f}_{i+1/2}^c = \frac{1}{2}\left(\mathbf{f}^c(\mathbf{q}_{i+1}) + \mathbf{f}^c(\mathbf{q}_i)\right) - \frac{1}{2} s_{\max}\left(\mathbf{q}_{i+1} - \mathbf{q}_i\right), \tag{19}$$

where the numerical dissipation $s_{\max} = \max\left(|\lambda_{i+1}^c|, |\lambda_i^c|\right)$ only accounts for the convective eigenvalues, thus no acoustic speed is involved.

The implicit terms appearing in the pressure sub-system (10) are approximated by means of finite difference operators with no numerical dissipation, thus one has

$$\left.\frac{\partial \mathbf{q}}{\partial x}\right|_i^{n+1} = \frac{\mathbf{q}_{i+1}^{n+1} - \mathbf{q}_{i-1}^{n+1}}{2\,\Delta x} + \mathcal{O}(\Delta x^2), \tag{20a}$$

$$\left.\frac{\partial}{\partial x}\left(h\frac{\partial \mathbf{q}}{\partial x}\right)\right|_i^{n,n+1} = \frac{1}{\Delta x^2}\left[h_{i-1}^n\; h_i^n\; h_{i+1}^n\right]\begin{bmatrix} 3/4 & -1 & 1/4 \\ 0 & 0 & 0 \\ 1/4 & -1 & 3/4 \end{bmatrix}\begin{bmatrix} \mathbf{q}_{i-1}^{n+1} \\ \mathbf{q}_i^{n+1} \\ \mathbf{q}_{i+1}^{n+1} \end{bmatrix} + \mathcal{O}(\Delta x^2). \tag{20b}$$

3.3 Extension to High Order of Accuracy

The semi-discrete first order scheme (13) supplemented with the spatial operators (18)–(20) is extended to high order of accuracy in space and time by means of semi-implicit IMEX Runge-Kutta methods [3] and quadrature-free CWENO reconstructions [9], respectively.

High order in time. The governing equations are written under the form of an autonomous system with initial condition $\mathbf{q}(t_0) = \mathbf{q}_0$:

$$\frac{\partial \mathbf{q}}{\partial t} = \mathcal{H}(\mathbf{q}_E(t), \mathbf{q}_I(t)), \tag{21}$$

where the spatial fluxes are contained in the flux term $\mathcal{H}(\mathbf{q}_E(t), \mathbf{q}_I(t))$ according to (8), hence involving both explicit and implicit terms, namely $\mathbf{q}_E(t)$ and $\mathbf{q}_I(t)$, respectively. Implicit-explicit (IMEX) Runge-Kutta schemes [24] are then used to advance the solution in time of system (21), following a method of lines (MOL) philosophy. After having set $\mathbf{q}_E = \mathbf{q}_I = \mathbf{q}^n$, the stage fluxes for $r = 1, \ldots, s$ are computed in the following way:

$$\mathbf{q}_E^r = \mathbf{q}_E^n + \Delta t \sum_{\ell=1}^{r-1} \tilde{a}_{r\ell} k_\ell, \qquad 2 \le r \le s, \tag{22a}$$

$$\tilde{\mathbf{q}}_I^r = \mathbf{q}_E^n + \Delta t \sum_{\ell=1}^{r-1} a_{r\ell} k_\ell, \qquad 2 \le r \le s, \tag{22b}$$

$$k_r = \mathcal{H}\left(\mathbf{q}_E^r, \tilde{\mathbf{q}}_I^r + \Delta t\, a_{rr}\, k_r\right). \qquad 1 \le r \le s. \tag{22c}$$

The coefficients $\tilde{a}_{r\ell}$ and $a_{r\ell}$ refer to the explicit and the implicit Runge-Kutta scheme, respectively, and they are collected in a double Butcher tableau. We employ stiffly accurate schemes [3], therefore the solution at the new time level is simply given by the solution of the last stage of the RK time stepping, that is $\mathbf{q}_E^s = \mathbf{q}_I^s = \mathbf{q}^{n+1}$. The interested reader is referred to [3].

High order in space. To increase the spatial accuracy, the numerical fluxes in the finite volume scheme (18) are fed with high order extrapolated data from the cells sharing the interface. A CWENO reconstruction [21] is performed because it allows for relatively compact stencils even for higher order reconstructions. Specifically, we rely on the very efficient dimension-by-dimension technique forwarded in [9], which ultimately yields a quadrature-free finite volume scheme. By denoting the reconstruction operator with $\mathbb{R}(\mathbf{q})$, the high order numerical fluxes are simply given by

$$\mathbf{f}^c_{i+1/2} = \frac{1}{2}\left(\mathbf{f}^c(\mathbb{R}(\mathbf{q}_{i+1})) + \mathbf{f}^c(\mathbb{R}(\mathbf{q}_i))\right) - \frac{1}{2}s_{\max}\left(\mathbb{R}(\mathbf{q}_{i+1}) - \mathbb{R}(\mathbf{q}_i)\right), \tag{23}$$

where the reconstruction operator must be evaluated at the interface $x_{i+1/2}$. High order finite difference operators are adopted for the implicit terms (20). Further details can be found in [9].

Remark on high Mach number flows. There is no advantage in our method for purely high Mach number flows, where indeed it would be much better to use classical explicit finite volume solvers. To track the shocks, the time step must be still determined taking into account the sound speed according to (7), at the computational price of the solution of the linear system (15). Nevertheless, our numerical scheme is stable even if the time step is chosen larger, namely according to (11), which is not the case for explicit schemes. This might turn to be useful in the case of coexisting different regimes, i.e. low and high Mach number flows, that may occur in the flow at the same time. In this situation, our approach still allows for a rather large time step, which will capture the stiff limit of the model while being stable across shocks.

3.4 Divergence-Free Involution in Multiple Space Dimensions

The extension of the semi-implicit IMEX scheme (13) to multiple space dimensions is carried out considering a Cartesian mesh in both y and z direction, therefore it is straightforward. However, in multiple space dimensions, we must take care of the solenoidal property of the magnetic field, that endows the MHD system with the following involution:

$$\nabla \cdot \mathbf{B} = 0. \tag{24}$$

To respect this condition at the discrete level, we rely on the constrained transport method presented in [14], which corrects the magnetic field by approximating the curl of the magnetic vector potential $\mathbf{A}$ such that $\mathbf{B} = \nabla \times \mathbf{A}$ with a fourth order finite difference scheme. The resulting magnetic field is then proven to be divergence-free by applying a discrete finite difference operator to the discrete curl operator. High order div-curl operators have been recently considered in [7], while curl-free structure preserving schemes have been designed in [8]. All the details of the divergence-free evolution of the magnetic field are reported in [14].

4 Numerical Results

In all the following numerical test problems, the time step is computed according to (11) with CFL $= 0.9$ and the ratio of specific heats is set to $\gamma = 5/3$. Furthermore, the magnetic field is verified to respect the divergence-free condition (24) up to machine precision by measuring the maximum divergence error over the whole domain using a finite difference approximation. Finally, the permeability of the magnetic field is normalized to unity.

4.1 Numerical Convergence Studies

The numerical convergence study is carried out by considering a smooth MHD vortex problem, according to the setup given in [17]. We run second and third order space-time accurate semi-implicit schemes until the final time $t_f = 1$. The results are reported in Table 1, demonstrating that the formal order of accuracy is achieved.

Furthermore, along the lines of [20], we run this problem for different values of the Mach number, namely we consider a vortex with $M = 1.55 \cdot 10^{-5}$, $M = 1.55 \cdot 10^{-4}$ and $M = 1.55 \cdot 10^{-3}$. From the analysis shown in Table 2, we can conclude that the

Table 1 Numerical convergence results for the ideal MHD equations equations using the semi-implicit finite volume schemes (SIFV) for second and third order of accuracy in space and time. The errors are measured in the L_2 norm and refer to the variable u (velocity component in the $x-$direction), p (pressure) and B_x (magnetic field component in the $x-$direction) at time $t_f = 0.1$

SIFV $O(2)$

$N_x = N_y$	$L_2(u)$	$O(u)$	$L_2(p)$	$O(p)$	$L_2(B_x)$	$O(B_x)$
24	7.633E-03	–	6.351E-03	–	2.350E-03	–
32	3.801E-03	2.42	3.212E-03	2.37	1.107E-03	2.16
48	1.512E-03	2.27	1.271E-03	2.29	4.191E-04	2.40
64	8.309E-04	2.08	6.821E-04	2.16	2.223E-04	2.20

SIFV $O(3)$

$N_x = N_y$	$L_2(u)$	$O(u)$	$L_2(p)$	$O(p)$	$L_2(B_x)$	$O(B_x)$
24	5.485E-03	–	5.091E-03	–	1.879E-03	–
32	2.364E-03	2.93	2.397E-03	2.62	7.675E-04	3.11
48	7.228E-04	2.92	8.494E-04	2.56	2.213E-04	3.07
64	3.062E-04	2.99	4.188E-04	2.42	9.620E-05	2.90

Table 2 Numerical convergence results for the ideal MHD equations using the semi-implicit finite volume schemes (SIFV) for third order of accuracy in space and time running the smooth vortex test cases at different Mach numbers. The errors are measured in the L_2 norm and refer to the variable B_x (magnetic field component in the $x-$direction) at time $t_f = 0.1$

	$M = 1.55 \cdot 10^{-5}$		$M = 1.55 \cdot 10^{-4}$		$M = 1.55 \cdot 10^{-3}$	
$N_x = N_y$	$L_2(B_x)$	$O(B_x)$	$L_2(B_x)$	$O(B_x)$	$L_2(B_x)$	$O(B_x)$
24	3.831E-04	–	3.832E-03	–	3.835E-02	–
32	4.461E-05	3.10	4.461E-04	3.10	4.466E-03	3.10
48	5.257E-06	3.08	5.255E-05	3.09	5.259E-04	3.09
64	6.448E-07	3.03	6.438E-06	3.03	4.450E-05	3.03

novel schemes are asymptotic preserving and asymptotic accurate, meaning that no loss of accuracy is observed for low Mach regimes. The distribution of magnetic pressure and hydrodynamics pressure are shown in Fig. 1, where the structure of the vortex is well preserved independently from the Mach number.

4.2 Riemann Problems

In this section, we apply the semi-implicit finite volume scheme to a set of four different Riemann problems of the ideal MHD equations taken from the literature [17, 18]. The aim of this set of test problems is to demonstrate the capability of the semi-implicit scheme to deal with shocks, thus not in the low Mach regime of the fluid. The initial left and right states, which are separated by a discontinuity located at x_d, are listed in Table 3. The computational domain for all Riemann problems is set to $\Omega = [0; 1]$ and the specific heat is defined by $\gamma = 2.0$ for RP1 and $\gamma = \frac{5}{3}$ for the rest. We use a discretization of 200 grid cells for the simulations with the semi-implicit method and the results are compared with those derived by a fully explicit finite volume method using the Rusanov flux on 1024 grid cells. Both methods have second order accuracy in time and space. The comparison between the solution obtained with the semi-implicit scheme and the reference solution is presented in Fig. 2. The results show that the semi-implicit scheme is able to properly capture and resolve the different waves. Only for B_y in RP2 and RP4 the resolution is not sufficient to reproduce every discontinuity in a similar manner as the reference solution. Overall, the results are consistent with those retrieved with other numerical methods in the literature [17, 18].

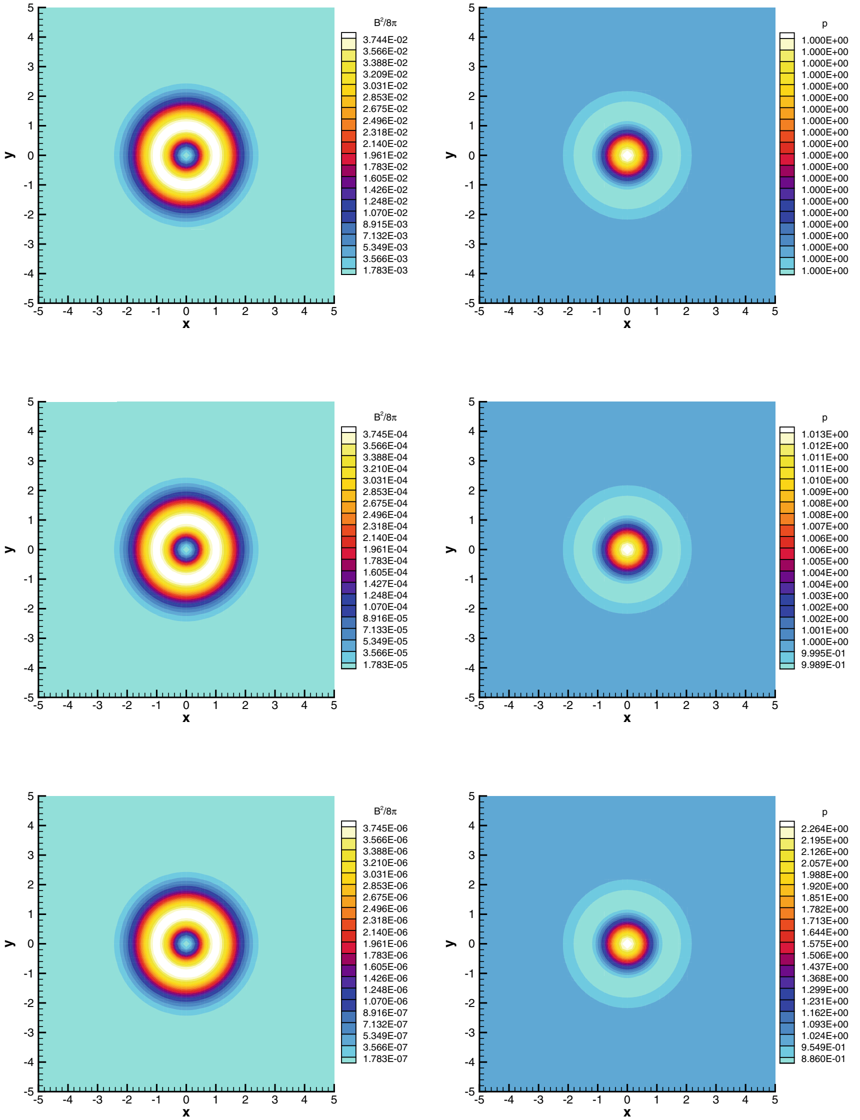


Fig. 1 Smooth MHD vortex. Numerical solution at $t_f = 1$ of magnetic pressure (left column) and hydrodynamics pressure (right column) for Mach number $M = 1.55 \cdot 10^{-3}$ (top), $M = 1.55 \cdot 10^{-4}$ (middle) and $M = 1.55 \cdot 10^{-5}$ (bottom)

Table 3 Initial values for the left and right state for different Riemann problems and the location of the initial discontinuity x_d

Case		x_d	t_f	ρ	u	v	w	p	B_x	B_y	B_z
RP1	L:	0.5	0.12	1.0	0.0	0.0	0.0	1.0	0.75	1.0	0.0
	R:			0.125	0.0	0.0	0.0	0.1	0.75	−1.0	0.0
RP2	L:	0.4	0.2	1.08	1.2	0.01	0.5	0.95	$2.0/\sqrt{4\pi}$	$3.6/\sqrt{4\pi}$	$2.0/\sqrt{4\pi}$
	R:			0.9891	−0.0131	0.0269	0.010037	0.97159	$2.0/\sqrt{4\pi}$	$4.0244/\sqrt{4\pi}$	$2.0026/\sqrt{4\pi}$
RP3	L:	0.4	0.15	1.7	0.0	0.0	0.0	1.7	$3.899398/\sqrt{4\pi}$	$3.544908/\sqrt{4\pi}$	0.0
	R:			0.2	0.0	0.0	−1.496891	0.2	$3.899398/\sqrt{4\pi}$	$2.785898/\sqrt{4\pi}$	$2.192064/\sqrt{4\pi}$
RP4	L:	0.5	0.16	1.0	0.0	0.0	0.0	1.0	1.3	1.0	0.0
	R:			0.4	0.0	0.0	0.0	0.4	1.3	−1.0	0.0

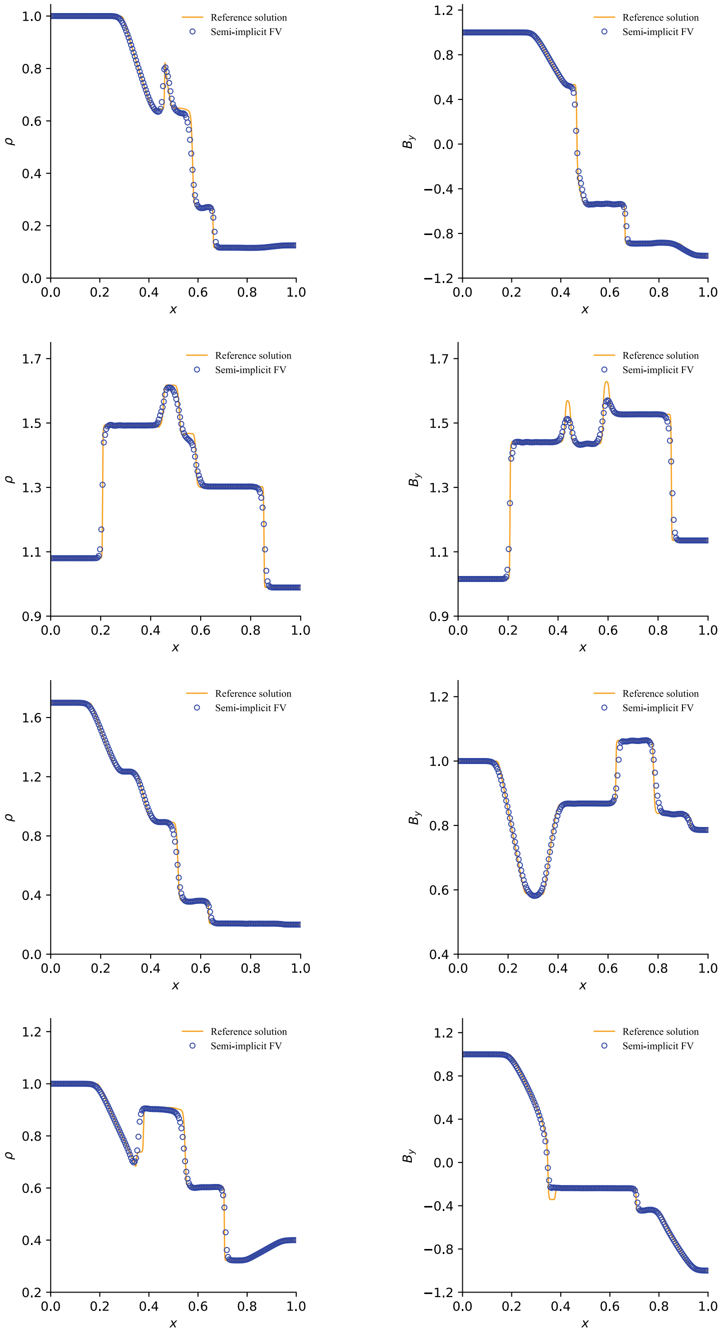


Fig. 2 Riemann problem RP1, RP2, RP3 and RP4 (from top to bottom row) at the final time $t = t_f$. Comparison of density (left column) and magnetic field component B_y (right column) against the reference solution

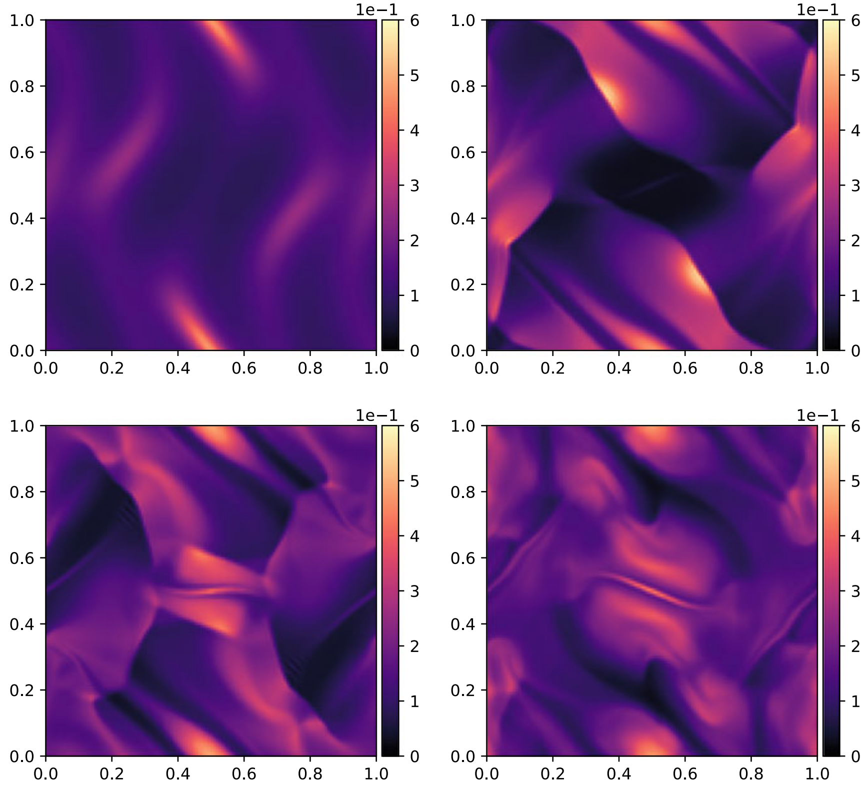


Fig. 3 Orszag-Tang vortex. Numerical solution of pressure at output time $t = 1/12$ (top left), $t = 1/3$ (top right), $t = 0.5$ (bottom left) and $t = 5/6$ (bottom right)

4.3 Orszag-Tang Vortex

A widely used test for the two-dimensional MHD equations is the Orszag Tang vortex [14, 17]. Starting with smooth initial data, over time shocks develop along the diagonal direction in combination with a vortex located at the center of the computational domain. On the spatial domain $\Omega = [0; 1]^2$ the initial condition for the state variables $\mathbf{q}$ is given by

$$\mathbf{q}(0, x, y) = \left(\frac{25}{36\pi},\ -\sin(2\pi y),\ \sin(2\pi x),\ 0.0,\ \frac{5}{12\pi},\ -\frac{1}{\sqrt{4\pi}}\sin(2\pi y),\ \frac{1}{\sqrt{4\pi}}\sin(4\pi x)\right) \tag{25}$$

and the magnetic vector potential $\mathbf{A}$ is initially defined by

$$\mathbf{A}(0, x, y) = \left(0.0,\ 0.0,\ \cos(2\pi y)/(4\pi^{3/2}) + \cos(4\pi x)/(8\pi^{3/2})\right). \tag{26}$$

Periodic boundary conditions are imposed on all sides. The computational domain is discretized by a 128 × 128 control volumes. In Fig. 3 the results for the pressure at different times computed by the third-order semi-implicit method are presented. The numerical method manages to capture the shocks that occur as time evolves. Overall, the results are qualitatively consistent with those in the literature [14, 17].

5 Conclusions

In this work we have presented a pressure-based semi-implicit scheme for the solution of the ideal MHD equations. An elliptic equation for the pressure is obtained to solve low Mach regimes very efficiently without adding any numerical dissipation in the implicit part. On the other hand, an explicit finite volume solver is adopted for handling the nonlinear convective terms, endowing our scheme with shock-capturing properties. The scheme is conservative for mass, momentum and total energy. High order of accuracy is achieved by means of a CWENO reconstruction in space and IMEX Runge-Kutta time stepping techniques. The accuracy and the robustness of the scheme have been demonstrated by performing a numerical convergence study for different Mach numbers and by solving a set of Riemann problems. Another benchmark in numerical MHD has been shown, namely the Orszag-Tang vortex test, proving the capability of the novel method to deal with complex magnetized flows and to provide results which are qualitatively in agreement with other existing numerical schemes in the literature.

Acknowledgements WB received financial support by Fondazione Cariplo and Fondazione CDP (Italy) under the project No. 2022-1895. WB also acknowledges fundings from the Italian Ministry of University and Research (MUR) with the PRIN Project 2022 No. 2022N9BM3N. CB acknowledges the support by the German Research Foundation (DFG) under the project No. KL 566/22-1.

References

1. Ascher, U.M., Ruuth, S.J., Spiteri, R.J.: Implicit-explicit Runge-Kutta methods for time-dependent partial differential equations. Appl. Numer. Math. **25**, 151–167 (1982)
2. Barsukow, W., Edelmann, P., Klingenberg, C., Röpke, F.: A low-Mach Roe-type solver for the Euler equations allowing for gravity source terms. ESAIM: Proc. Surv. **58**, 27–39 (2017)
3. Boscarino, S., Filbet, F., Russo, G.: High order semi-implicit schemes for time dependent partial differential equations. J. Sci. Comput. **68**, 975–1001 (2016)
4. Boscarino, S., Pareschi, L.: On the asymptotic properties of IMEX Runge-Kutta schemes for hyperbolic balance laws. J. Comput. Appl. Math. **316**, 60–73 (2017)
5. Boscarino, S., Pareschi, L., Russo, G.: A unified IMEX Runge-Kutta approach for hyperbolic systems with multiscale relaxation. SIAM J. Numer. Anal. **55**(4), 2085–2109 (2017)
6. Boscarino, S., Russo, G.: On a class of uniformly accurate IMEX Runge-Kutta schemes and applications to hyperbolic systems with relaxation. SIAM J. Sci. Comput. **31**, 1926–1945 (2009)

7. Boscheri, W., Dimarco, G., Pareschi, L.: Locally Structure-Preserving div-curl operators for high order Discontinuous Galerkin schemes. J. Comput. Phys. **486**, 112130 (2023)
8. Boscheri, W., Dumbser, M., Ioriatti, M., Peshkov, I., Romenski, E.: A structure-preserving staggered semi-implicit finite volume scheme for continuum mechanics. J. Comput. Phys. **424**, 109866 (2021)
9. Boscheri, W., Pareschi, L.: High order pressure-based semi-implicit IMEX schemes for the 3D Navier-Stokes equations at all Mach numbers. J. Comput. Phys. **434**, 110206 (2021)
10. Boscheri, Walter, Tavelli, Maurizio: High order semi-implicit schemes for viscous compressible flows in 3d. Appl. Math. Comput. **434**, 127457 (2022)
11. Casulli, V.: Semi-implicit finite difference methods for the two-dimensional shallow water equations. J. Comput. Phys. **86**, 56–74 (1990)
12. Casulli, V.: A semi-implicit finite difference method for non-hydrostatic free-surface flows. Int. J. Num. Meth. Fluids **30**, 425–440 (1999)
13. Chen, W., Wu, K., Xiong, T.: High order asymptotic preserving finite difference weno schemes with constrained transport for MHD equations in all sonic mach numbers. Astron. Astrophys. (2022)
14. Christlieb, Andrew J., Rossmanith, James A., Tang, Qi.: Finite difference weighted essentially non-oscillatory schemes with constrained transport for ideal magnetohydrodynamics. J. Comput. Phys. **268**, 302–325 (2014)
15. Cui, Wenqian, Yaobin, Ou., Ren, Dandan: Incompressible limit of full compressible magnetohydrodynamic equations with well-prepared data in 3-d bounded domains. J. Math. Anal. Appl. **427**(1), 263–288 (2015)
16. Dellacherie, S.: Analysis of Godunov type schemes applied to the compressible Euler system at low Mach number. J. Comput. Phys. **229**, 978–1016 (2010)
17. Dumbser, M., Balsara, D.S., Tavelli, M., Fambri, F.: A divergence-free semi-implicit finite volume scheme for ideal, viscous, and resistive magnetohydrodynamics. Int. J. Numer. Methods Fluids **89**, 16–42 (2019)
18. Fambri, F.: A novel structure preserving semi-implicit finite volume method for viscous and resistive magnetohydrodynamics. Int. J. Numer. Methods Fluids **93**, 3447–3489 (2021)
19. Jiang, S., Ju, Q., Li, F.: Incompressible limit of the compressible magnetohydrodynamic equations with periodic boundary conditions. Commun. Math. Phys. **297**, 371–400 (2010)
20. Leidi, G., Birke, C., Andrassy, R., Higl, J., Edelmann, P.V.F., Wiest, G., Klingenberg, C., Röpke, F.K.: A finite-volume scheme for modeling compressible magnetohydrodynamic flows at low Mach numbers in stellar interiors. Astron. Astrophys. **668**, A143 (2022)
21. Levy, D., Puppo, G., Russo, G.: Central WENO schemes for hyperbolic systems of conservation laws. M2AN Math. Model. Numer. Anal. **33**(3), 547–571 (1999)
22. Mamashita, Tomohiro, Kitamura, Keiichi, Minoshima, Takashi: Slau2-hlld numerical flux with wiggle-sensor for stable low mach magnetohydrodynamics simulations. Comput. Fluids **231**, 105165 (2021)
23. Osher, S., Solomon, F.: A partially implicit method for large stiff systems of Ode's with only few equations introducing small time-constants. SIAM J. Numer. Anal. **13**, 645–663 (1976)
24. Pareschi, L., Russo, G.: Implicit-explicit Runge-Kutta schemes and applications to hyperbolic systems with relaxation. J. Sci. Comput. **25**, 129–155 (2005)
25. Saad, Y., Schultz, M.: GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems. SIAM J. Sci. Stat. Comput. **7**, 856–869 (1986)
26. Thomann, A., Puppo, G., Klingenberg, C.: An all speed second order well-balanced IMEX relaxation scheme for the Euler equations with gravity. J. Comput. Phys. **420**, 109723 (2020)

AeroSPEED: A High Order Acoustic Solver for Aeroacoustic Applications

Alberto Artoni, Paola F. Antonietti, Roberto Corradi, Ilario Mazzieri, Nicola Parolini, Daniele Rocchi, Paolo Schito, and Francesco F. Semeraro

Abstract We propose AeroSPEED, a solver based on the Spectral Element Method (SEM) that solves the aeroacoustic Lighthill's wave equation. First, the fluid solution is computed employing a cell centered Finite Volume method. Then, AeroSPEED maps the sound source coming from the flow solution onto the acoustic grid, where finally the Lighthill's wave equation is solved. An ad-hoc projection strategy is adopted to apply the flow source term in the acoustic solver. A model problem with a manufactured solution and the Noise Box test case are used as benchmark for the acoustic problem. We studied the noise generated by the complex flow field around tandem cylinders as a relevant aeroacoustic application. AeroSPEED is an effective and accurate solver for both acoustics and aeroacoustic problems.

Keywords Acoustics · Aeroacoustics · Spectral element method · Finite volume · Lighthill's equation

1 Introduction

Aeroacoustics is the field of acoustics that studies the noise induced by flows. Due to the different scales involved in the flow and acoustics, usually aeroacoustic problems are posed in a segregated manner [12]. First, a flow problem is solved. Since finite volumes are largely employed in the industrial framework to solve CFD problems, in this work, the open-source finite volume library OpenFOAM [14] is adopted. Then, the flow solution is post-processed to generate the sound source term of the

A. Artoni · P. F. Antonietti (✉) · I. Mazzieri · N. Parolini
MOX, Dipartimento di Matematica, Politecnico di Milano, Piazza Leonardo da Vinci 32, Milano 20133, Italy
e-mail: paola.antonietti@polimi.it

A. Artoni
e-mail: alberto.artoni@polimi.it

R. Corradi · D. Rocchi · P. Schito · F. F. Semeraro
Dipartimento di Meccanica, Politecnico di Milano, Via La Masa 1, Milano 20156, Italy

E. Franck et al. (eds.), *Finite Volumes for Complex Applications X—Volume 1, Elliptic and Parabolic Problems*, Springer Proceedings in Mathematics & Statistics 432,
https://doi.org/10.1007/978-3-031-40864-9_3

Lighthill's wave equation. With this purpose, we have developed AeroSPEED [1, 9], a spectral element based solver that maps the computed sound source term onto the acoustic grid and solves the inhomogeneous Lighthill wave equation. The spectral element method is well suited to solve wave propagation problems, since it provides high accuracy and guarantees both low numerical dispersion and dissipation errors. We validate the open-source acoustic solver AeroSPEED on a model problem based on a manufactured solution, comparing it with COMSOL [2], a commercial software based on Lagrangian Finite Element Method (FEM). As an additional acoustic test case, we considered a geometry representing a simplified cockpit of a car (Noise Box). Next, we apply our aeroacoustic solver AeroSPEED to study the noise induced by the turbulent flow around two tandem circular cylinders.

2 The Coupled Aeroacoustic Model Problem

Given two open, bounded domains $\Omega_F \subseteq \Omega_A \subseteq \mathbb{R}^d$, having sufficiently regular boundaries $\partial\Omega_F$ and $\partial\Omega_A$ respectively (see Fig. 1), we consider the flow on a rigid body at high Reynolds and low Mach numbers. We are interested in the noise generated by an incompressible and acoustically compact flow, meaning that the feedback between the acoustic pressure and the hydrodynamic pressure can be neglected. Hence, the coupled aeroacoustic problem can be posed in a segregated manner. The segregated approach considers the following sequence of problems.

Flow Problem. The fluid flow is governed by the incompressible Navier Stokes equations: for $t \in (0, T]$, find the velocity field $\mathbf{u}(\mathbf{x}, t) : \Omega_F \times (0, T] \to \mathbb{R}^3$ and the pressure field $p(\mathbf{x}, t) : \Omega_F \times (0, T] \to \mathbb{R}$ such that

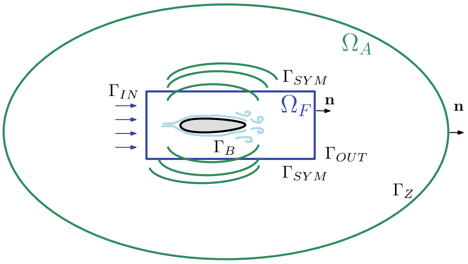


Fig. 1 Domain of the aeroacoustic problem. Ω_F is the fluid domain, while Ω_A is the acoustic domain

$$
\begin{aligned}
\frac{\partial \mathbf{u}}{\partial t} + \nabla \cdot (\mathbf{u} \otimes \mathbf{u}) - \nabla \cdot (\nu \nabla \mathbf{u}) + \nabla \left(\frac{p}{\rho_0} \right) &= 0, && \text{in } \Omega_F \times (0, T], \\
\nabla \cdot \mathbf{u} &= 0, && \text{in } \Omega_F \times (0, T], \\
\mathbf{u} &= \mathbf{0}, && \text{on } \Gamma_B, \\
\mathbf{u} &= \mathbf{g}, && \text{on } \Gamma_{IN}, \\
\nu \nabla \mathbf{u} \cdot \mathbf{n} - p\mathbf{n} &= \mathbf{0}, && \text{on } \Gamma_{OUT}, \\
\mathbf{u} \cdot \mathbf{n} &= \mathbf{0}, && \text{on } \Gamma_{SYM}, \\
\nabla(\mathbf{u} - (\mathbf{u} \cdot \mathbf{n})\mathbf{n}) \cdot \mathbf{n} &= \mathbf{0}, && \text{on } \Gamma_{SYM},
\end{aligned} \tag{1}
$$

with initial condition $\mathbf{u}(\mathbf{x}, 0) = \mathbf{0}$, and where $\mathbf{n}$ is the outward unit normal vector to $\partial \Omega_F$, ν is the kinematic viscosity, ρ_0 is the fluid density and $\mathbf{g}$ is the inlet Dirichlet datum. Moreover, we suppose the boundary of the fluid domain to be decomposed in the inlet boundary Γ_{IN}, the outlet boundary Γ_{OUT}, the body boundary Γ_B and the boundary Γ_{SYM}, such that $\partial \Omega_F = \Gamma_{IN} \cup \Gamma_{OUT} \cup \Gamma_B \cup \Gamma_{SYM}$.

Aeroacoustic Source. Based on the Lighthill analogy [12], from the fluid velocity $\mathbf{u}$ we compute the Lighthill's tensor as $\mathbf{T} = \rho_0 \mathbf{u} \otimes \mathbf{u}$. The Lighthill's tensor has support only on the fluid domain $\Omega_F \subseteq \Omega_A$, and it depends only on the solution $\mathbf{u}$ of problem (1). The Lighthill's tensor represents the sound source term and the coupling term between the flow problem (1) and the acoustic problem (2).

Acoustic Problem. We consider in Ω_A the following non-homogeneous acoustic wave equation: for $t \in (0, T]$, find the density field $\rho(\mathbf{x}, t) : \Omega_A \times (0, T] \to \mathbb{R}$ such that

$$
\begin{aligned}
\frac{\partial^2 \rho}{\partial t^2} - c_0^2 \Delta \rho &= f, && \text{in } \Omega_A \times (0, T], \\
c_0^2 \frac{\partial \rho}{\partial \mathbf{n}} &= 0, && \text{on } \Gamma_B \times (0, T], \\
\frac{1}{\rho_0} \frac{\partial \rho}{\partial \mathbf{n}} &= -\frac{1}{Z} \frac{\partial \rho}{\partial t}, && \text{on } \Gamma_Z \times (0, T],
\end{aligned} \tag{2}
$$

with initial conditions $\rho(\mathbf{x}, 0) = 0$, $\frac{\partial \rho}{\partial t}(\mathbf{x}, 0) = 0$, where c_0 is the sound speed and Z is the impedance of an external wall. The boundary $\partial \Omega_A$ has been split as $\partial \Omega_A = \Gamma_Z \cup \Gamma_B$ where Γ_B is the body boundary where we set a sound hard boundary condition, while Γ_Z is the external boundary where we apply an impedance boundary condition. We recall that if Z is the characteristic impedance, i.e. $Z = \rho_0 c_0$, we have a non-reflective boundary condition, which is necessary for free-field wave propagation problems. When dealing with aeroacoustic problems the sound source is $f = \nabla \cdot \nabla \cdot \mathbf{T}$, obtaining the so called Lighthill's wave equation.

3 Numerical Scheme

We introduce the spectral element method for the spatial discretization of (2) with a generic source term f, highlighting the aeroacoustic case where $f = \nabla \cdot \nabla \cdot \mathbf{T}$ inside Ω_F, while $f = 0$ in $\Omega_A \setminus \Omega_F$.

3.1 Spectral Element Dicretization

Given the acoustic domain Ω_A, we introduce a conforming decomposition $\mathcal{T}_A$ made by hexaedral elements κ_A and we denote the characteristic mesh size as $h_A = \max\limits_{\kappa_A \in \mathcal{T}_A} h_{\kappa_A}$, being h_{κ_A} the diameter of the element κ_A. Let $\widehat{\kappa}$ be the reference element $\widehat{\kappa} = [-1, 1]^3$, and assume that for any hexaedral element $\kappa_A \in \mathcal{T}_A$ there exists a suitable trilinear invertible map $\boldsymbol{\theta}_{\kappa_A} : \widehat{\kappa} \rightarrow \kappa_A$, such that its Jacobian $\mathbf{J}_{\kappa_A}$ is positive. We now introduce the following finite-dimensional space:

$$V_A = \left\{ v \in C^0(\overline{\Omega}_A) \cap H^1(\Omega_A) : v|_{\kappa_A} \circ \boldsymbol{\theta}_{\kappa_A}^{-1} \in \mathbb{Q}_r(\widehat{\kappa}), \forall \kappa_A \in \mathcal{T}_A \right\}, \tag{3}$$

where $\mathbb{Q}_r(\widehat{\kappa})$ is the space of polynomials of degree less than or equal to $r \geq 1$ in each coordinate direction. Next, for any $u, w \in V_A$, we define the following bilinear form by means of the Gauss-Legendre-Lobatto (GLL) quadrature rule:

$$(u, w)_{\kappa_A}^{NI} = \sum_{i,j,k=0}^{r} u(\boldsymbol{\theta}_{\kappa_A}(\boldsymbol{\xi}_{i,j,k}^{GLL}))w(\boldsymbol{\theta}_{\kappa_A}(\boldsymbol{\xi}_{i,j,k}^{GLL}))\omega_{i,j,k}^{GLL} |\det(\mathbf{J}_{\kappa_A})| \approx (u, w)_{\kappa_A}, \tag{4}$$

and we denote with

$$(u, w)_{\mathcal{T}_A}^{NI} = \sum_{\kappa_A \in \mathcal{T}_A} (u, w)_{\kappa_A}^{NI} \quad \forall u, w \in V_A,$$

where $\boldsymbol{\xi}^{GLL}$ are the GLL quadrature nodes, ω^{GLL} the corresponding weights defined on $\widehat{\kappa}$ and NI stands for numerical integration.

3.2 Discretization of the Acoustic Problem

Weak Formulation. We derive the weak formulation of the inhomogeneous wave equation in (2): *for* $t \in (0; T]$, *find* $\rho(\mathbf{x}, t) \in H^1(\Omega_A)$ *such that* $\forall w \in H^1(\Omega_A)$:

$$\left(\frac{\partial^2 \rho}{\partial t^2}, w \right)_{\Omega_A} + c_0^2 (\nabla \rho, \nabla w)_{\Omega_A} + \frac{\rho_0 c_0^2}{Z} \int_{\Gamma_Z} \frac{\partial \rho}{\partial t} w \, ds = \mathcal{L}(w), \tag{5}$$

with initial conditions $\rho = \dfrac{\partial \rho}{\partial t} = 0$ in $\Omega_A \times \{0\}$, and $\mathcal{L}(w)$ is a suitable linear operator. When dealing with general inhomogeneous acoustic problems, the operator is

$$\mathcal{L}(w) = (f, w)_{\Omega_A}, \tag{6}$$

while when dealing with Lighthill's wave equation the term $(\nabla \cdot \nabla \cdot \mathbf{T}, w)_{\Omega_A}$ is usually integrated by parts, and we have that

$$\mathcal{L}(w) = -(\nabla \cdot \mathbf{T}, \nabla w)_{\Omega_A} + ((\nabla \cdot \mathbf{T}) \cdot \mathbf{n}, w)_{\Gamma_B} + ((\nabla \cdot \mathbf{T}) \cdot \mathbf{n}, w)_{\partial\Omega_F \setminus \Gamma_B}. \tag{7}$$

We assume that in the far field $((\nabla \cdot \mathbf{T}) \cdot \mathbf{n}, w)_{\partial\Omega_F \setminus \Gamma_B}$ is negligible, since the velocity $\mathbf{u}$ can be considered uniform and constant. If we consider the local momentum equation on the boundary Γ_B multiplied by the normal direction $\mathbf{n}$ we have:

$$\rho_0 \frac{\partial \mathbf{u}}{\partial t} \cdot \mathbf{n} + \nabla \cdot (\rho_0 \mathbf{u} \otimes \mathbf{u}) \cdot \mathbf{n} + \nabla p \cdot \mathbf{n} = 0. \tag{8}$$

Assuming a solid body, meaning $\mathbf{u} \cdot \mathbf{n} = 0$, we have $((\nabla \cdot \mathbf{T}) \cdot \mathbf{n}, w)_{\Gamma_B} = (-\nabla p \cdot \mathbf{n}, w)_{\Gamma_B}$. For acoustically rigid bodies we impose $(\nabla p \cdot \mathbf{n}, w)_{\Gamma_B} = 0$. Hence we have:

$$\mathcal{L}(w) = -(\nabla \cdot \mathbf{T}, \nabla w)_{\Omega_A}, \tag{9}$$

see for instance [1, 15].

Semi-Discrete Spectral Element Formulation. For the sake of simplicity, we assume that $\partial\Omega_A = \Gamma_N$, and hence $\Gamma_Z = \emptyset$. The semi-discrete spectral element formulation of problem (5) with numerical integration (SEM-NI) reads: *for any time $t \in (0; T]$ find $\rho_h \in V_A$ such that:*

$$\left(\frac{\partial^2 \rho_h}{\partial t^2}, w_h\right)^{NI}_{\mathcal{T}_A} + c_0^2 (\nabla \rho_h, \nabla w_h)^{NI}_{\mathcal{T}_A} = \mathcal{L}_h(w_h) \quad \forall w_h \in V_A, \tag{10}$$

with $\rho_h = \dfrac{\partial \rho_h}{\partial t} = 0$ in $\Omega_A \times \{0\}$, where $\mathcal{L}_h(w) = (f, w)^{NI}_{\mathcal{T}_A}$ for the purely acoustic case, while $\mathcal{L}_h(w) = -(\nabla \cdot \mathbf{T}, \nabla w)^{NI}_{\mathcal{T}_A}$ for the aeroacoustic case.

Computation of the Aeroacoustic Source Term. Let $\mathcal{T}_F$ be a polyhedral decomposition of the fluid domain Ω_F and let

$$V_F = \left\{ v_F \in L^2(\Omega_F) : v_F|_{\kappa_F} \in \mathbb{P}^0(\kappa_F), \forall \kappa_F \in \mathcal{T}_F \right\} \tag{11}$$

be the space of piecewise discontinuous functions. The sound source term $\nabla \cdot (\rho_0 \mathbf{u} \otimes \mathbf{u})$ is computed as a post-process of the numerical solution of problem (1) on the fluid grid $\mathcal{T}_F$ via a Gauss discretization with a linear reconstruction. Namely, given the velocity $\mathbf{u}_h^k \in V_F$ at time t^k, we compute the cell value of the sound source term on the cell $\kappa_F \in \mathcal{T}_F$ as:

$$\frac{1}{|\kappa_F|}\int_{\kappa_F}\rho_0\nabla\cdot(\mathbf{u}_h^k\otimes\mathbf{u}_h^k)d\mathbf{x}=\frac{1}{|\kappa_F|}\int_{\partial\kappa_F}\rho_0\mathbf{u}_h^k(\mathbf{u}_h^k\cdot\mathbf{n}_{\mathcal{F}})ds$$
$$\approx\frac{1}{|\kappa_F|}\sum_{\mathcal{F}\in\partial\kappa_F}\rho_0\mathbf{u}_{\mathcal{F}}(\mathbf{u}_{\mathcal{F}}\cdot\mathbf{n}_{\mathcal{F}})|\mathcal{F}|, \qquad (12)$$

where $\mathbf{u}_{\mathcal{F}}=\mathbf{u}(\mathbf{x}_{\mathcal{F}})$, $\mathbf{n}_{\mathcal{F}}$ is the unit normal to the face $\mathcal{F}\in\partial\kappa_F$ and where we applied a mid-point quadrature rule using the mid-point $\mathbf{x}_{\mathcal{F}}$ of face $\mathcal{F}$. The value of the velocity at the face centre $\mathbf{x}_{\mathcal{F}}$ is computed with a linear interpolation.

Projection of the Aeroacoustic Source Term. Let $q_F\in V_F$ be a function defined on the fluid grid $\mathcal{T}_F$ such that $q_F=\sum_{i=1}^{N_F}\widehat{q}_{F,i}\phi_{F,i}$, where $\{\phi_{F,i}\}_i^{N_F}$ is the set of N_F basis functions associated to V_F, and $\widehat{q}_{F,i}$ are the corresponding expansion coefficients. We define the L^2-projection of the field $q_F\in V_F$ into V_A as: *find* $q_A\in V_A$ *s.t.*

$$(q_A,\phi_{A,i})_{\mathcal{T}_A}=(q_F,\phi_{A,i})_{\mathcal{T}_A}\quad\forall\phi_{A,i}\in V_A, \qquad (13)$$

where $q_A\in V_A$ is a function defined on the discrete acoustic space such that $q_A=\sum_{i=1}^{N_A}\widehat{q}_{A,i}\phi_{A,i}$, where $\{\phi_{A,i}\}_i^{N_A}$ is the set of N_A basis functions, and $\widehat{q}_{A,i}$ are the corresponding expansion coefficients. Since q_F is a piecewise constant over $\mathcal{T}_F$, namely $q_F\in V_F$, we recast problem (13) as follows:

$$\sum_{\kappa_A\in\mathcal{T}_A}(q_A,\phi_{A,i})_{\kappa_A}=\sum_{\kappa_A\in\mathcal{T}_A}\left(\sum_{\ell=1}^{N_F}\widehat{q}_{F,\ell}\phi_{F,\ell},\phi_{A,i}\right)_{\kappa_A}$$
$$=\sum_{\kappa_A\in\mathcal{T}_A}\sum_{\ell=1}^{N_F}\widehat{q}_{F,\ell}(1,\phi_{A,i})_{\kappa_A\cap\kappa_{F,\ell}}, \qquad (14)$$

where we have used that $\kappa_{F,\ell}=\mathrm{supp}(\phi_{F,\ell})$. The discrete algebraic counterpart of (14) becomes

$$\mathbf{M}^{AA}\widehat{\mathbf{q}}_A=\mathbf{M}^{AF}\widehat{\mathbf{q}}_F, \qquad (15)$$

where $\mathbf{M}^{AA}\in\mathbb{R}^{N_A\times N_A}$ with

$$\mathbf{M}^{AA}_{i,j}=\sum_{\kappa_A\in\mathcal{T}_A}(\phi_{A,j},\phi_{A,i})_{\kappa_A},\qquad i,j=1,\ldots,N_A, \qquad (16)$$

is the full mass matrix and $\mathbf{M}^{AF}\in\mathbb{R}^{N_A\times N_F}$ is defined as

$$\mathbf{M}^{AF}_{i,\ell}=\sum_{\kappa_A\in\mathcal{T}_A}\int_{\kappa_A\cap\kappa_{F,\ell}}\phi_{A,i}\,\mathrm{d}\mathbf{x},\qquad i=1,\ldots,N_A,\ \ell=1,\ldots,N_F. \qquad (17)$$

The vector $\widehat{\mathbf{q}}_A$ in (15) collects the expansion coefficients of the projected acoustic field q_A, while $\widehat{\mathbf{q}}_F$ collects the expansion coefficients of the donor fluid field q_F. Further details on the implementation of the projection method are given in [1].

Algebraic Formulation of the Semi-discrete Problem. We introduce the mass and stiffness matrix $\mathbf{M}, \mathbf{K} \in \mathbb{R}^{N_A \times N_A}$:

$$\mathbf{M}_{i,j} = \sum_{\kappa_A \in \mathcal{T}_A} (\phi_{A,j}, \phi_{A,i})^{NI}_{\kappa_A}, \qquad \mathbf{K}_{i,j} = \sum_{\kappa_A \in \mathcal{T}_A} (\nabla\phi_{A,j}, \nabla\phi_{A,i})^{NI}_{\kappa_A}, \tag{18}$$

for $i, j = 1, \ldots, N_A$. We remark that the mass matrix $\mathbf{M}$ computed with the quadrature formula in Eq. (4) becomes diagonal and hence $\mathbf{M} \neq \mathbf{M}^{AA}$. Since $\rho_h \in V_A$, we write $\rho_h = \sum_{i=1}^{N_A} \widehat{\rho}_{A,i}\phi_{A,i}$, where $\left\{\phi_{A,i}\right\}_i^{N_A}$ is the set of N_A basis functions associated to V_A. Collecting the expansion coefficients $\widehat{\rho}_{A,i}$ into the vector $\boldsymbol{\rho}_h$, we obtain the following algebraic semi-discrete formulation:

$$\mathbf{M}\ddot{\boldsymbol{\rho}}_h + c_0^2\mathbf{K}\boldsymbol{\rho}_h = \mathbf{f}, \tag{19}$$

supplemented with the initial conditions $\boldsymbol{\rho}_h = \mathbf{0}$ and $\dot{\boldsymbol{\rho}}_h = \mathbf{0}$.

For an inhomogeneous acoustic problem we define

$$\mathbf{f}_i = \mathcal{L}_h(\phi_{A,i}) = \sum_{\kappa_A \in \mathcal{T}_A} (f, \phi_{A,i})^{NI}_{\kappa_A}, \quad i = 1, \ldots, N_A, \tag{20}$$

while for the aeroacoustic problem we have that:

$$\mathbf{f}_i = \Big[\sum_{\ell = x,y,z} \mathbf{C}^\ell \widehat{\mathbf{q}}_{A,\ell} \Big]_i, \quad i = 1, \ldots, N_A. \tag{21}$$

For the aeroacoustic case in fact, given $\widehat{\mathbf{q}}_{A,\ell}$ solution of the projection problem (14) with $\widehat{\mathbf{q}}_{F,\ell} = [\nabla \cdot \mathbf{T}]_\ell$, computed as described in Eq. (12), with $l = x, y, z$ representing each component, we have that:

$$\begin{aligned} \mathcal{L}_h(\phi_{A,i}) &= \sum_{\ell=x,y,z} \left(\sum_{\kappa_A \in \mathcal{T}_A} -(q_{A,\ell}, [\nabla\phi_{A,i}]_\ell)^{NI}_{\kappa_A} \right) \\ &= \sum_{\ell=x,y,z} \left(\sum_{\kappa_A \in \mathcal{T}_A} -\Big(\sum_{j=1}^{N_A} \widehat{q}_{A,\ell}\phi_{A,j}, [\nabla\phi_{A,i}]_\ell \Big)^{NI}_{\kappa_A} \right) = \Big[\sum_{\ell=x,y,z} \mathbf{C}^\ell \widehat{\mathbf{q}}_{A,\ell} \Big]_i, \end{aligned} \tag{22}$$

with $i = 1, \ldots, N_A$, and where $\mathbf{C}^\ell$ is defined as

$$\mathbf{C}^\ell_{i,j} = \sum_{\kappa_A \in \mathcal{T}_A} (\phi_{A,j}, [\nabla\phi_{A,i}]_\ell)^{NI}_{\kappa_A}, \qquad \text{for } i, j = 1, \ldots, N_A. \tag{23}$$

Time Discretization. For the time discretization of problem (19) we divide the temporal interval $(0, T]$ into N subintervals, such that $T = N\Delta t$, and we set $t^k = k\Delta t$, with $k = 0, \dots, N-1$ and introduce the auxiliary variables $\boldsymbol{v}_h^k = \dot{\boldsymbol{\rho}}_h(t^k)$, $\boldsymbol{a}_h^k = \ddot{\boldsymbol{\rho}}_h(t^k)$. Furthermore, since the mass matrix $\mathbf{M}$ is not singular, we can represent Eq. (19) as:

$$\ddot{\boldsymbol{\rho}}_h = \mathcal{A}(\boldsymbol{\rho}_h, t), \tag{24}$$

where $\mathcal{A}(\boldsymbol{\rho}_h, t) = \mathbf{M}^{-1}(\mathbf{f} - c_0^2\mathbf{K}\boldsymbol{\rho}_h)$. We now employ the Newmark method to discretize Eq. (24):

$$\begin{aligned} \boldsymbol{\rho}_h^{k+1} &= \boldsymbol{\rho}_h^k + \Delta t\boldsymbol{v}_h^k + \Delta t^2\Big(\beta_N\mathcal{A}^{k+1} + \left(\frac{1}{2} - \beta_N\right)\mathcal{A}^k\Big), \\ \boldsymbol{v}_h^{k+1} &= \boldsymbol{v}_h^k + \Delta t\Big(\gamma_N\mathcal{A}^{k+1} + (1-\gamma_N)\mathcal{A}^k\Big), \end{aligned} \tag{25}$$

where $0 \le \beta_N \le \dfrac{1}{2}$ and $0 \le \gamma_N \le 1$ are parameters of the Newmark scheme, and where $\mathcal{A}^k = \mathcal{A}(\boldsymbol{\rho}_h^k, t^k)$.

4 Curle Analogy

By following the derivation in [3], we recall the Curle aeroacoustic solution for problem (2), that will be considered for comparison in the numerical results presented in Sect. 6. Given an observer located at $\mathbf{x}$ at the time t, a volume V and a body $B \subset V$, we have that:

$$\begin{aligned} p(\mathbf{x}, t) = &\int_{-\infty}^{+\infty}\int_V \frac{\partial^2 T_{ij}(\mathbf{y}, \tau)}{\partial x_i \partial x_j} G(\mathbf{x}, t|\mathbf{y}, \tau)\mathrm{d}\mathbf{y}\mathrm{d}\tau \\ &+ \int_{-\infty}^{+\infty}\int_{\partial B}\left(p(\mathbf{y}, \tau)\frac{\partial G(\mathbf{x}, t|\mathbf{y}, \tau)}{\partial \mathbf{n}} - G(\mathbf{x}, t|\mathbf{y}, \tau)\frac{\partial p(\mathbf{y}, \tau)}{\partial \mathbf{n}}\right) d\partial B\mathrm{d}\tau, \end{aligned} \tag{26}$$

where G is a suitable Green function, V is the control volume, $\mathbf{n}$ is the outward unit normal to the boundary ∂B. We denote with $\mathbf{r} = \mathbf{x} - \mathbf{y}$, being r its modulus. We choose as Green function $G(\mathbf{x}, t|\mathbf{y}, \tau) = \dfrac{1}{4\pi r}\delta(\tau - t + \dfrac{r}{c_0})$ in Eq. (26) to obtain:

$$\begin{aligned} p(\mathbf{x}, t) = &\frac{1}{4\pi}\int_V \frac{1}{r}\Big[\frac{\partial^2 T_{ij}}{\partial x_i \partial x_j}\Big]d\mathbf{y} \\ &+ \int_{\partial B}\frac{1}{4\pi r}\left[\left(\frac{1}{c_0}\frac{\partial p}{\partial t} + \frac{p}{r}\right)\frac{\mathbf{r}}{r}\cdot\mathbf{n} - \frac{\partial p}{\partial \mathbf{n}}\right]d\partial B, \end{aligned} \tag{27}$$

where $[\cdot]$ means that the function has to be evaluated at the retarded time $t - \dfrac{r}{c_0}$. Next, we perform the following simplifications (see [4] for details). First, the volume term containing the Lighthill tensor $\mathbf{T}$ is neglected. Then the retarded times are neglected. This assumption is reasonable if the considered sound sources are compact, that means if the characteristic length of the emitting object D is smaller then the characteristic length λ of the acoustic wave, namely if $D \ll \lambda$. Furthermore, since the object is considered acoustically rigid, i.e. $\frac{\partial p}{\partial \mathbf{n}}\Big|_{\partial B} = 0$, we have that:

$$p(\mathbf{x}, t) = \int_{\partial B} \frac{1}{4\pi r} \left(\left(\frac{1}{c_0} \frac{\partial p}{\partial t} + \frac{p}{r} \right) \frac{\mathbf{r}}{r} \cdot \mathbf{n} \right) d\partial B. \tag{28}$$

By neglecting the viscous forces and considering $\mathbf{F} = \displaystyle\int_{\partial B} p\mathbf{n}$, we have that:

$$p(\mathbf{x}, t) = \frac{1}{4\pi} \frac{\mathbf{r}}{r^2} \cdot \left(\frac{\mathbf{F}}{r} + \frac{1}{c_0} \frac{\partial \mathbf{F}}{\partial t} \right). \tag{29}$$

5 Numerical Results for Acoustic Problems

We consider the inhomogeneous wave equation described in (2) and we compare our software AeroSPEED based on the spectral element approximation introduced in Sect. 3.1 with the commercial software COMSOL [2] based on the Lagrangian FEM.

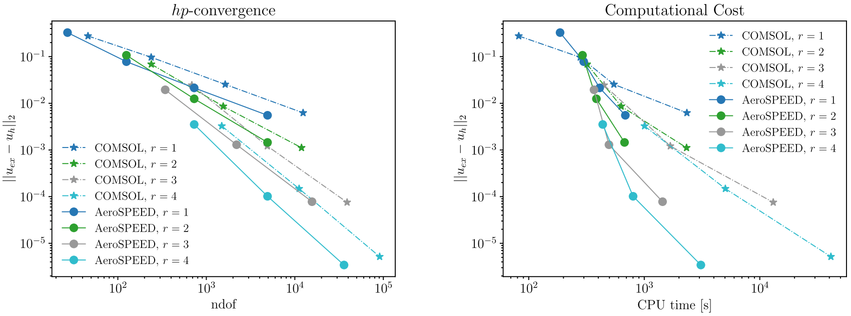
5.1 Verification Test Case

As a first test case, we consider a simple model problem where we verify the performance of both AeroSPEED and COMSOL, in terms of accuracy and computational efficiency.

Acoustic Setup. Given the manufactured solution

$$u_{ex} = \sin(\pi t) \sin(4\pi(x-1)(y-1)(z-1)) \sin(4\pi xyz), \tag{30}$$

we solve an inhomogeneous wave equation on the cube $\Omega_A = (0, 1)^3$, with Neumann boundary conditions on $\Gamma_B = \partial\Omega_A$, with $c_0 = 1\text{m s}^{-1}$. We employ a very fine time step $\Delta t = 1 \times 10^{-6}$s and we use a second order Newmark scheme with parameters $\gamma_N = 0.5$ and $\beta_N = 0.25$. We solve the test case both in AeroSPEED and in COMSOL, changing the refinement of the acoustic grid and the polynomial degree of the



(a) (b)

Fig. 2 Comparison between the SEM solver AeroSPEED and the Lagrangian FEM solver COMSOL. **a** Computed errors vs number of degrees of freedom (ndof). **b** Computed error versus CPU time. The computational tests were performed on 4 cores Intel(R) Xeon(R) Gold 6226 CPU at 2.70 GHz

underlying polynomial approximation and we compute the error $E_2 = ||u_{ex} - u_h||_2$ at the final time $T = 0.5$ s.

Acoustic Results. We report in Fig. 2 the computed errors versus the number of degrees of freedom (left) and the CPU time (right) obtained with the AeroSPEED and COMSOL solvers varying the polynomial approximation degree $r = 1, 2, 3, 4$ of a sequence of meshes with comparable granularity. The expected convergence rates are obtained for both the underlying SEM and FEM approximations, respectively. For a comparable number of degrees of freedom, the SEM approximation is more accurate and less expensive. The results reported in Fig. 2 (right) clearly indicate that AeroSPEED is able to achieve the same error in a much shorter computational time. Moreover, as the underlying polynomial approximation degree increases, AeroSPEED becomes more and more efficient compared to COMSOL.

5.2 *Noise Box Test Case*

We consider a second test case to asses the capabilities of AeroSPEED in solving acoustic problems in a confined geometry and we compare the obtained numerical solution with the one provided by COMSOL.

Acoustic Setup. We consider a geometry that represents a simplified car cockpit, the so called Noise Box, see Fig. 3, introduced in [7]. Each wall is modeled as a real wall (with both partially reflective and partially absorbing behaviour), by setting a wall impedance of $Z = 32206\,\mathrm{pa\,s\,m^{-1}}$. As forcing term we consider a monopole sound source $f(\mathbf{x}, t) = \delta(\mathbf{x} - \mathbf{x}_S)\sin(2\pi f_0 t)$, where $\delta(\mathbf{x} - \mathbf{x}_S)$ is the Dirac delta centered in $\mathbf{x}_S = (1.15, 0.595, 0.065)$ m and $f_0 = 162\,\mathrm{H_z}$. We set the density of air to be $\rho_0 = 1.204\,\mathrm{kg\ m^{-3}}$ and the speed of sound $c_0 = 343\,\mathrm{m\,s^{-1}}$. For the space

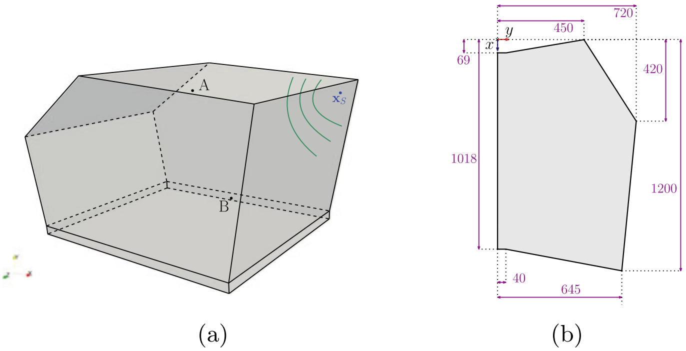


Fig. 3 **a** Three-dimensional view of the domain of the Noise Box. A and B are the positions of the selected microphones, where A = (0.424, 0.595, 0.151) m and B = (0.9, 0.224, 0.528) m. **b** Quoted computational domain of the Noise Box. The spanwise length is 0.825 m. In the figure, units are expressed in millimeters

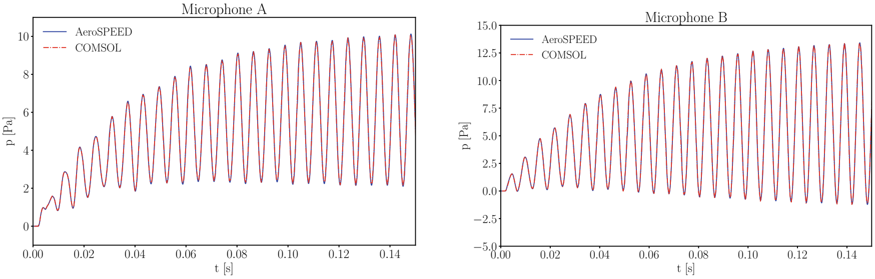
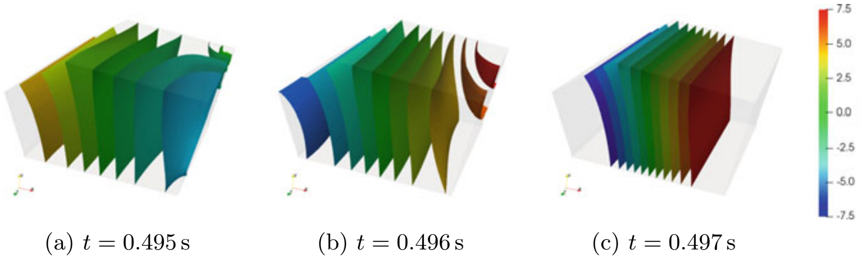


Fig. 4 Computed acoustic pressure measured by microphone A and B with both AeroSPEED and COMSOL

discretization we set the polynomial degree $r = 2$ and we fix $\Delta x = 0.04$ m. For the time discretization we set $\gamma_N = 0.5$ and $\beta_N = 0.25$, with $\Delta t = 5 \times 10^{-6}$ s, with a final time of $T = 0.5$ s. We solve for the same setup both with COMSOL and AeroSPEED and we compare the two results.

Acoustic Results. From the results reported in Fig. 4a we note the initial transient state, up to around $t \approx 0.05$ s. The acoustic monopole is injecting energy in the system, that is not fully dissipated, up until $t \approx 0.1$ s. At that time, the system has reached a stationary regime, where the amount of energy dissipated by the system is balanced by the amount of energy injected. The numerical solution obtained with AeroSPEED perfectly matches the numerical solution obtained with COMSOL. In Fig. 5 we see the stationary pressure waves inside the Noise Box from different snapshots of the solution.



(a) $t = 0.495\,\mathrm{s}$ (b) $t = 0.496\,\mathrm{s}$ (c) $t = 0.497\,\mathrm{s}$

Fig. 5 Snapshots of the computed pressure fluctuations $p' = p - \overline{p}$, where $\overline{p}$ is the average pressure, inside the Noise Box, for $t = 0.495, 0.496, 0.497$ s. The selected contour levels are from -7.5 to 7.5 Pa with a step of 1.5 Pa

6 Numerical Results for an Aeroacoustic Application

We consider an aeroacoustic application, namely the noise generated by two tandem cylinders, a test case that have been subject of a dedicated workshop [8]. The flow simulation has been performed with OpenFOAM [14], while the aeroacoustic coupling has been implemented in AeroSPEED [1].

6.1 Turbulent Flow Around a Tandem Cylinder

Simulating the turbulent flow around two tandem cylinders at high Reynolds number is a challenging problem due to the unsteadiness and complex flow structures to be captured. The separation point on the front cylinder moves on the surface, generating a shear layer that rolls up forming a periodic vortex shedding that impinges on the rear cylinder. As result, a tonal and broadband noise are generated. Proper turbulence models are crucial to simulate at a reasonable computational cost such a complex physics.

Fluid Setup. The two tandem cylinders problem configuration involves two cylinders of equal diameter $D = 0.057\,\mathrm{m}$ aligned along the streamwise direction at a distance of $3.7D$. A sketch of the computational domain is reported in Fig. 6a. At the inlet a fixed velocity of $U_\infty = 44\,\mathrm{m\,s^{-1}}$ is set, corresponding to a Reynolds number $Re = 1.66 \times 10^5$. On Γ_{C1} and Γ_{C2}, no slip conditions are imposed. At the outlet, a zero gradient condition is set. On the remaining boundaries, a symmetry condition is imposed. We choose a fixed time step of $\Delta t = 1.25 \times 10^{-5}$ s, we set the final time to $T = 0.35$ s, and we employ a second order backward difference formula. The height of the first cell near the wall corresponds to $y^+ \approx 30$, and proper wall functions are prescribed, see [11]. Following [5], we employ a DDES $k - \omega$SST model to simulate the turbulent flow, see for instance [6] for more details.

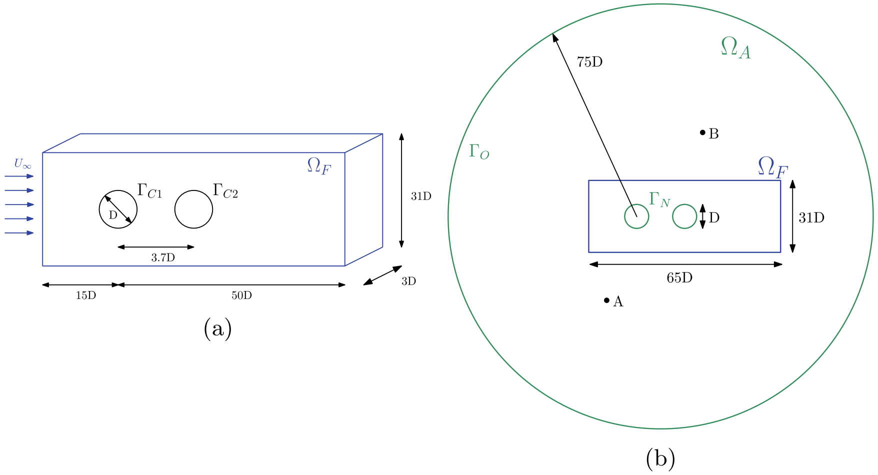


Fig. 6 Tandem cylinders computational domain. **a** Fluid computational domain. **b** Aeroacoustics computational domain. The center of the computational domain is set at the center of the front cylinder. The points A, B are microphone probes located at A = (–8.33D, 27.82D), and B = (9.11D, 32.49D), with $D = 0.057$ m. The front cylinder is denoted by Γ_{C1} and the rear cylinder with Γ_{C2}

Acoustic Setup. Since the acoustic problem can be considered bi-dimensional, we take as sound source only the average along the spanwise direction. A sketch of the computational domain for the aeroacoustic case is depicted in Fig. 6b. On the cylinders $\Gamma_{C1} \cup \Gamma_{C2} = \Gamma_N$ rigid wall boundary conditions are imposed, while at the far field Γ_O absorbing boundary conditions are considered. The fluid sound source is mapped each four time steps, namely the computational time step for the Lighthill's wave equation is $\Delta t = 5 \times 10^{-5}$ s. The chosen polynomial degree is $r = 2$ and the spacing at the far field is $\Delta x \approx 0.04$ m.

Flow Validation. The average flow field is characterized by a mostly symmetric recirculation regions after the cylinders, see Fig. 7 (left). The first recirculation length is about $2D$, aligned with the results of [5]. A visualization of the vortex structures in the flow field, see Fig. 7 (right), is made by employing the Q criterion, where $Q = \frac{1}{2}\Big(\operatorname{tr}(\nabla\mathbf{u})^2 + \operatorname{tr}(\nabla\mathbf{u}\nabla\mathbf{u})\Big)$. In Fig. 8, we compare the prediction on the force coefficients with the results collected in [8]. The results are quite heterogeneous due to the complexity of the problem and the numerous different strategies among the research groups. We denote with C_D and C_L the mean drag and lift coefficients, and their root mean squared rms values as $\widetilde{C_L}$, $\widetilde{C_D}$, respectively. All the computed integral results are within the standard deviation from the literature data. For further comparison, we consider Fig. 9, where we plot the values of the pressure coefficient C_p along the cylinders and we compared it with the experiments [8, 10] and with the computations performed in [5]. The main differences are located in the aft of both cylinders, and are due to the prediction of a different pressure recovery, resulting

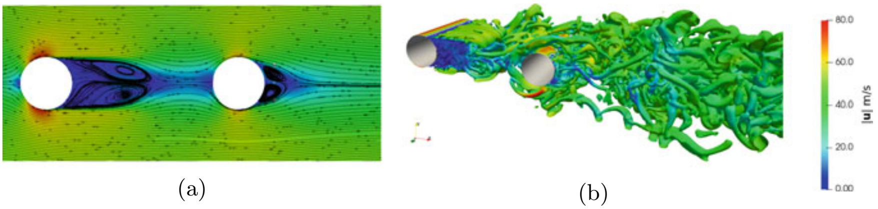


Fig. 7 **a** Average velocity magnitude $|\mathbf{u}|$ on the symmetry plane with streamlines. **b** Isosurfaces of Q $=$ 1000 at t $=$ 0.35 s colored with velocity magnitude

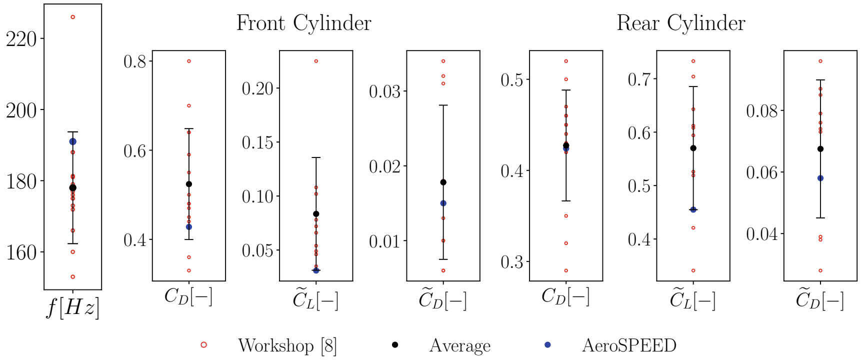


Fig. 8 Comparison of the average drag forces and the rms for lift and drag of the two cylinders. Data have been taken from [8]. The lift frequency is common between the two cylinders. The bar is centred on the mean value of the available literature data, and the length of the bar is one standard deviation

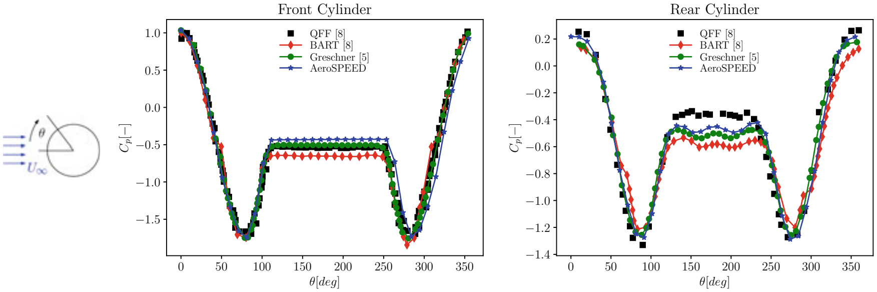


Fig. 9 Average C_P distribution on the surface of the front and rear cylinders, where $C_P = \dfrac{\overline{p} - p_{ref}}{\frac{1}{2}\rho_0 U_\infty^2}$

in small shifts in the separation point locations. Overall, our flow predictions are aligned with the references.

Aeroacoustic Validation. In Fig. 10, a snapshot of the pressure fluctuations induced by the tandem cylinder is shown. The acoustic pressure fluctuations are dominated by a dipole pattern induced by the lift force acting on the rear cylinder. As suggested by

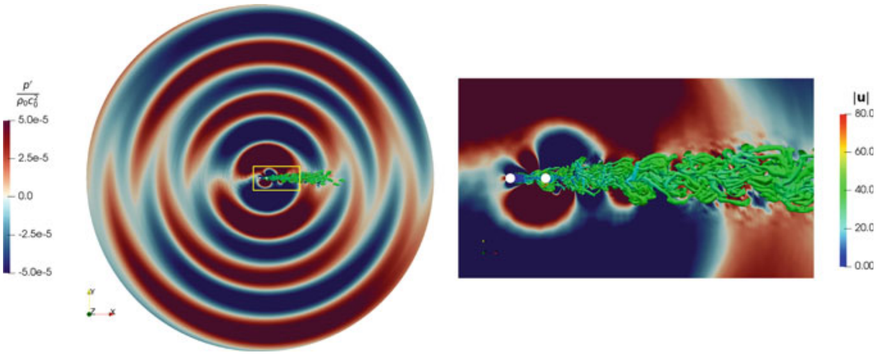


Fig. 10 Snapshot of the fluctuating pressure (from the acoustic computations), and Q criterion colored with the velocity magnitude (from the flow computations)

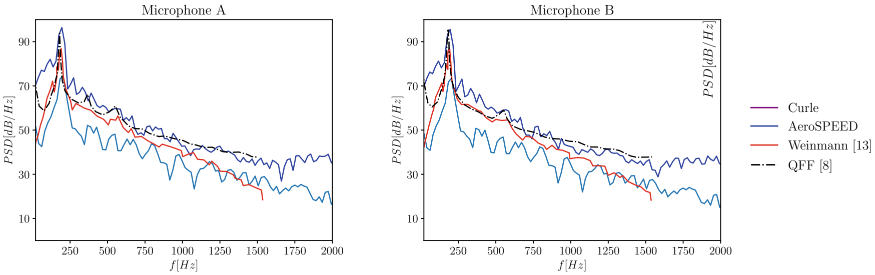


Fig. 11 Comparison of the sound spectra at microphone A = (–8.33D, 27.82D) and B = (9.11D, 32.49D)

Fig. 8, the main contribution to the sound generation is from the rear cylinder, being its $\widetilde{C}_L$ much larger than the front cylinder one. Also, we can compare the different structures coming from the flow with the bigger structures solved by the acoustics. In Fig. 11 we compare the sound spectra obtained with different methodologies, such as the Curle method described in Sect. 4, another Curle analogy with a spanwise corrections proposed in [13], experimental data from QFF [8] and the results computed with AeroSPEED. We observe that, although all the aeroacoustic solutions predict the peak coming from the lift frequency of the rear cylinder, AeroSPEED better matches the PSD values of the experimental data.

7 Conclusion

We introduce AeroSPEED, a solver for aeroacoustic problems that couples a finite volume solution onto a spectral element space and solves the Lighthill's wave equation. The high-order spectral acoustic solver was compared with the commercial software COMSOL on a model problem with a manufactured solution. The accuracy of the numerical results agree with theoretical estimates and the performance of the two solvers is compared in terms of accuracy versus computational time. The spectral element solution obtained with AeroSPEED is able to guarantee higher accuracy with lower computational time. We then applied our solver to simulate the acoustic propagation inside a simplified car cockpit (the Noise Box). The solution obtained with the two different solvers for the pressure signals in two different locations inside the domain perfectly match. Next, we studied a more complex application where the noise is generated by the highly unsteady flow around tandem cylinders. We compared the results obtained with AeroSPEED with experimental and numerical tests performed by many research groups on the tandem cylinder benchmark case, proving the prediction capabilities of the proposed approach also for relevant and more challenging aeroacoustic problems.

References

1. Artoni, A., Antonietti, P. F., Mazzieri, I., Parolini, N., Rocchi, D.: A hybrid finite volume – spectral element method for aeroacoustic problems (2023). arXiv:2302.03370
2. COMSOL Multiphysics v. 6.1. COMSOL AB, Stockholm, Sweden. www.comsol.com
3. Curle, N.: The influence of solid boundaries upon aerodynamic sound. In: Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences, vol. 231, no. 1187 (1955)
4. Epikhin, A., Evdokimov, I., Kraposhin, M., Kalugin, M., Strijhak, S.: Development of a dynamic library for computational aeroacoustics applications using the OpenFOAM open source package. Procedia Comput. Sci. **66**, 150-157 (2015). https://www.sciencedirect.com/science/article/pii/S1877050915033670
5. Greschner, B., Eschricht, D., Mockett, C., Thiele, F.: Turbulence modelling effects on tandem cylinder interaction flow and analysis of installation effects on broadband noise using chimera technique. In: 30th AIAA Applied Aerodynamics Conference (2012)
6. Gritskevich, M.S., Garbaruk, A.V., Schütze, J., Menter, F.R.: Development of DDES and IDDES formulations for the k-ω shear stress transport model. Flow Turbul. Combust. **88**(3), 431–449 (2011)
7. Liu, L.: Design, simulation and testing of a coupled plate-cavity system targeted for vehicle interior noise analysis and control. Ph.D. Thesis, Politecnico di Milano (2021)
8. Lockard, D.: Summary of the tandem cylinder solutions from the benchmark problems for airframe noise computations-I workshop. AIAA 2011-353. In: 49th AIAA Aerospace Sciences Meeting including the New Horizons Forum and Aerospace Exposition (2011)
9. Mazzieri, I., Stupazzini, M., Guidotti, R., Smerzini, C.: SPectral Elements in Elastodynamics with Discontinuous Galerkin: a non-conforming approach for 3D multi-scale problems. Int. J. Numer. Meth. Eng. **95**(12), 991–1010 (2013)
10. Neuhart, D., Jenkins, L., Choudhari, M., Khorrami, M.: Measurements of the flowfield interaction between tandem cylinders, AIAA 2009-3275. In: 15th AIAA/CEAS Aeroacoustics Conference (2009)
11. Spalding, D.B.: A single formula for the "Law of the Wall". J. Appl. Mech. (1961)
12. Schoder, S., Kaltenbacher, M.: Hybrid aeroacoustic computations: state of art and new achievements. J. Theor. Comput. Acoust. **27** (2019)
13. Weinmann, M., Sandberg, R.D., Doolan, C.: Tandem cylinder flow and noise predictions using a hybrid RANS/LES approach. Int. J. Heat Fluid Flow **50** (2014)

14. Weller, H.G., Tabor, G., Jasak, H., Fureby, C.: A tensorial approach to computational continuum mechanics using object-oriented techniques. Comput. Phys. **12**(6), 620–631 (1998)
15. Kaltenbacher, M., Escobar, M., Becker, S., Ali, I.: Computational aeroacoustics based on Lighthill's acoustic analogy. In: Marburg, S., Nolte, B. (eds.) Computational Acoustics of Noise Propagation in Fluids - Finite and Boundary Element Methods. Springer, Berlin, Heidelberg (2008)

Finite Volumes for a Generalized Poisson-Nernst-Planck System with Cross-Diffusion and Size Exclusion

Clément Cancès, Maxime Herda, and Annamaria Massimini

Abstract We present two finite volume approaches for modeling the diffusion of charged particles, specifically ions, in constrained geometries using a degenerate Poisson-Nernst-Planck system with cross-diffusion and volume filling. Both methods utilize a two-point flux approximation and are part of the exponentially fitted scheme framework. The only difference between the two is the selection of a Stolarsky mean for the drift term originating from a self-consistent electric potential. The first version of the scheme, referred to as (SQRA), uses a geometric mean and is an extension of the squareroot approximation scheme. The second scheme, (SG), utilizes an inverse logarithmic mean to create a generalized version of the Scharfetter-Gummel scheme. Both approaches ensure the decay of some discrete free energy. Classical numerical analysis results—existence of discrete solution, convergence of the scheme as the grid size and the time step go to 0—follow. Numerical simulations show that both schemes are effective for moderate Debye lengths, with the (SG) scheme demonstrating greater robustness in the small Debye length limit.

Keywords Drift-diffusion · Cross-diffusion · Exponential fitting · Free energy decay

MSC 2010 65M08 · 65M12 · 35K51

C. Cancès (✉) · M. Herda
Inria, Univ. Lille, CNRS, UMR 8524 - Laboratoire Paul Painlevé, F-59000 Lille, France
e-mail: clement.cances@inria.fr

M. Herda
e-mail: maxime.herda@inria.fr

A. Massimini
Technische Universität Wien, Institute of Analysis and Scientific Computing, Wiedner Hauptstraße 8–10, 1040 Wien, Austria
e-mail: annamaria.massimini@asc.tuwien.ac.at

E. Franck et al. (eds.), *Finite Volumes for Complex Applications X—Volume 1, Elliptic and Parabolic Problems*, Springer Proceedings in Mathematics & Statistics 432,
https://doi.org/10.1007/978-3-031-40864-9_4

1 The Continuous Generalized Poisson-Nernst-Planck Model

Motivated by the transfer of ions in confined geometries, Burger *et al.* introduced in [3] a model accounting for cross-diffusion and size-exclusion effects. In this model, I species, the volume fractions of which being denoted by $U = (u_i)_{1 \le i \le I}$, are subject to diffusion as well as to electric forces induced by a self-consistent electrostatic potential. Denote by $\Omega \subset \mathbb{R}^d$ a bounded connected polyhedral domain, then the conservation of the volume occupied by the species i writes

$$\partial_t u_i + \nabla \cdot F_i = 0, \qquad i = 1, \dots, I, \tag{1}$$

with the flux of the species i being (formally) given by

$$F_i = -D_i \left(u_0 \nabla u_i - u_i \nabla u_0 + u_0 u_i z_i \nabla \phi\right) = -D_i u_i u_0 \nabla \left(\log\left(\frac{u_i}{u_0}\right) + z_i \phi\right). \tag{2}$$

In the above expression, $D_i > 0$ denotes the diffusion coefficient of the species i. The quantity

$$u_0 = 1 - \sum_{i=1}^{I} u_i \tag{3}$$

shall be thought as the volume fraction of available space for the ions, possibly occupied by a mobile and electro-neutral solvent. Denoting by z_i the charge of species i and by $\lambda > 0$ the (scaled) Debye length, then the electrostatic potential solves the Poisson equation

$$-\lambda^2 \Delta \phi = \sum_{i=1}^{I} z_i u_i + f \tag{4}$$

for some prescribed background charge density f. We consider boundary conditions of mixed type for the electric potential. More precisely, we assume that the boundary $\partial\Omega$ of the domain can be split into a insulator part Γ^N and its complement Γ^D on which Dirichlet boundary condition is imposed:

$$\nabla \phi \cdot n = 0 \quad \text{on } \Gamma^N \qquad \text{and} \quad \phi = \phi^D \text{ on } \Gamma^D. \tag{5}$$

Throughout this paper, we will assume that $f \in L^\infty(\Omega)$ and that ϕ^D is the trace of an $L^\infty \cap H^1(\Omega)$ function (which we also denote by ϕ^D). Neither f nor ϕ^D depend on time. Boundary conditions of various types can be considered for the conservation laws (1)–(2), like for instance Robin type boundary condition modeling electrochemical reaction thanks to Butler-Volmer type formula, see for instance [5], or boundary conditions of mixed Dirichlet-Neumann type as in [11]. In the presentation of the scheme, we assume for simplicity that the system is isolated, in the sense that

$$F_i \cdot n = 0 \quad \text{on } \partial\Omega, \qquad i = 1, \dots, I. \tag{6}$$

The system is finally complemented with initial conditions $u_i(t = 0) = u_i^0$ with

$$u_i^0 \geq 0 \quad \text{and} \quad \int_\Omega u_i^0 > 0 \quad \text{for} \quad i = 0, \dots, I \quad \text{and} \quad \sum_{i=0}^{I} u_i^0 = 1. \tag{7}$$

Let us now describe the entropy (or formal gradient flow) structure of the model. Introduce the Slotboom variables $w_i = \frac{u_i}{u_0} e^{z_i \phi}$, then the fluxes (2) rewrite as

$$F_i = -D_i u_0^2 e^{-z_i \phi} \nabla w_i, \qquad i = 1, \dots, I. \tag{8}$$

Multiplying (1) by $\log w_i = \log \frac{u_i}{u_0} + z_i \phi$, integrating over Ω and summing over $i = 1, \dots, I$ yields

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{H} + 4 \int_\Omega \sum_{i=1}^{I} D_i u_0^2 e^{-z_i \phi} |\nabla \sqrt{w_i}|^2 = 0, \tag{9}$$

where, denoting the mixing (neg)entropy density function $H : \mathbb{R}_+^{I+1} \to \mathbb{R}_+^{I+1}$ by

$$H(U) = u_0 \log(u_0) + \sum_{i=1}^{I} u_i \log(u_i) + \log(I + 1) \geq 0,$$

the free energy $\mathcal{H}$ is given by

$$\mathcal{H} = \int_\Omega H(U) + \frac{\lambda^2}{2} \int_\Omega |\nabla \phi|^2 - \lambda^2 \int_{\Gamma^D} \phi^D \nabla \phi \cdot n.$$

Assume that u_0 is nonnegative (as established in [11] and proved in the discrete case later on), then the second term in (9) is non-negative. As a consequence, the free energy decays along time, as a manifestation of the second principle of thermodynamics. Observe that $\mathcal{H}$ need not be non-negative but may be bounded uniformly from below by a constant depending only on λ, f and ϕ^D.

A finite volume scheme has been studied in [4]. Even though the scheme mainly behaves well in practice, its mathematical study is very partial since requiring strong assumptions such as constant diffusion coefficients $D_i = D$ for all i, or no charge $z_i = 0$. Moreover, since the scheme proposed in [4] uses upwinding for the mobilities, numerical experiments exhibit a mere first order convergence in space. An alternative finite element method using the so-called electrochemical potentials $\mu_i = \log(w_i)$ rather than the u_i as primary variables has been analyzed in [12]. This latter scheme is by construction free energy diminishing without further restriction on the physical parameters, and is shown to converge towards a weak solution as the mesh size and the time step tend to 0 (up to quadrature error terms). Second order convergence w.r.t. the mesh size is observed, but the nonlinear system to be solved at each time step is stiffer than for the finite volume scheme because of the use of the electrochemical potentials as variables, so that no clear gain was observed in comparison with the upstream mobility finite volumes. The finite volume scheme proposed in [1], in which the fluxes F_i are approximated thanks to the second expression of (2) also leads to singular numerical fluxes expressions. Our goal here is to propose and to analyze a scheme which shares the best with the aforementioned approaches: decay of the free energy and unconditional convergence are established, second order accuracy in space and well-behaved nonlinear system for moderately small Debye length.

2 Two TPFA Finite Volume Schemes

First, we introduce the time discretization and the spatial mesh of the domain Ω. The mesh will be assumed to be admissible in the sense of [9], in the sense that it fulfills the so-called *orthogonality condition*.

Let $\mathcal{T}$ denote a family of non-empty, disjointed, convex, open and polygonal *control volumes* $K \in \mathcal{T}$, whose Lebesgue measure is denoted by m_K. We also assume that control volumes partition the domain in the sense that $\overline{\Omega} = \bigcup_{K\in\mathcal{T}} \overline{K}$. Further, we call $\mathcal{E}$ a *family of edges/faces*, where $\sigma \in \mathcal{E}$ is a closed subset of $\overline{\Omega}$ contained in a hyperplane of $\mathbb{R}^d$. Each σ has a strictly positive $(d-1)$-dimensional Hausdorff (or Lebesgue) measure, denoted by m_σ. We use the abbreviation $K|L = \partial K \cap \partial L$ for the intersection between two distinct control volumes which is either empty or reduces to a face contained in $\mathcal{E}$. The subset of all interior faces is denoted by

$$\mathcal{E}_{\text{int}} = \{\sigma \in \mathcal{E} \text{ s. t. } \sigma = K|L \text{ for some } K, L \in \mathcal{T}\}.$$

For any $K \in \mathcal{T}$, we assume that there exists a subset $\mathcal{E}_K$ of distinct elements of $\mathcal{E}$ such that the boundary of a control volume can be described by $\partial K = \bigcup_{\sigma\in\mathcal{E}_K} \sigma$ and, consequently, it follows that $\mathcal{E} = \bigcup_{K\in\mathcal{T}} \mathcal{E}_K$. Additionally, we assume that boundary edges $\mathcal{E}_{\text{ext}} = \mathcal{E} \setminus \mathcal{E}_{\text{int}}$ are either subsets of Γ^D or Γ^N. To each control volume $K \in \mathcal{T}$ we assign a *cell center* $x_K \in K$ which satisfies the *orthogonality condition*: If K, L share a face $\sigma = K|L$, then the vector $x_L - x_K$ is orthogonal to $\sigma = K|L$. The triplet $(\mathcal{T}, \mathcal{E}, \{x_K\}_{K\in\mathcal{T}})$ is called an *admissible mesh*.

We introduce the notation d_σ for the Euclidean distance between x_K and x_L if $\sigma = K|L$ or between x_K and the affine hyperplane spanned by σ if $\sigma \subset \partial\Omega$. We also denote by $d_{K\sigma} = \text{dist}(x_K, \sigma)$, so that $d_\sigma = d_{K\sigma} + d_{L\sigma}$ if $\sigma = K|L \in \mathcal{E}_{\text{int}}$ and $d_\sigma = d_{K\sigma}$ if $\sigma \in \mathcal{E}_K \cap \mathcal{E}_{\text{ext}}$. The *transmittivity* of the edge $\sigma \in \mathcal{E}$ is defined by $a_\sigma = \frac{m_\sigma}{d_\sigma}$. The size of the mesh is $h = \max_{K\in\mathcal{T}} \text{diam}(K)$ where $\text{diam}(K)$ denotes the diameter of the cell K. The regularity of the mesh is defined by

$$\zeta = \max_{K\in\mathcal{T}} \left(\text{card } \mathcal{E}_K \ ; \ \max_{\sigma\in\mathcal{E}_K} \frac{\text{diam}(K)}{d_{K\sigma}} \right).$$

For the time discretization we decompose the time interval $\mathbb{R}_+ := [0, +\infty)$ into a sequence of increasing number of time steps $0 = t^0 < t^1 < \cdots$ with a stepsize

$$\tau^n = t^n - t^{n-1}$$

at time step $n \in \mathbb{N} \setminus \{0\}$. We finally introduce $\Delta t = \sup_{n\in\mathbb{N}\setminus\{0\}} \tau^n$, which we assume to be finite.

We are now in position to define the finite volume scheme. Let us start with the discretization of the Poisson equation (4) and (5), which relies on a classical two-point flux approximation

$$\lambda^2 \sum_{\sigma\in\mathcal{E}_K} a_\sigma(\phi^n_K - \phi^n_{K\sigma}) = m_K \left(f_K + \sum_{i=1}^{I} z_i u^n_{i,K} \right), \qquad K \in \mathcal{T}, \tag{10}$$

where f_K is (possibly an approximation of) the mean value of f on the cell K, and where

$$\phi^n_{K\sigma} = \begin{cases} \phi^n_L & \text{if } \sigma = K|L \in \mathcal{E}_{\text{int}}, \\ \phi^n_K & \text{if } \sigma \subset \Gamma^N, \\ \phi^D_\sigma = \frac{1}{m_\sigma}\int_\sigma \phi^D & \text{if } \sigma \subset \Gamma^D. \end{cases}$$

The Eq. (1) is discretized using a backward Euler method in time and finite volumes in space, leading to

$$\frac{u^n_{i,K} - u^{n-1}_{i,K}}{\tau^n} m_K + \sum_{\sigma\in\mathcal{E}_K} F^n_{i,K\sigma} = 0, \quad i = 1, \ldots, m, \quad K \in \mathcal{T}. \tag{11}$$

In accordance with (6), we set $F^n_{K\sigma} = 0$ if $\sigma \subset \partial\Omega$. For $\sigma = K|L$ an internal edge, then we define

$$F^n_{i,K\sigma} = a_\sigma D_i \left(u^n_{i,K} u^n_{0,L} \mathfrak{B}\left(z_i(\phi^n_L - \phi^n_K)\right) - u^n_{i,L} u^n_{0,K} \mathfrak{B}\left(z_i(\phi^n_K - \phi^n_L)\right)\right), \tag{12}$$

with

$$u^n_{0,K} = 1 - \sum_{i=1}^{I} u^n_{i,K}, \qquad K \in \mathcal{T}. \tag{13}$$

Formula (12) involves a function $\mathfrak{B} \in C^1(\mathbb{R}; \mathbb{R})$ which is (strictly) positive and satisfies $\mathfrak{B}(0) = 1$ and $\mathfrak{B}'(0) = -1/2$.

The continuous system (1) and (2) was originally derived in [3] thanks to a hopping process, suggesting the choice

$$\mathfrak{B}(y) = e^{-y/2}, \tag{SQRA}$$

leading to a scheme referred to as the square-root approximation (SQRA) scheme in what follows, in reference to [6, 13, 15]. Another natural choice for the function $\mathfrak{B}$ is the Bernoulli function

$$\mathfrak{B}(y) = \frac{y}{e^y - 1}, \tag{SG}$$

the corresponding scheme being referred to as the Scharfetter-Gummel (SG) scheme although its construction is not based on the original idea of [16]. We rather take advantage of the free-energy diminishing character of the SG scheme highlighted in [8].

In order to close the system, it remains to define the discrete counterpart to u_0 as follows:

$$u_{i,K}^0 = \frac{1}{m_K}\int_K u_i^0, \qquad K \in \mathcal{T},\ i = 0, \dots, I. \tag{14}$$

Then we infer from (7) that

$$\sum_{i=0}^{I} u_{i,K}^0 = 1 \text{ for all } K \in \mathcal{T}, \text{ and } \sum_{K\in\mathcal{T}} u_{i,K}^0 m_K = \int_\Omega u_i^0 > 0 \quad \text{for } i = 0, \dots, I. \tag{15}$$

In what follows, we denote by $U_K^n = \left(u_{i,K}^n\right)_{i=0,\dots,I}$ for $K \in \mathcal{T}$ and $n \geq 0$.

3 Stability and Convergence Properties of the Schemes

The goal of this section is to show that the nonlinear system corresponding to the scheme (10)–(13) admits at least one solution, and that beyond local conservativity, this solution preserves at the discrete level some key features of the model, namely the positivity of the volume fractions and the decay of the free energy. The grid $\mathcal{T}$ and the time steps $(\tau^n)_{n\geq 1}$ remain fixed.

Since our scheme is locally conservative, i.e., $F_{K\sigma}^n + F_{L\sigma}^n = 0$ for all $\sigma = K|L \in \mathcal{E}_{\text{int}}$, then summing (11) over K shows by induction and thanks to (14) that

$$\sum_{K\in\mathcal{T}} u_{i,K}^n m_K = \sum_{K\in\mathcal{T}} u_{i,K}^{n-1} m_K = \sum_{K\in\mathcal{T}} u_{i,K}^0 m_K = \int_\Omega u_i^0 > 0. \tag{16}$$

Since we are interested in discrete solutions with positive volume fractions $u_{i,K}^n$, we perform an eventually harmless modification of the flux formula (12) into

$$\begin{aligned} F_{i,K\sigma}^n = a_\sigma D_i \Big(& \left(u_{i,K}^n\right)^+ \left(u_{0,L}^n\right)^+ \mathfrak{B}\left(z_i(\phi_L^n - \phi_K^n)\right) \\ & - \left(u_{i,L}^n\right)^+ \left(u_{0,K}^n\right)^+ \mathfrak{B}\left(z_i(\phi_K^n - \phi_L^n)\right) \Big). \end{aligned} \tag{17}$$

Proposition 31 *Let $n \geq 1$, and let $\left(U_K^{n-1}\right)_{K\in\mathcal{T}}$ be such that*

$$u_{i,K}^{n-1} \geq 0, \quad \sum_{i=0}^{I} u_{i,K}^{n-1} = 1 \quad \forall K \in \mathcal{T}, \quad \text{and} \quad \sum_{K\in\mathcal{T}} u_{i,K}^{n-1} m_K > 0. \tag{18}$$

Then any solution $\left(U_K^n, \phi_K^n\right)_{K\in\mathcal{T}, n\geq 1}$ to the modified scheme with (17) instead of (12) satisfies $u_{i,K}^n > 0$ for all $i = 0, \ldots, I$ and all $K \in \mathcal{T}$.

Proof Let us start by establishing the positivity of $u_{0,K}^n$. Assume for contradiction that there exists a cell $K \in \mathcal{T}$ such that $u_{0,K}^n \leq 0$. Then we deduce from formula (17) that $F_{i,K\sigma}^n \geq 0$ for all $\sigma \in \mathcal{E}_K$ and all $i = 1, \ldots, I$. Because of (13) and (18), this implies that

$$0 \geq u_{0,K}^n = u_{0,K}^{n-1} + \frac{\tau^n}{m_K} \sum_{i=1}^{I} \sum_{\sigma\in\mathcal{E}_K} F_{i,K\sigma}^n \geq 0.$$

In particular, all the fluxes $F_{i,K\sigma}^n$, $i = 1, \ldots, I$ and $\sigma \in \mathcal{E}_K$ are equal to 0. In view of formula (17) and of the strict positivity of $\mathfrak{B}$, this implies either that $u_{i,K}^n \leq 0$ for all i, which yields a contradiction with (13), or that $u_{0,L}^n \leq 0$ for all the cells L sharing an edge $\sigma = K|L$ with K. Since Ω is connected, one would obtain that $u_{0,K}^n = 0$ for all $K \in \mathcal{T}$ and thus that $\sum_{K\in\mathcal{T}} u_{0,K}^n m_K = 0$. This contradicts (16), and thus we necessarily have that $u_{0,K}^n > 0$ for all $K \in \mathcal{T}$.

With the positivity of $u_{0,K}^n$, $K \in \mathcal{T}$, at hand, let us focus on the $u_{i,K}^n$ for an arbitrary $i = 1, \ldots, I$. Similarly, we assume that there exists some $K \in \mathcal{T}$ such that $u_{i,K}^n \leq 0$. Then owing to (17), we infer that $F_{i,K\sigma}^n \leq 0$ for all $\sigma \in \mathcal{E}_K$, and then that

$$0 \geq u_{i,K}^n = u_{i,K}^{n-1} - \frac{\tau^n}{m_K} \sum_{\sigma\in\mathcal{E}_K} F_{i,K\sigma}^n \geq 0.$$

This leads to $u_{i,K}^n = 0$ and to $F_{i,K\sigma}^n = 0$ for all $\sigma \in \mathcal{E}_K$. Since we already know that $u_{0,K}^n > 0$, we deduce from (17) that $u_{i,L}^n \leq 0$ for all cell L sharing a cell $\sigma = K|L$ with K. As above, this implies that $u_{i,K}^n = 0$ for all $K \in \mathcal{T}$, which contradicts (16). Then $u_{i,K}^n > 0$ for all $K \in \mathcal{T}$, concluding the proof of Proposition 31.

A consequence of previous proposition is that a solution to the modified scheme with (17) instead of (12) is also a solution to the original scheme (10)–(13). We did assume that the background charge density f and thus its discrete counterpart $(f_K)_{K\in\mathcal{T}}$ are uniformly bounded, and that ϕ^D belongs to $L^\infty \cap H^{1/2}(\Gamma^D)$. Therefrom, we deduce some uniform discrete $L^\infty(H^1(\Omega))$ estimate on $(\phi_K)_{K\in\mathcal{T}}$ from [9, Lemma 13.4], while [7, Proposition A.1] gives a uniform bound

$$|\phi_K^n| \leq C, \qquad K \in \mathcal{T},\ n \geq 0, \tag{19}$$

since the right-hand side of the discrete Poisson equation (10) is uniformly bounded. These a priori estimates are sufficient to prove the existence of a solution to the scheme thanks to a topological degree argument we do not detail here. We end up with the following proposition.

Proposition 32 *There exists at least one solution to the numerical scheme* (10)–(13) *such that $u^n_{i,K} > 0$ for all $i = 0, \dots, I$, for all $K \in \mathcal{T}$ and all $n \geq 1$.*

Next proposition is about the thermodynamical consistency of our scheme and the decay of a discrete counterpart of the free energy.

Proposition 33 *Let $\left(U^n_K, \phi^n_K\right)_{K\in\mathcal{T},n\geq 1}$ be a solution to the scheme* (10)–(13) *as in Proposition 32, then define for $n \geq 0$ the discrete free energy at the n^{th} time step*

$$\mathcal{H}^n_{\mathcal{T}} = \sum_{K\in\mathcal{T}} m_K H(U^n_K) + \frac{\lambda^2}{2}\sum_{\sigma\in\mathcal{E}} a_\sigma(\phi^n_K - \phi^n_{K\sigma})^2 + \lambda^2 \sum_{\sigma\in\mathcal{E}^D} a_\sigma \phi^D_\sigma(\phi^n_K - \phi^D_\sigma), \tag{20}$$

the discrete electrochemical potentials $\mu^n_{i,K} = \log\left(\frac{u^n_{i,K}}{u^n_{0,K}}\right) + z_i\phi^n_K$ of species i, and

$$\mathcal{D}^n_{\mathcal{T}} = \sum_{i=1}^{I}\sum_{\sigma\in\mathcal{E}_{int}} F^n_{i,K\sigma}(\mu^n_{i,K} - \mu^n_{i,L})$$

the discrete dissipation, which is nonnegative for both choices (SQRA) and (SG) of function $\mathfrak{B}$. Then there holds

$$\mathcal{H}^n_{\mathcal{T}} + \tau^n\mathcal{D}^n_{\mathcal{T}} \leq \mathcal{H}^{n-1}_{\mathcal{T}}, \qquad n \geq 1. \tag{21}$$

Proof With both choices (SQRA) and (SG) for the function $\mathfrak{B}$, the fluxes (12) enter the framework of the exponentially fitted schemes. Indeed, denoting by

$$w^n_{i,K} = \frac{u^n_{i,K}}{u^n_{0,K}} e^{z_i\phi^n_K} = \exp(\mu^n_{i,K}) \quad \text{for } K \in \mathcal{T} \text{and } i = 1, \dots, I$$

(which is well defined since $u^n_{0,K} > 0$), then the fluxes (12) can be reformulated as

$$F^n_{i,K\sigma} = a_\sigma\, D_i\, u^n_{0,K}\, u^n_{0,L} \mathfrak{M}(e^{-z_i\phi^n_K}, e^{-z_i\phi^n_L})\left(w^n_{i,K} - w^n_{i,L}\right) \tag{22}$$

for some mean function $\mathfrak{M}$ depending on the choice of $\mathfrak{B}$ (see [14]). More precisely,

$$\mathfrak{M}(a, b) = \sqrt{ab} \quad \text{for (SQRA)}, \quad \text{and} \quad \mathfrak{M}(a, b) = \frac{\log(1/a) - \log(1/b)}{1/a - 1/b} \quad \text{for (SG)},$$

for $a, b > 0$ with $a \neq b$, and $\mathfrak{M}(a, a) = a$. As a consequence of the positivity of $u^n_{0,K}$ and of the monotonicity of the exponential function, one easily infers that

$$\mathcal{D}^n_{i,\sigma} := F^n_{i,K\sigma}(\mu^n_{i,K} - \mu^n_{i,L}) \geq 0, \qquad \forall i = 1, \ldots, I,\ \sigma = K|L \in \mathcal{E}_{\text{int}}, \tag{23}$$

whence the nonnegativity of $\mathcal{D}^n$.

Define by $\mu^n_{i,K} = \log\left(\frac{u^n_{i,K}}{u^n_{0,K}}\right) + z_i \phi^n_K = \log(w^n_{i,K})$ the electrochemical potential of species i, then multiplying the discrete conservation law (11) by $\tau^n \mu^n_{i,K}$, and summing over $i = 1, \ldots, I$ and $K \in \mathcal{T}$ provides thanks to discrete integration by parts

$$\mathcal{A}^n_{\mathcal{T}} + \mathcal{B}^n_{\mathcal{T}} + \tau^n \mathcal{D}^n_{\mathcal{T}} = 0, \tag{24}$$

where we have set

$$\begin{aligned}\mathcal{A}^n_{\mathcal{T}} &= \sum_{i=1}^{I} \sum_{K\in\mathcal{T}} \left(u^n_{i,K} - u^{n-1}_{i,K}\right) \log\left(\frac{u^n_{i,K}}{u^n_{0,K}}\right) m_K \\ &\overset{(13)}{=} \sum_{i=0}^{I} \sum_{K\in\mathcal{T}} \left(u^n_{i,K} - u^{n-1}_{i,K}\right) \log\left(u^n_{i,K}\right) m_K,\end{aligned}$$

and

$$\begin{aligned}\mathcal{B}^n_{\mathcal{T}} &= \sum_{i=1}^{I} \sum_{K\in\mathcal{T}} \left(u^n_{i,K} - u^{n-1}_{i,K}\right) z_i \phi^n_K m_K \\ &\overset{(10)}{=} \lambda^2 \sum_{K\in\mathcal{T}} \phi^n_K \sum_{\sigma\in\mathcal{E}_K} a_\sigma \left(\phi^n_K - \phi^{n-1}_K - (\phi^n_{K\sigma} - \phi^{n-1}_{K\sigma})\right).\end{aligned}$$

Then we deduce from the convexity of H that

$$\mathcal{A}^n_{\mathcal{T}} \geq \sum_{K\in\mathcal{T}} \left(H(U^n_K) - H(U^{n-1}_K)\right) m_K, \tag{25}$$

while reorganizing the term $\mathcal{B}^n$ gives

$$\begin{aligned}\mathcal{B}^n_{\mathcal{T}} = \lambda^2 \sum_{\sigma\in\mathcal{E}} a_\sigma \left(\phi^n_K - \phi^{n-1}_K - (\phi^n_{K\sigma} - \phi^{n-1}_{K\sigma})\right)(\phi^n_K - \phi^n_{K\sigma}) \\ + \lambda^2 \sum_{\sigma\in\mathcal{E}^D} a_\sigma \phi^D_\sigma (\phi^n_K - \phi^{n-1}_K).\end{aligned}$$

Then using the elementary convexity inequality $a(a-b) \geq (a^2 - b^2)/2$ in the above term and combining the result with (25) in (24) provides the desired result (21).

Proposition 33 is interesting in itself, but it also contains important information for proving the convergence of the scheme, as in particular the discrete $L^2_{\text{loc}}(H^1)$ estimates on the discrete counterparts of u_0 and $\sqrt{u_i u_0}$. We prove these estimates in

Lemma 35. As an intermediate result we need a uniform bound on the discrete free energy.

Lemma 34 *There exists $C > 0$ depending only on Ω, ϕ^D, λ, f, $(z_i)_i$, and ζ such that, for all $N \geq 1$, there holds $|\mathcal{H}_{\mathcal{T}}^N| \leq C$.*

Proof Because of the bound $0 \leq u_{i,K}^n \leq 1$ for all i and K, it is clear that the first two contributions of (20) remain uniformly bounded. Concerning the last contribution observe that if one defines ϕ_K^D and ϕ_σ^D as the averages of ϕ^D on $K \in \mathcal{T}$ and $\sigma \in \mathcal{E}$ respectively, then

$$\sum_{\sigma\in\mathcal{E}^D} a_\sigma \phi_\sigma^D (\phi_K^n - \phi_\sigma^D)$$
$$= \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}^D} a_\sigma (\phi_\sigma^D - \phi_K^D)(\phi_K^n - \phi_{K\sigma}^n) + \sum_{K\in\mathcal{T}} \phi_K^D \sum_{\sigma\in\mathcal{E}^D} a_\sigma (\phi_K^n - \phi_{K\sigma}^n)$$

which yields using Young's inequality for the first term and the Poisson equation for the second one that

$$|\lambda^2 \sum_{\sigma\in\mathcal{E}^D} a_\sigma \phi_\sigma^D (\phi_K^n - \phi_\sigma^D)|$$
$$\leq C(\|\nabla\phi^D\|_{L^2}^2 + \|\phi^D\|_{L^1}(1 + \|f\|_{L^\infty})) + \frac{\lambda^2}{4} \sum_{\sigma\in\mathcal{E}} a_\sigma (\phi_K^n - \phi_{K\sigma}^n)^2$$

for some C depending only on the domain, λ, ζ and $(z_i)_i$.

Lemma 35 *There exists $C > 0$ depending only on Ω, ϕ^D, λ, f, $(z_i)_i$, $(D_i)_i$ and ζ such that, for all $N \geq 1$, there holds*

$$\sum_{n=1}^N \tau^n \sum_{i=1}^I \sum_{\sigma\in\mathcal{E}_{int}} a_\sigma \left(\sqrt{u_{i,K}^n u_{0,K}^n} - \sqrt{u_{i,L}^n u_{0,L}^n}\right)^2$$
$$+ \sum_{n=1}^N \tau^n \sum_{\sigma\in\mathcal{E}_{int}} a_\sigma \left(\sqrt{u_{0,K}^n} - \sqrt{u_{0,L}^n}\right)^2$$
$$+ \sum_{n=1}^N \tau^n \sum_{\sigma\in\mathcal{E}_{int}} a_\sigma \left(u_{0,K}^n - u_{0,L}^n\right)^2 \leq C(1 + \sum_{n=1}^N \tau^n).$$

Proof One gets from the elementary inequality $(a - b)(\log(a) - \log(b)) \geq 4(\sqrt{a} - \sqrt{b})^2$ applied to (23) that

$$\mathcal{D}_{i,\sigma}^n \geq 4 a_\sigma D_i \mathfrak{R}(e^{-z_i\phi_K^n}, e^{-z_i\phi_L^n})$$
$$\times \left(\sqrt{u_{i,K}^n u_{0,L}^n} e^{\frac{z_i}{4}(\phi_K^n - \phi_L^n)} - \sqrt{u_{i,L}^n u_{0,K}^n} e^{\frac{z_i}{4}(\phi_L^n - \phi_K^n)}\right)^2$$

with $\mathfrak{R}(e^{-z_i\phi^n_K}, e^{-z_i\phi^n_L}) = \mathfrak{M}(e^{-z_i\phi^n_K}, e^{-z_i\phi^n_L})e^{\frac{z_i}{2}(\phi^n_K+\phi^n_L)}$ being equal to 1 for the choice (SQRA) of $\mathfrak{B}$ but not for (SG). However, thanks to (19) and since $D_i > 0$ for all i, there holds

$$2D_i\mathfrak{R}(e^{-z_i\phi^n_K}, e^{-z_i\phi^n_L}) \geq \kappa$$

for some $\kappa > 0$ uniform w.r.t. K, i and n. As a consequence, using furthermore that $(a+b)^2 \geq \frac{1}{2}a^2 - b^2$,

$$\begin{aligned}\mathcal{D}^n_{i,\sigma} \geq &\kappa a_\sigma \cosh^2\left(\frac{z_i}{4}(\phi^n_K - \phi^n_L)\right)\left(\sqrt{u^n_{i,K}u^n_{0,L}} - \sqrt{u^n_{i,L}u^n_{0,K}}\right)^2 \\ &- \kappa a_\sigma \left(\sqrt{u^n_{i,K}u^n_{0,L}} + \sqrt{u^n_{i,L}u^n_{0,K}}\right)^2 \sinh^2\left(\frac{z_i}{4}(\phi^n_K - \phi^n_L)\right).\end{aligned}$$

Since $|\phi^n_K| \leq C$ owing to (19), one has $\sinh^2\left(\frac{z_i}{4}(\phi^n_K - \phi^n_L)\right) \leq C(\phi^n_K - \phi^n_L)^2$. Using moreover that $0 < u^n_{i,K}, u^n_{0,K} < 1$ and that $\cosh(a) \geq 1$, one gets that

$$\mathcal{D}^n_{i,\sigma} \geq a_\sigma\kappa\left(\sqrt{u^n_{i,K}u^n_{0,L}} - \sqrt{u^n_{i,L}u^n_{0,K}}\right)^2 - Ca_\sigma(\phi^n_K - \phi^n_L)^2.$$

Since

$$\begin{aligned}\left(\sqrt{u^n_{i,K}u^n_{0,L}} - \sqrt{u^n_{i,L}u^n_{0,K}}\right)^2 = &\left(\sqrt{u^n_{i,K}u^n_{0,K}} - \sqrt{u^n_{i,L}u^n_{0,L}}\right)^2 \\ &- (u^n_{i,K} - u^n_{i,L})(u^n_{0,K} - u^n_{0,L}),\end{aligned}$$

then summing over $i = 1, \dots, I$ and $\sigma \in \mathcal{E}_{\text{int}}$ and using (13) leads to

$$\begin{aligned}\mathcal{D}^n_{\mathcal{T}} \geq &\kappa \sum_{i=1}^{I}\sum_{\sigma\in\mathcal{E}_{\text{int}}} a_\sigma\left(\sqrt{u^n_{i,K}u^n_{0,K}} - \sqrt{u^n_{i,L}u^n_{0,L}}\right)^2 \\ &+ \kappa\sum_{\sigma\in\mathcal{E}_{\text{int}}} a_\sigma\left(u^n_{0,K} - u^n_{0,L}\right)^2 - C\sum_{\sigma\in\mathcal{E}_{\text{int}}} a_\sigma(\phi^n_K - \phi^n_{K\sigma})^2.\end{aligned}$$

Invoking again the arguments developed in the discussion preceding Proposition 32 to get a uniform discrete $L^\infty(H^1)$ estimate on $\left(\phi^n_K\right)_{K,n}$, we obtain that

$$\begin{aligned}\mathcal{D}^n_{\mathcal{T}} \geq &\kappa \sum_{i=1}^{I}\sum_{\sigma\in\mathcal{E}_{\text{int}}} a_\sigma\left(\sqrt{u^n_{i,K}u^n_{0,K}} - \sqrt{u^n_{i,L}u^n_{0,L}}\right)^2 \\ &+ \kappa\sum_{\sigma\in\mathcal{E}_{\text{int}}} a_\sigma\left(u^n_{0,K} - u^n_{0,L}\right)^2 - C.\end{aligned} \tag{26}$$

Moreover, the inequality $\sum_{i=0}^{I}\sqrt{u^n_{i,K}u^n_{i,L}} \leq 1$ gives that

$$\begin{aligned}
&\sum_{i=1}^{I}\sum_{\sigma\in\mathcal{E}_{\mathrm{int}}} a_\sigma\left(\sqrt{u_{i,K}^n u_{0,K}^n}-\sqrt{u_{i,L}^n u_{0,L}^n}\right)^2\\
&\quad\geq \sum_{\sigma\in\mathcal{E}_{\mathrm{int}}} a_\sigma\left((1-u_{0,K}^n)u_{0,K}^n+(1-u_{0,L}^n)u_{0,L}^n-2(1-\sqrt{u_{0,K}^n u_{0,L}^n})\sqrt{u_{0,K}^n u_{0,L}^n}\right)\\
&\quad= \sum_{\sigma\in\mathcal{E}_{\mathrm{int}}} a_\sigma\left(\sqrt{u_{0,K}^n}-\sqrt{u_{0,L}^n}\right)^2-\sum_{\sigma\in\mathcal{E}_{\mathrm{int}}} a_\sigma\left(u_{0,K}^n-u_{0,L}^n\right)^2,
\end{aligned}$$

whence we also deduce that

$$\mathcal{D}_{\mathcal{T}}^n \geq \kappa\sum_{\sigma\in\mathcal{E}_{\mathrm{int}}} a_\sigma\left(\sqrt{u_{0,K}^n}-\sqrt{u_{0,L}^n}\right)^2 - C.$$

To conclude the proof, it eventually remains to remark from (21) and Lemma 34 that there exists C depending neither on h, Δt, N nor on the initial data $U^0=\left(u_i^0\right)_{0\leq i\leq I}$ (provided it fulfills (7)) such that $\sum_{n=1}^{N}\tau^n\mathcal{D}_{\mathcal{T}}^n\leq C$. Combining this with (26) yields the desired result.

One also deduces the following discrete $L^2_{\mathrm{loc}}(L^2)^d$ estimates on the fluxes, which amount to some discrete $L^2_{\mathrm{loc}}(H^1)'$ estimate on time increments of the discrete counterpart to $\partial_t u_i$.

Lemma 36 *There exists C depending only on Ω, ϕ^D, λ, f, $(z_i)_i$, $(D_i)_i$ and ζ such that*

$$\sum_{i=1}^{I}\sum_{n=1}^{N}\sum_{\sigma\in\mathcal{E}_{int}}\frac{d_\sigma}{m_\sigma}\left|F_{i,K\sigma}^n\right|^2\leq C(1+\sum_{n=1}^{N}\tau^n). \tag{27}$$

Proof One splits the flux (2) into two parts corresponding to convection and diffusion respectively:

$$F_{i,K\sigma}^n = F_{i,K\sigma}^{\mathrm{conv},n}+F_{i,K\sigma}^{\mathrm{diff},n},$$

with

$$\begin{aligned}
F_{i,K\sigma}^{\mathrm{conv},n} &= a_\sigma D_i\frac{u_{i,K}^n u_{0,L}^n+u_{i,L}^n u_{0,K}^n}{2}\left[\mathfrak{B}\big(z_i(\phi_L^n-\phi_K^n)\big)-\mathfrak{B}\big(z_i(\phi_K^n-\phi_L^n)\big)\right],\\
F_{i,K\sigma}^{\mathrm{diff},n} &= a_\sigma D_i\frac{u_{i,K}^n u_{0,L}^n-u_{i,L}^n u_{0,K}^n}{2}\left[\mathfrak{B}\big(z_i(\phi_L^n-\phi_K^n)\big)+\mathfrak{B}\big(z_i(\phi_K^n-\phi_L^n)\big)\right].
\end{aligned}$$

The flux $(F_{i,K\sigma}^n)_{\sigma,n}$ is bounded in $L^2_{\mathrm{loc}}(L^2)^d$ in the sense of (27) if both $(F_{i,K\sigma}^{\mathrm{conv},n})_{\sigma,n}$ and $(F_{i,K\sigma}^{\mathrm{diff},n})_{\sigma,n}$ are. For the choice (SG) of the function $\mathfrak{B}$, then $\mathfrak{B}(-y)-\mathfrak{B}(y)=y$, while $\mathfrak{B}(-y)-\mathfrak{B}(y)=y+\mathcal{O}(y^2)$ for (SQRA), so that

$$F_{i,K\sigma}^{\mathrm{conv},n} = a_\sigma D_i\frac{u_{i,K}^n u_{0,L}^n+u_{i,L}^n u_{0,K}^n}{2}z_i(\phi_K^n-\phi_L^n)+\mathcal{O}\left(a_\sigma(\phi_K^n-\phi_L^n)^2\right), \tag{28}$$

the remainder term being null for (SG). The $L^2_{\rm loc}(L^2)^d$ character of the above expression directly follows from the uniform bound on $u^n_{i,K}$, $0 \le i \le I$ and from the discrete $L^\infty(H^1)$ bound on $(\phi^n_K)_{K,n}$ inherited from the control of the energy $\mathcal{H}^n_{\mathcal{T}}$, to be combined with (19) to control the remainder term.

Concerning the diffusive term, one has for both choices (SQRA) and (SG) of the function $\mathfrak{B}$ that

$$1 \le \frac{1}{2}\left[\mathfrak{B}\big(z_i(\phi^n_L - \phi^n_K)\big) + \mathfrak{B}\big(z_i(\phi^n_K - \phi^n_L)\big)\right] \le 1 + \mathcal{O}\left(\left(\phi^n_K - \phi^n_L\right)^2\right).$$

Therefore, one gets that

$$F^{\mathrm{diff},n}_{i,K\sigma} = a_\sigma D_i \left(u^n_{i,K}u^n_{0,L} - u^n_{i,L}u^n_{0,K}\right)\left(1 + \mathcal{O}\left(\left(\phi^n_K - \phi^n_L\right)^2\right)\right). \tag{29}$$

Since $u^n_{i,K}u^n_{0,L} - u^n_{i,L}u^n_{0,K} = u^n_{i,K}u^n_{0,K} - u^n_{i,L}u^n_{0,L} + (u^n_{i,K} + u^n_{i,L})(u^n_{0,L} - u^n_{0,K})$, Lemma 35 provides the desired $L^2_{\rm loc}(L^2)$ bound on $F^{\mathrm{diff},n}_{i,K\sigma}$, hence Lemma 36.

The above estimates are sufficient to establish the convergence of the numerical scheme. For a given mesh $\mathcal{T}$ and a given time discretization $\tau = (\tau^n)_{n\ge1}$, we denote by $u_{i,\mathcal{T},\tau}$ and $\phi_{\mathcal{T},\tau}$ the piecewise constant reconstructions defined by

$$u_{i,\mathcal{T},\tau}(t,x) = u^n_{i,K} \quad \text{and} \quad \phi_{\mathcal{T},\tau}(t,x) = \phi^n_K \quad \text{if } (t,x) \in K \times (t^{n-1}, t^n].$$

Theorem 37 *Let $(\mathcal{T}_\ell)_{\ell\ge1}$ be a sequence of admissible discretizations of Ω (satisfying the orthogonality condition), such that h_ℓ goes to 0 as ℓ tends to $+\infty$ while the mesh regularity factor ζ_ℓ remains bounded uniformly w.r.t. ℓ, and let $(\tau_\ell)_{\ell\ge1} = \left(\left(\tau^n_\ell\right)_{n\ge1}\right)_{\ell\ge1}$ be a sequence of sequences of time steps such that $\Delta t_\ell = \max_n \tau^n_\ell$ goes to 0 as ℓ tends to $+\infty$. Then there exists a weak solution (U,ϕ) such that, up to a subsequence,*

$$\phi_{\mathcal{T},\tau} \underset{h,\Delta t\to0}{\longrightarrow} \phi \quad \textit{in the } L^\infty(\mathbb{R}_+ \times \Omega)\textit{-weak-}\star \textit{ sense and a.e. in } \mathbb{R}_+ \times \Omega, \tag{30}$$

$$u_{i,\mathcal{T},\tau} \underset{h,\Delta t\to0}{\longrightarrow} u_i \quad \textit{in the } L^\infty(\mathbb{R}_+ \times \Omega)\textit{-weak-}\star \textit{ sense}, \qquad i = 0,\dots,M, \tag{31}$$

with furthermore $u_{0,\mathcal{T},\tau}$ and $u_{i,\mathcal{T},\tau}(u_{0,\mathcal{T},\tau})^{1/2}$ converging a.e. in $\mathbb{R}_+ \times \Omega$ towards their respective limits u_0 and $u_i(u_0)^{1/2}$ which belong to $L^2_{loc}(H^1)$.

The proof is technical and will be detailed in a forthcoming contribution. It borrows ideas to the proof proposed in [4] and relies on compactness arguments (in particular on the degenerate Aubin-Lions lemma [4, Lemma 10]) as well as on a suitable notion of weak solution. Indeed, yet another reformulation of the fluxes is needed, like for instance

$$F_i = -D_i\left(\nabla(u_0u_i) - 4u_i\sqrt{u_0}\,\nabla\sqrt{u_0} + u_iu_0z_i\nabla\phi\right).$$

This last formulation is suitable to establish the convergence since it clearly belongs to $L^2_{\text{loc}}(L^2)^d$ as the product of gradient terms the approximation of which being weakly convergent in $L^2_{\text{loc}}(L^2)^d$ with bounded zeroth order term the approximation of which being strongly convergent.

4 Numerical Results

The nonlinear system corresponding to the scheme is solved thanks to a Newton-Raphson method with stopping criterion $\|\mathcal{F}^n_{\mathcal{T}}((U^n_K)_{K\in\mathcal{T}}, (\phi^n_K)_{K\in\mathcal{T}})\|_\infty \leq 10^{-8}$, the components of $\mathcal{F}^n_{\mathcal{T}}$ being given by the left-hand side of (11).

The goal of our first numerical test is to show that both schemes corresponding to (SQRA) and (SG) are second order accurate w.r.t. the mesh size. To this end, we consider the one-dimensional domain $\Omega = (0, 1)$, in which $I = 2$ different ions evolve, both with the same diffusion coefficient $D_1 = D_2 = 1$. Their (normalized) charge is set to $z_1 = 2$ and $z_2 = 1$, yielding repulsive interaction. No background charge is considered, i.e. $f = 0$, whereas Dirichlet boundary conditions are imposed for the electric potential on both sides of the interval, that are $\phi^D(t, 0) = 10$ and $\phi^D(t, 1) = 0$. We consider a moderately small Debye length $\lambda^2 = 10^{-2}$. We start at initial time $t = 0$ with the following configurations: $u^0_1(x) = 0.2 + 0.1(x - 1)$ and $u^0_2 \equiv 0.4$.

A reference solution is computed on a grid made of 1638400 cells and with a constant time step $\tau = 10^{-3}$, to which are compared solutions computed on successively refined grids but with the same constant time step. The profile of the solution at the final time $T = 1$ is depicted on Figs. 1 and 2. The relative space-time L^1 error is plotted as a function of the number of cells on Fig. 3, showing some second order accuracy in space, as specified in the introductory discussion. For such a moderately small value of $\lambda^2 = 10^{-2}$, both schemes exhibit a very similar behavior in terms of

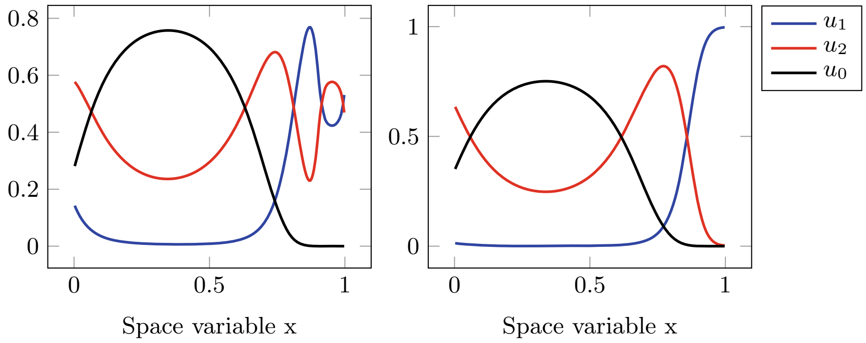


Fig. 1 Concentration profiles $u_1(T, x)$, $u_2(T, x)$ and $u_0(T, x)$ at times $T = 1$ (left) and $T = 5000$ (right)

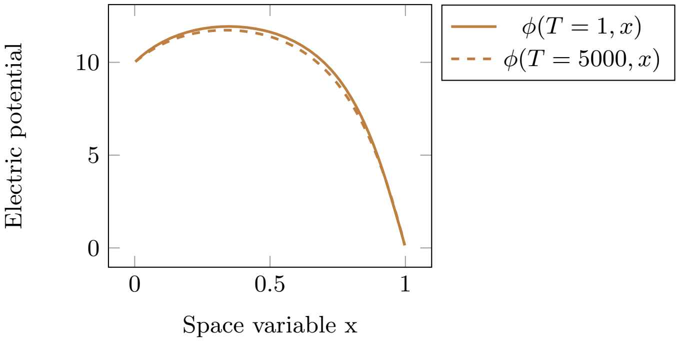
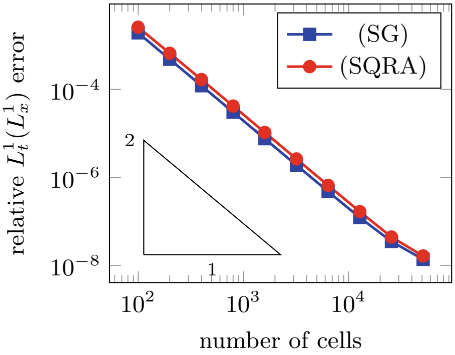


Fig. 2 Electric potential profile $\phi(T, x)$ at times $T = 1$ (solid) and $T = 5000$ (dashed)

Fig. 3 Convergence of the schemes under space grid refinement



accuracy, but also in terms of nonlinear resolution. More precisely, the number of Newton iterations required to solve a time step remains between 6 for the very first iterations and 2 for larger times is mainly insensitive to the mesh size.

Nevertheless, there is an important difference in the numerical behavior of the two schemes in the small Debye length regime. Indeed, when λ^2 become small, then excepted for very particular values of the data, the variations of $\phi_{\mathcal{T},\tau}$ across the interfaces $\mathcal{E}$ become very large because of (10). Therefore, the drift becomes too large to evaluate its exponential, making the computation with the (SQRA) scheme fail. Since $B(y) \sim -y$ as y tends to $-\infty$, the situation is much less problematic with the (SG) scheme, for which computation of the solution corresponding to $\lambda = 10^{-6}$ is feasible without any specific treatment. However, since the drift becomes large, the use of a reduce time step is required to ensure the convergence of Newton's methods.

The long-time limit of the continuous model has been exhibited in [3]. The model reduces to a nonlinear elliptic equation on the electric potential ϕ, from which one deduces the concentration profiles. However, no quantitative estimate concerning the convergence towards equilibrium. We then perform a numerical study still with the same parameters as previously (in particular with $\lambda^2 = 10^{-2}$). The steady solution is computed by choosing a very large final time $T_\infty = 5.10^5$ in the simulation. We

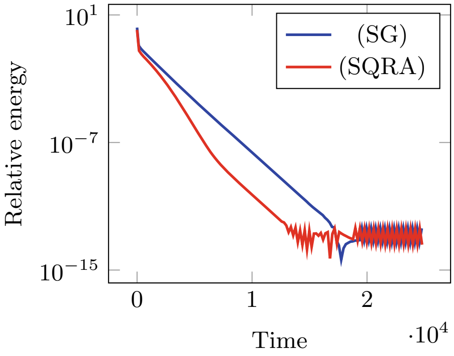


Fig. 4 Convergence towards the steady long-time behavior in terms of relative energy $\mathcal{H}_{\mathcal{T}}^{\mathrm{rel},n}$

denote by $\mathcal{H}_{\mathcal{T}}^{\infty}$ the corresponding discrete free energy. The relative energy at time t^n is then defined as $\mathcal{H}_{\mathcal{T}}^{\mathrm{rel},n} = \mathcal{H}_{\mathcal{T}}^{n} - \mathcal{H}_{\mathcal{T}}^{\infty}$. The energy decay stated in Proposition 33 ensures that $\mathcal{H}_{\mathcal{T}}^{\mathrm{rel},n} \geq 0$ up to numerical errors related to the resolution of the nonlinear systems. One observes on Fig. 4 that the (SQRA) scheme dissipates faster energy than the (SG) scheme, the latter exhibiting an almost perfect but rather slow exponential convergence towards the steady state as long as the numerical precision has not been reached. Note that in opposition to the classical Poisson-Nernst-Planck problems arising in semi-conductor physics [2, 10], there is no theoretical foundation to the exponential decay we observe here. The rigorous proof of such an exponential convergence for the continuous and discretized degenerate Poisson-Nernst-Planck problem should be investigated in future works.

Acknowledgements CC and MH acknowledge support from the LabEx CEMPI (ANR-11-LABX-0007) and the ministries of Europe and Foreign Affairs (MEAE) and Higher Education, Research and Innovation (MESRI) through PHC Amadeus 46397PA. AM acknowledges support from the Austrian Science Fund (FWF), grants P33010 and F65, and from the multilateral project of the Austrian Agency for International Cooperation in Education and Research (OeAD), grant FR 01/2021. CC also acknowledges support from the European Union's Horizon 2020 research and innovation programme under grant agreement No 847593 (EURAD program, WP DONUT) and from the COMODO (ANR-19-CE46-0002) project. This work has also received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme, ERC Advanced Grant no. 101018153.

References

1. Bailo, R., Carrillo, J.A., Hu, J.: Bound-preserving finite-volume schemes for systems of continuity equations with saturation. SIAM J. Appl. Math. **83**(3), 1315–1339 (2023). arXiv:2110.08186
2. Bessemoulin-Chatard, M., Chainais-Hillairet, C.: Exponential decay of a finite volume scheme to the thermal equilibrium for drift-diffusion systems. J. Numer. Math. **25**(3), 147–168 (2017)
3. Burger, M., Schlake, B., Wolfram, M.T.: Nonlinear Poisson-Nernst-Planck equations for ion flux through confined geometries. Nonlinearity **25**(4), 961 (2012)

4. Cancès, C., Chainais-Hillairet, C., Gerstenmayer, A., Jüngel, A.: Finite-volume scheme for a degenerate cross-diffusion model motivated from ion transport. Numer. Methods Part. Differ. Equ. **35**(2), 545–575 (2019)
5. Cancès, C., Chainais-Hillairet, C., Merlet, B., Raimondi, F., Venel, J.: Mathematical analysis of a thermodynamically consistent reduced model for iron corrosion. Z. Angew. Math. Phys. **74**, 96 (2023)
6. Cancès, C., Venel, J.: On the square-root approximation finite volume scheme for nonlinear drift-diffusion equations. Comptes Rendus. Mathématique **361**, 525–558 (2023)
7. Cancès, C., Chainais-Hillairet, C., Fuhrmann, J., Gaudeul, B.: A numerical analysis focused comparison of several finite volume schemes for a unipolar degenerated drift-diffusion model. IMA J. Numer. Anal. **41**(1), 271–314 (2021)
8. Chatard, M.: Asymptotic behavior of the Scharfetter-Gummel scheme for the drift-diffusion model. In: Finite volumes for complex applications VI. Probl. Perspect. **1**, **2**. Springer Proceedings in Mathematics, vol. 4, pp. 235–243. Springer, Heidelberg (2011)
9. Eymard, R., Gallouët, T., Herbin, R.: Finite volume methods. In: Handbook of Numerical Analysis, vol. VII, pp. 713–1020. North-Holland, Amsterdam (2000)
10. Gajewski, H., Gröger, K.: Semiconductor equations for variable mobilities based on Boltzmann statistics or Fermi-Diracs statistics. Math. Nachr. **140**, 7–36 (1989)
11. Gerstenmayer, A., Jüngel, A.: Analysis of a degenerate parabolic cross-diffusion system for ion transport. J. Math. Anal. Appl. **461**(1), 523–543 (2018)
12. Gerstenmayer, A., Jüngel, A.: Comparison of a finite-element and finite-volume scheme for a degenerate cross-diffusion system for ion transport. Comput. Appl. Math. **38**(3), Article 108, 23 (2019)
13. Heida, M.: Convergences of the squareroot approximation scheme to the Fokker-Planck operator. Math. Models Methods Appl. Sci. **28**(13), 2599–2635 (2018)
14. Heida, M., Kantner, M., Stephan, A.: Consistency and convergence of a family of finite volume discretizations of the Fokker-Planck operator. ESAIM: Math. Model Numer. Anal. **55**(6), 3017–3042 (2021)
15. Lie, H.C., Fackeldey, K., Weber, M.: A square root approximation of transition rates for a Markov state model. SIAM J. Matrix Anal. Appl. **34**, 738–756 (2013)
16. Scharfetter, D.L., Gummel, H.K.: Large-signal analysis of a silicon read diode oscillator. IEEE Trans. Electron Devices **16**(1), 64–77 (1969)

Magic SIAC Toolbox: A Codebase of Effective, Efficient, and Flexible Filters

Xulia Docampo-Sánchez and Jennifer K. Ryan

Abstract Filtering is a powerful tool in CFD that can aid in accurately and efficiently predicting the governing physics in simulations, leading to improved designs. Filters can remove subgrid scale high-frequency physics so that only large scale structures remain in the filtered solution, alleviate aliasing error, and mitigate Gibbs phenomenon. They can even extract hidden accuracy. The same ideas are useful in data compression, post-processing, and machine learning. Well-designed filters, such as the one that gives rise to the Smoothness-Increasing Accuracy-Conserving (SIAC) post-processing filters, can be used to extract hidden information in certain numerical simulations, creating even more accurate representations of the data. They can be adapted for boundaries, unstructured grids, and non-smooth solutions. Furthermore, well-designed filters have the potential to accurately capture multi-scale physics, and are flexible enough to combine simulation information with experimental data. The SIAC Magic Toolbox provides a codebase for efficient, effective, flexible filters for general data. It takes in two data files: one data file consisting of information on the mesh and a second data file consisting of information from the corresponding approximation, either modal or nodal data. If desired, the user can choose parameters that correlate to the amount of dissipation, accuracy, and scaling. Otherwise these parameters are set as default parameters. The toolbox then returns the filtered information in the same format.

Keywords Filtering · Post-Processing · Superconvergence · Accuracy extraction

X. Docampo-Sánchez
INDOMINUS Advanced Solutions, Parque Tecnológico de Valladares-Vigo, Pontevedra, Spain
e-mail: xulia.docampo@indominus.eu

J. K. Ryan (✉)
Department of Mathematics, KTH Royal Institute of Technology, Lindstedtsvägen 25, 114 28 Stockholm, Sweden
e-mail: jryan@kth.se
URL: https://siac-magic.gitlab.io/web/

E. Franck et al. (eds.), *Finite Volumes for Complex Applications X—Volume 1, Elliptic and Parabolic Problems*, Springer Proceedings in Mathematics & Statistics 432,
https://doi.org/10.1007/978-3-031-40864-9_5

1 Introduction and Background

In this article, we introduce the SIAC Magic Toolbox [7], a software package to enable practical utilization of mathematical filters that allow for revealing hidden information contained in data. These filters have proven useful in streamline and vortex visualization [13], aeroacoustics [22], shock regularization [26], adaptivity [6], and data extraction [19, 20]. Smoothness-Increasing Accuracy-Conserving filtering is a filter family that generalizes a post-processing technique introduced for finite element approximations to elliptic equations by Bramble and Schatz [3] and extended to linear hyperbolic equations and discontinuous Galerkin methods by Cockburn, Luskin, Shu, and Süli [5]. They are based on rich mathematical theory that ties together utilizing moments, Fourier information, and information from dual equations. A summary of the developments of SIAC as well as applications can be found in [8].

While most of the previous developments of SIAC filtering have concentrated on developments in the theory, here we introduce a Julia package that allows for applying SIAC filters. The SIAC Magic Toolbox only requires the data values and mesh information. The user can then test the effect of different moment conditions, smoothness of the filter, and scaling on their data. The toolbox will return the filtered values.

We start by introducing the basics of SIAC filters and then proceed to introduce the SIAC Magic Toolbox, discuss its computational performance, and provide examples generated by the toolbox.

1.1 Smoothness-Increasing Accuracy-Conserving (SIAC) Filters

To illustrate the ability of SIAC to perform on a given data set, it is useful to outline how SIAC works for general data as well as through the error estimates.

Assume that $f_h(\mathbf{x}),\ \mathbf{x} = (x_1, x_2, \dots, x_d)$ is the set of given data that approximates a function $f(\mathbf{x})$ on a discrete set of points that comprise a mesh and where h describes the data spacing. Then the filtered data is given by convolving f_h with a kernel function, $K_H(\cdot)$. In a continuous setting, this is written as

$$f_h^*(\mathbf{x}) = K_H(\mathbf{x}) \star f_h(\mathbf{x}) = \int_{\mathbb{R}} K_H(\mathbf{x}-\mathbf{y}) f_h(\mathbf{y}) d\mathbf{y}. \tag{1}$$

where H represents the kernel scaling. In a discrete setting where only point values are given, the discrete convolution is done using exact quadrature.

The filtered error can be decomposed into a term that only depends on how the kernel is designed (the potential for data extraction) and a term that relies on the discretization error:

$$\|f - f_h^*(\mathbf{x})\| \leq \underbrace{\|f - K_H\,() \star f\|}_{\text{Kernel design}} + \underbrace{\|K_H\,() \star (f - f_h())\|}_{\text{Discretization Error}} \leq \mathcal{O}(H^{r+1}) + \mathcal{O}(h^s) \tag{2}$$

where $\|\cdot\|$ is some norm. The first term in the error estimate is *only* controlled by the filter/post-processor design. The second term is determined by the discretization error and how the filter/post-processor is linked to the discretization error through the scaling. As noted in Mock and Lax [18] and Bramble and Schatz [3], the goal is to obtain an optimal balance between these two terms (and hence h and H). This allows for taking advantage of the patterns of information from the choice of data representation and thus damping nonphysical noise.

In the following section we discuss the formulation and the underlying mechanism in the design of Smoothness-Increasing Accuracy-Conserving (SIAC) filters.

SIAC Formulation. Assume that the kernel is comprised of $r+1$ (scaled) function translates of a given function,

$$K\,(\cdot) = \sum_{\gamma=1}^{r+1} c_\gamma \psi_{\mathbf{T}_\gamma}(\cdot), \tag{3}$$

where $K_H\,(\cdot) = \frac{1}{H} K\,\left(\frac{\cdot}{H}\right)$. $K_H\,(\cdot)$ can be viewed as a normalized probability density function. Here, c_γ are the weights of some kernel basis function, ψ. The weights are obtained by solving a system of equations defined by enforcing consistency and (mechanical/statistical) moments,

$$\int_{\mathbb{R}} K\,(x-y) y^m \, dy = x^m, \qquad m = 0, 1, 2, \ldots, r, \tag{4}$$

as is also done in image processing. Hence, this gives information about the mean, variance, standard deviation, etc. Further, it ensures that the first term is only controlled by the number of moments, *regardless of the choice of kernel basis function,*

$$\|f - K_H\,() \star f\| \leq C_m H^{r+1}.$$

Mock and Lax [18] describe the importance of satisfying moment conditions and pre-processing data in order to recover accuracy for discontinuous functions. The pre-processing of data is important for methods not based on Galerkin orthogonality.

In the Cartesian coordinate system, utilizing a compact kernel basis function can aid in computational efficiency. Hence, ψ is usually chosen to be a B-spline kernel,

$$\boldsymbol{B}_{T,1} = \chi_{[-\frac{1}{2},\frac{1}{2})}, \qquad \boldsymbol{B}_{T,n} = \boldsymbol{B}_{T,n-1} \star \boldsymbol{B}_{T,1},$$

where $\mathbf{T}_n$ represents the knot matrix for the n^{th} order spline (i.e. B-spline breaks) [16]. For a symmetric kernel (r even), the general form of the knot matrix is

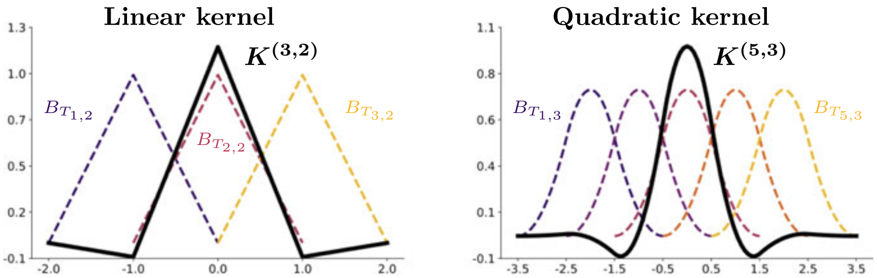


Fig. 1 Example B-spline kernels (solid lines) consisting of $2p + 1$ B-Splines (dashed lines) of order $p + 1$ for $p = 1$ (left) and $p = 2$ (right)

$$\mathbf{T} = \begin{pmatrix} -\frac{n+r}{2} & \frac{-(n+r)+2}{2} & \dots & \frac{n-r}{2} \\ -\frac{n+r-2}{2} & \frac{-(n+r)+4}{2} & \dots & \frac{n+2-r}{2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{r-n}{2} & \frac{r-n+2}{2} & \dots & \frac{n+r}{2} \end{pmatrix}, \tag{5}$$

where each row gives the B-Spline breaks of the γ^{th} B-spline [16] ($\gamma = 1, \dots, r + 1$). Examples of the usual kernels used for $p = 1,\ 2$ are given in Fig. 1.

As discussed by Bramble and Schatz [3] and Thomeé [24], choosing the kernel basis function to be B-Splines has an added advantage when it comes to extracting information as derivatives can be written as divided differences of lower order splines,

$$\frac{\partial^\alpha}{\partial x^\alpha} \boldsymbol{B}_{\mathbf{T}_n}(x) = \partial_H^\alpha \boldsymbol{B}_{\mathbf{T}_n - \alpha}(x),$$

where ∂_H^α represents the α^{th} divided difference. This is an essential property in order to maximize the accuracy extracting capability for piecewise polynomial data. The general B-spline SIAC kernel is then

$$K_H^{(r+1,n)}(x) = \sum_{\gamma=1}^{r+1} c_\gamma B_{\mathbf{T}_\gamma,n}(x). \tag{6}$$

Here, $B_{\mathrm{T},n}$ represents n^{th}-order central B-spline with knot sequence, $\mathbf{T}$ and smoothness $n - 2$. The scaling, H, is generally tied to the spacing of the data.

For the symmetric kernel with equally spaced knots, the Fourier transform of the SIAC kernel illustrates the different parameters that make for a flexible filter. The Fourier transform is given by

$$\mathcal{F}(K) = \widehat{K}(\xi) = \underbrace{\text{sinc}\left(\frac{\xi}{2}\right)^n}_{\text{controls dissipation}} \underbrace{\left(c_{\frac{r+2}{2}} + 2\sum_{\gamma=1}^{\lceil \frac{r}{2} \rceil} c_\gamma \cos\left(\left(\gamma - \frac{r+2}{2}\right)\xi\right)\right)}_{\text{moment conditions}} \tag{7}$$

Thus there are three parameters that affect the performance of the filter: the number of kernel basis functions which is tied to the number of moments, the smoothness that is tied to the amount of dissipation, and the scaling that takes advantage of the patterns of information contained in the data. Note that this satisfies the classical definition of a spectral filter [25].

Why SIAC Works. The ability of SIAC to extract hidden data is not magic and more complete theoretical discussions are given in [9, 14, 16, 19–22]. A review article summarizing the theory and previous applications can be found in [8]. The theoretical framework essentially relies on the ability to take advantage of the patterns of noise, which is hidden in the Fourier information. This pattern of the error will be found in any data that is represented by piecewise polynomials and results in obtaining different convergence orders when analyzing the errors in the solution versus the dispersion and dissipation errors. The pattern of the error for discontinuous Galerkin methods was studied by Adjerid et al. in [1], where the authors examine the leading term of the error. It is also discussed in [19, 20]. This results in a discrepancy in the error analysis when investigating through Taylor series versus Fourier space. For example, for data from discontinuous Galerkin solutions, the typical convergence rate is $\mathcal{O}(h^{p+1/2})$ and the dispersion and dissipation errors are of $\mathcal{O}(h^{2p+1,2p+2})$ [2, 11, 23, 27]. In [11], it was shown that the non-physical oscillations are damped exponentially fast in time, allowing the physical eigenvalue to dominate the wave propagation. It is the information from these eigenvalues that the SIAC filter extracts. Because it utilizes convolution, it can translate the higher accuracy in Fourier space to physical space. This ensures that the filter works for data generated from both linear and nonlinear equations. Note that for data generated from Galerkin-type methods, the translation of the information from Fourier space to physical space is done by encoding the information of the dual equation by utilizing the information in the "noise" that is measured by the negative-order norm,

$$\|f - f_h\|_{H^{-\ell}(\Omega)} = \sup_{\Phi \in \mathcal{C}_0^\infty(\Omega)} \frac{(f - f_h, \Phi)}{\|\Phi\|_{H^\ell}}, \tag{8}$$

where Φ is the solution to the dual equation at the final time. As the dual may not be unique, such as for non-linear equations, we take the supremum over all solutions that are continuous with compact support. Assuming that the solution has enough smoothness to allow for recovering sufficient accuracy (i.e. $\mathcal{C}^s$, where s is the convergence rate), the optimal orders of accuracy are given in Table 1.

Table 1 Methods and highest order of convergence on locally translation invariant mesh as given in [3, 5, 12, 14, 24, 28]

Method	B-splines		Norm	Possible convergence order
	Number	Order		
Elliptic & Parabolic				
Ritz-Galerkin [3, 10, 24]	$2p-1$	p-1	$L^2,\ L^\infty$	$2p-2,\ p \geq 3$
Spectral Element [15, 28]			$L^2,\ L^\infty$	$p+2,\ (2p+1)(2p+3) > 2M^2$
Linear hyperbolic				
Standard Galerkin [5]	$2p+1$	p+1	L^2	$2p,\ p \geq 1$
Discontinuous Galerkin [5, 14]	$2p+1$	p+1	$L^2,\ L^\infty$	$2p+1,\ p \geq 1$
Active Flux [12]	3	1	L^∞	4

Remark 1 It is possible to extend this discussion to extracting accuracy for derivative information. This can be seen by replacing u with $v = \frac{\partial u}{\partial \mathbf{x}}$, where $= (\alpha_1,\ \alpha_2,\ \ldots,\ \alpha_d)$. To achieve the same superconvergence order as the approximation itself, the B-splines are of order $n + |\alpha|$.

Remark 2 The second estimate in Eq. (2) is affected by the underlying equation, numerical scheme, and smoothness, $n-2$.

The traditional extension to multi-dimensions is via a tensor product:

$$K_{\mathbf{H}}(\mathbf{x}) = K_{\Delta x_1}(x_1) K_{\Delta x_2}(x_2) \cdots K_{\Delta x_d}(x_d).$$

However, a more computationally efficient multi-dimensional kernel is the Line SIAC kernel [9], which is a rotated one-dimensional kernel,

$$K_H() = K_\Gamma() \Rightarrow f_h^\star(\overline{x}, \overline{y}) = \int_\Gamma K_\Gamma\left(\frac{\Gamma(0) - \Gamma(t)}{h_t}\right) f(\Gamma(t)) dt \qquad (9)$$

with the filtering performed along a line. In 2D, the line is defined by $\Gamma(t) = (\overline{x}, \overline{y}) + \lambda(\cos(\theta),\ \sin(\theta))$, with an angle of rotation $\theta = \tan\left(\frac{\Delta y}{\Delta x}\right)$. An illustration of the difference in the support size for the two-dimensional tensor product filter versus the line filter is given in Fig. 2. For the kernel consisting of 5 B-splines of order 3 over a structured mesh, the tensor product filter requires 196 $2D$ integrals while the line filter requires 21 $1D$ integrals. For both implementations, the integrals are computed using exact quadrature. This necessitates the integration to respect both the B-spline breaks as well as element interfaces. For an approximation of polynomial degree p using a B-spline kernel of order n in each smooth region, the quadrature is computed

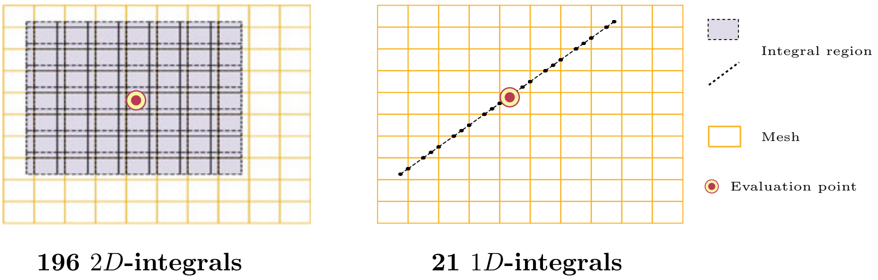


Fig. 2 Footprints for the tensor product (left) and line (right) SIAC kernels on a uniform mesh. The kernel given is $K_H^{(5,3)}$ () (a 5 B-spline kernel of order 3)

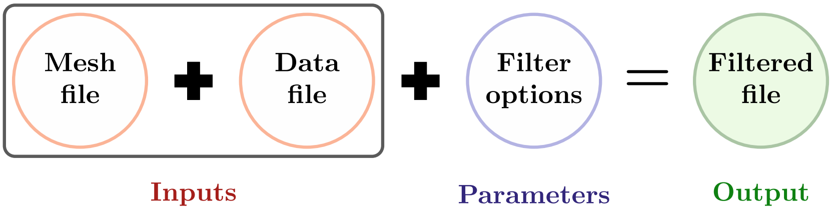
using $\lceil \frac{p+n}{2} \rceil$ quadrature points per region. It is possible to utilize inexact quadrature, but this needs many more points and will affect the performance of the filter.

2 The SIAC Magic Toolbox

In this section we outline the necessary components for using the SIAC Magic Toolbox as well as the computational challenges. We also present a comparison of the computational performance in terms of CPU time and the ability to run on many cores.

2.1 How It Works

The SIAC magic toolbox [7] is a standalone tool written in Julia [4]. It takes in two files: one containing information on the data and one containing the corresponding mesh information. The output is then the filtered values. Currently, it is supporting conforming meshes constructed of simplices. Support for non-conforming meshes and higher-order elements will be added in the future.



All the necessary tools for filtering are contained in the MSIAC package. Inside of Julia, it only requires activating the package. A `test.jl` file is also included to check that the package was installed correctly. The main function that reads in the data and filters the solution is given in Toolbox Code 1. The user can specify the parameters pertaining to dissipation, moments, and scaling, or choose to use default values provided. Additionally, it's possibly for the user to load their own mesh and data files from gmesh or vtk (c.f. Toolbox 1). The toolbox utilizes Julia's multi-threading and distributed memory capability.

```
function filter_data (mesh, data, parameters)
    modes  = l2_projection(data)
    kernel = set_kernel(parameters)
    for point in data
        map = find_kernel_breaks(mesh, point, kernel)   # Footprint
        point* = sum( gauss(map, kernel, modes))        # Convolution
    end
end
```

Toolbox Code 1: *Main SIAC function. The modal information is first constructed from the given function values at the quadrature points. The kernel is then constructed using either default or user-specified values. Within the for loop, the kernel footprint is first defined and then the convolution is performed using quadrature*

In the Toolbox 1 example, the data is assumed to be sampled at quadrature points (Legendre, Lobatto, or Radau) and is converted to modal information as shown in Toolbox Code 1. The mesh file contains information on the mesh map. That is, element indices, connectivity, and boundaries. In Toolbox Code 3, the assumed mesh data structure is shown together with an illustration for one element, 5, that has neighboring elements 2, 6, 8, and 4 and is defined by nodes 6, 7, 11, and 10. The node structure contains information on the coordinates and the surrounding nodes as well as a type designation.

2.2 Computational Performance

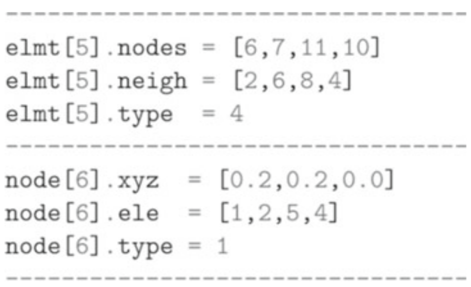
Collecting the spline breaks and element interfaces presents the most challenging aspect of implementing the filters. This is because it requires determining the intersection of the mesh on which the data is given as well as the reduced points of continuity of the kernel (i.e. kernel breaks). This determines the filter footprint and hence requires evaluating more integrals than the number of elements in the support of the kernel. This is the reason that most of the CPU time is spent computing the filter footprint, making the tensor product filter intractable in multi-dimensions without using HPC. The Line filter aids in alleviating this difficulty and remains a one-dimensional filter in multiple dimensions.

To determine the intersections of the kernel breaks and element interfaces, the algorithm avoids querying the point location and instead utilizes sorted knot matrices

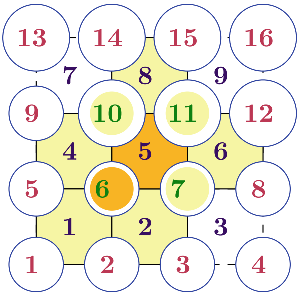
```
data = load_data(meshFile,fieldFile);
#--- For more Options:
#        degree {select degree}
#        modalExp {default expansions: "legendre" for quads and
   "hierarchy" (modified legendre) for tris}
#        structured {defaults as false}
#        reverse {defaults as false}
data = load_data(meshFile,fieldFile, modalExp="Pk",
   structured=false, reverse=false);
data = load_data(meshFile,fieldFile, degree=p);
```

Toolbox Code 2: *The user can load pre-existing data from .vtu files and prescribe the necessary parameters or utilize default values*

```
struct element
        nodes :: List of ordered vertex indexes forming element
        neigh :: List element neighbors
        type  :: 4 (quad) 3 (triangle)
end
struct node
        xyz  :: vertex coordinates
        ele  :: List of elements surrounding vertex
        type :: 1 (interior) 0 (boundary edge) -1 (boundary corner)
end
```



```
--------------------------------
elmt[5].nodes = [6,7,11,10]
elmt[5].neigh = [2,6,8,4]
elmt[5].type  = 4
--------------------------------
node[6].xyz  = [0.2,0.2,0.0]
node[6].ele  = [1,2,5,4]
node[6].type = 1
--------------------------------
```



Toolbox Code 3: *The assumed mesh data structure (top) and an illustration (bottom) for one element (=5) and one node (6)*

to finding the direction along a line where the next mesh or kernel break will be located. This is illustrated in Fig. 3 and the main function is shown in Toolbox Code 4. For example, if there is one B-spline break per element and the breaks are translates of $\bar{\mathbf{x}}$, with $\bar{\mathbf{x}}$ being along the mesh diagonal in element e, the line segments in e are $(e_{-,-}, \bar{\mathbf{x}})$ and $(\bar{\mathbf{x}}, e_{+,+})$, with $e_{-,-}$ representing the lower left corner and $e_{+,+}$ representing the upper right corner. As an example, consider a kernel consisting of three uniform B-splines of order two, i.e. $r = 2,\ n = 2$. The knot matrix is given by

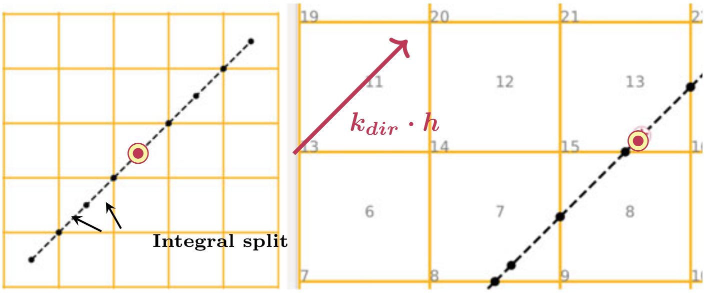


Fig. 3 Illustration of the line kernel footprint, integral breaks, and search direction, k_{dir}

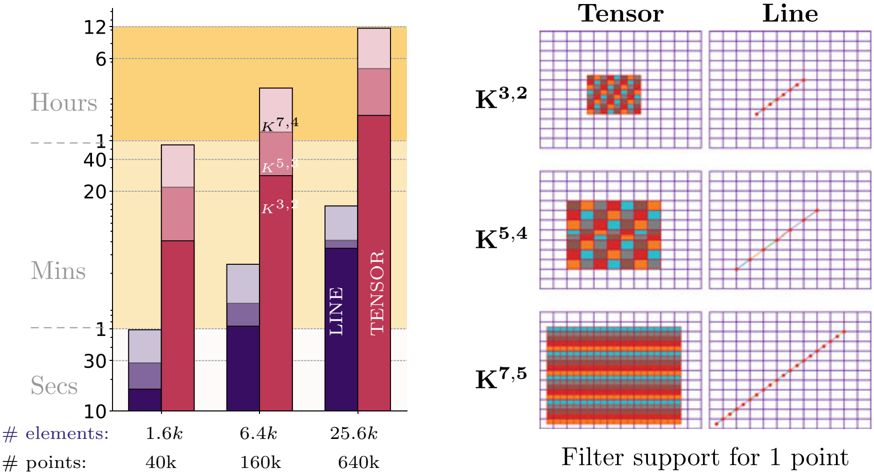
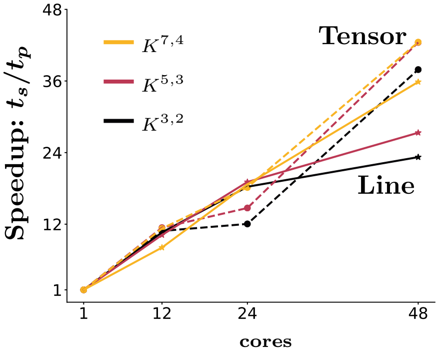


Fig. 4 Performance of code for both the tensor product and Line filter using three different kernels traditionally used to extract information for discontinuous Galerkin approximations. The shades of red represent the performance of the tensor product filter for the $K^{3,2}$ $K^{5,3}$ and $K^{7,4}$ kernels (Darkest → lightest), while the corresponding shades of purple show the performance of the line filter

$$T = \begin{pmatrix} -2 & -1 & 0 \\ -1 & 0 & 1 \\ 0 & 1 & 2 \end{pmatrix}. \tag{10}$$

The sorted knot matrices, $T_{-} = (-2, -1, 0)$ and $T_{+} = (0, 1, 2)$, give the direction along the line to search for the next kernel/data mesh intersection, denoted k_{dir}.

In Fig. 4, CPU times using a Dell XPS with four cores are given for both filters when post-processing a solution sampled at 25 points per element. The shades of red represent the performance of the tensor product filter for the $K^{3,2}$ $K^{5,3}$ and $K^{7,4}$ kernels, while the shades of purple show the performance of the line filter. While the times indicate that the line filter is faster, the tensor product filter benefits more from



25.6k elements, 640k points

	Line (mins)		Tensor (hrs)	
CPUs/ Order	4	48	4	48
2	5.8	0.5	1.7	0.1
3	6.9	0.7	4.7	0.3
4	14.5	1.1	11.4	0.8

Fig. 5 Parallelization performance of the tensor product SIAC filter versus the Line SIAC filter on MareNostrum4, with a maximum of 48 cores (1 node). Data originated from a discontinuous Galerkin approximation to Burgers equation

parallelization. This can be observed in Fig. 5 (left) where the speedup factors for each filter are plotted using up to 48 cores from the MareNostrum4 supercomputer: the speedup for the tensor product filter scales almost linearly. For the line filter, the calculation times per point are low. Note that the overall CPU times (right) for the line filter are measured in minutes while the times for the tensor product filter are measured in hours. The results indicate that both filters benefit from parallelization, and that the tensor product filter should be used in an HPC-environment.

```
for t in (T-,T+)
    xp = eval point, e=13
    for i in t
        xn = kdir * h * i
        Spn = segment(xp,xn)
        if xn in elmt[e]
            store(xn,elmt[e])
            xp = xn
        else
            (j,xp) = intersect(Spn,elmt [13])
            store(xp,e,elmt[e].neigh[j])
            e = elmt[e].neigh[j]     #8
        end
    end
end
```

Toolbox Code 4: *mesh intersections. The algorithm takes in an evaluation point and determines the line segments that are in the kernel support along the direction* k_{dir}

Table 2 One- and two-dimensional capabilities of the SIAC Magic Toolbox

			Mesh			
	Kernel		Uniform		Non-uniform	
Filter	Symmetric	Shifted	Quads	Triangles	Quads	Triangles
Tensor	🟩	🟩	🟩	🟩	🟧	🟧
Line	🟩	🟧	🟩	🟩	🟧	🟧

🟩 Theory & implemented 🟧 Implemented, no theory

2.3 Current Capabilities

The current capabilities of the SIAC Magic Toolbox are given in Table 2, which also connects the code corresponding to published (and unpublished) theory. Additionally, the 3D Line filter is available for structured and shape regular hexahedral and tetrahedral elements. For the data input file, nodal or modal information can be specified. While the toolbox has not been tested for non-conforming meshes, in principle, it should be able to handle such mesh types since the kernel footprint algorithm loops around all the neighbours at every element interface. Curved meshes are not supported at the moment but will be in the future.

3 Numerical Examples

In the following section, we present results demonstrating the ability of the SIAC Magic Toolbox to effectively raise the order of convergence of data arising from simulations. In each example, data was produced from a discontinuous Galerkin approximation. Error contours for both the tensor-product SIAC filter and Line SIAC filter are shown.

Table 3 Errors for the $2D$ Burgers equation with source term for the provided data as well as the Line SIAC and tensor-product SIAC filters

	N	$\|\|e\|\|_{L^2}$	Rate	$\|\|e\|\|_{L^2}$	Rate	$\|\|e\|\|_{L^2}$	Rate
$p = 1$	40^2	3.02e-03	2.01	3.34e-04	3.32	1.37e-04	3.05
	80^2	7.52e-04	2.00	4.74e-05	2.82	1.63e-05	3.07
$p = 2$	40^2	4.01e-05	3.00	9.37e-06	6.04	4.66e-07	4.65
	80^2	4.93e-06	3.02	2.95e-07	4.99	1.52e-08	4.93
DG			**Line**		**Tensor Product**		

Burgers Equation with Source Term In the first example, we present the results of post-processing data that is taken from a discontinuous Galerkin approximation to Burgers equation with a source term,

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x}\left(\frac{u^2}{2}\right) + \frac{\partial}{\partial y}\left(\frac{u^2}{2}\right) = f(x, y).$$

The approximation is taken over a uniform quadrilateral mesh and the source term is taken such that the exact solution is $u(x, y, t) = \sin(\pi(x + y))$, where $(x, y) \in [-1, 1]^2$. At the final time, the data is passed into the SIAC Magic toolbox and is filtered using both the Line SIAC filter and tensor-product filter. Both filters are taken such that they satisfy $2p$ moment conditions and have smoothness of order $p - 1$, and hence composes a filter utilizing $2p + 1$ B-splines of order $p + 1$. In Table 3 the errors are presented for piecewise linear and piecewise quadratic approximations and the corresponding error contours are given in Fig. 6. For this nonlinear equation, both filters are able to raise the convergence rate from $p + 1$ to $2p + 1$, with the errors for tensor-product SIAC filtered approximation being lower than those for LSIAC.

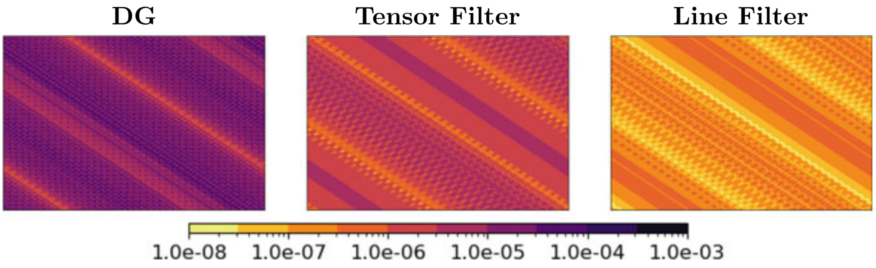


Fig. 6 Error contours for the $2D$ Burgers with source term for the provided $p = 2$ data on a 40×40 mesh. The data is read into the toolbox and the error contours for the tensor-product SIAC and Line SIAC filters are shown

3.1 Perturbation of Quadrilateral Elements

We next explore the ability of the Line SIAC and tensor product SIAC filters to perform on L^2-projected functions defined on random perturbations of a uniform quadrilateral grids for $p = 1$ and $p = 2$ on a 40×40 mesh. We consider the functions $u(x, y) = e^{-(x^2+y^2)}$ as well as $u(x, y) = \sin(x)\cos(y)$. The mesh and corresponding function and error contours are presented in Fig. 7. Both the tensor product and line filters reduce the errors from the initial data.

3.2 LSIAC Filtering: Triangular Meshes and Paraview

In this example we utilize data in a Paraview format on structured and unstructured triangular meshes. We present the error contours for piecewise linear data ($p = 1, n = 20^2$) for an L^2-projected sine wave in Fig. 8.

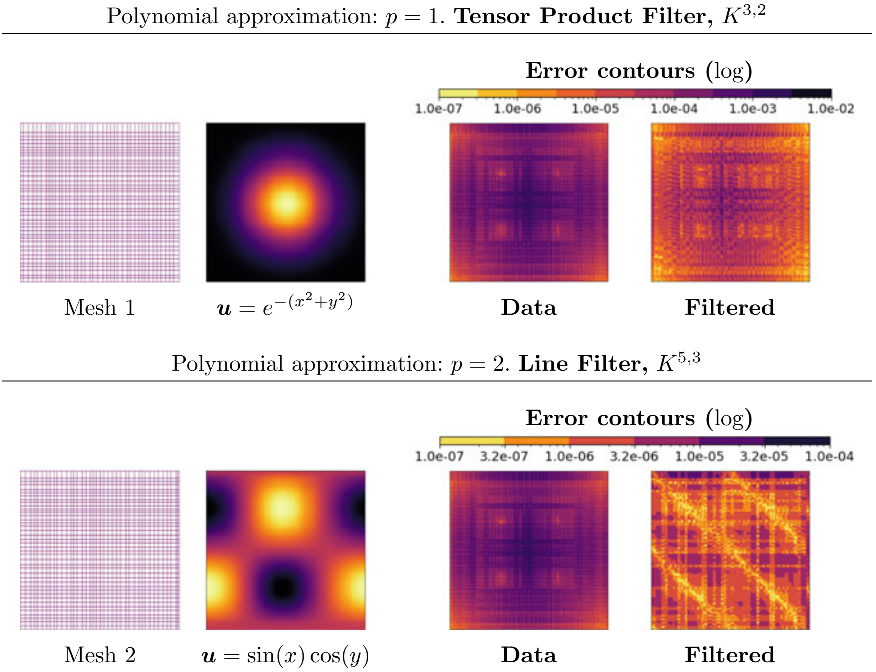


Fig. 7 *Performance of the tensor product and line filters on two different* meshes consisting of 40×40 perturbed quadrilateral elements when post-processing L^2-projected analytic functions corresponding to $u(x, y) = e^{-(x^2+y^2)}$ (top) and $u(x, y) = \sin(x)\cos(y)$ (bottom). The filters use three B-splines of order two

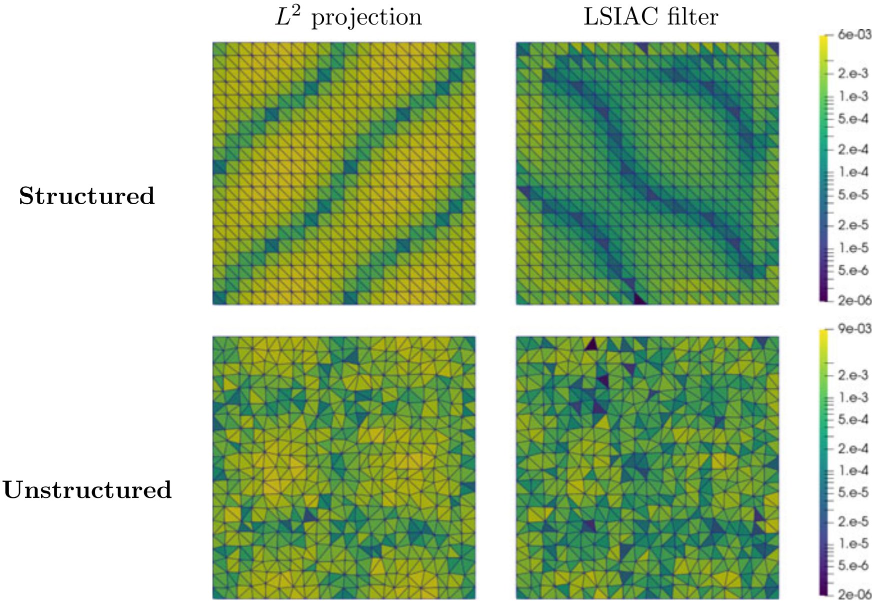


Fig. 8 Error contours for the LSIAC (right) filtered solutions on a structured and unstructured triangular mesh. The given data is an L^2-projection of a sine function and is imported from Paraview

4 Summary

The SIAC Magic Toolbox has been created to enable easy access to flexible filtering tools and increased use in practical applications. The toolbox only requires the data values and mesh information. The user is able to test different parameters related to the dissipation, accuracy, and scaling. This allows the parameters to be tuned to provide optimal effectiveness for a given application. Alternatively, the user can use the default values defined by the toolbox. The toolbox will return the filtered values.

Use of the toolbox automatically ensures consistency, r moments, and $n-1$ smoothness. Enforcement of r moments allows for convergence up to order $r+1$ (and $r+2$ if the filter is symmetric) while providing the capability to capture features of the data. Essentially, SIAC filters work by reducing noise in the data by capturing appropriate patterns of information that correspond to the physically relevant eigenvalues. It is applicable for periodic and non-periodic boundary conditions as well as structured and unstructured meshes. Because it is based on convolution, it is effective for both linear and nonlinear equations as well as different data types such as finite element methods, including discontinuous Galerkin and Spectral Element methods, as well as Active Flux methods. It can easily incorporate data from experiments through quasi-interpolation [17].

Acknowledgements The authors would like to thank Matt Picklo and Soraya Terrab for help in testing the code. The work of the first author was supported by European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 893378. The HPC experiments were run on the BSC MareNostrum 4 supercomputer. The work of the second author was supported by the US Air Force Office of Scientific Research Computational mathematics under grant number FA9550-20-1-0166 and US National Science Foundation Computational mathematics under grant number DMS-2110745.

References

1. Adjerid, S, Devine, K.D. Flaherty, J.E., Krivodonova, L.: A Posteriori Error Estimation for Discontinuous Galerkin Solutions of Hyperbolic Problems. Comp. Meth. Appl. Mech. Eng. **191**, 1097–1112 (2002)
2. Ainsworth, M.: Dispersive and dissipative behaviour of high order discontinuous Galerkin finite element methods. J. Comp. Phys. **198**, 106–130 (2004)
3. Bramble, J.H., Schatz, A.H.: Higher order local accuracy by averaging in the finite element method. Math. Comp. **31**, 94–111 (1977)
4. Bezanson, J., Edelman, A., Karpinski, S., Shah, V.B.: Julia: a fresh approach to numerical computing. SIAM Rev. **59**, 65–98 (2017). https://julialang.org
5. Cockburn, C., Luskin, M., Shu, C.-W., Suli, E.: Enhanced accuracy by post-processing for finite element methods for hyperbolic equations. Math. Comp. **72**, 577–606 (2003)
6. Dedner, A., Giesselmann, J. Pryer, T., Ryan, J.K.: Residual estimates for post-processors in elliptic problems. J. Sci. Comp. **88** (2021)
7. Docampo Sánchez, J., Ryan, J.K.: SIAC Magic Toolbox. https://siac-magic.gitlab.io/web/
8. Docampo Sánchez, J., Jacobs, G. , Li, X., Ryan, J.K.: Enhancing accuracy with a convolution filter: what works and why! Comp. Fluids **213** (2020)
9. J. Docampo Sánchez, Ryan, J.K., Mirzargar, M., Kirby, R.M.: Multi-dimensional filtering: reducing the dimension through rotation. SIAM J. Sci. Comp. **39**, A2179–A2200 (2017)
10. Douglas, J., Dupont, T., Wheeler, M.F.: A quasi-projection analysis of Galerkin methods for parabolic and hyperbolic equations. Math. Comp. **32**, 345–362 (1978)
11. Guo, W., Zhong, X., Qiu, J.-M.: Superconvergence of discontinuous Galerkin and local discontinuous Galerkin methods: eigen-structure analysis based on Fourier approach. J. Comp. Phys. **235**, 458–485 (2013)
12. Helzel, C., Kerkmann, D., Ryan, J.: An active flux cut cell method with SIAC filter (2023)
13. Jallepalli, A., Docampo Sánchez, J., Ryan, J.K., Haimes, R. Kirby, R.M.: On the treatment of field quantities and elemental continuity in FEM solutions. IEEE TVCG **24**, 903–912 (2018)
14. Ji, L., van Slingerland, P., Ryan, J.K., Vuik, C.: Superconvergent error estimates for a position-dependent smoothness-increasing accuracy-conserving filter for DG solutions. Math. Comp. **83**, 2239–2262 (2014)
15. Li, H., Appelö, D., Zhang, X.: Accuracy of spectral element method for wave, parabolic and Schrödinger equations. SINUM **60**, 339–363 (2022)
16. Li, X., Ryan, J.K., Kirby, R.M., Vuik, C.: Smoothness-Increasing Accuracy-Conserving (SIAC) filters for derivative approximations of discontinuous Galerkin (DG) solutions over nonuniform meshes and near boundaries. J. Comput. Appl. Math **294**, 275–296 (2016)
17. Mirzargar, M., Ryan, J.K., Kirby, R.M.: Smoothness-Increasing Accuracy-Conserving (SIAC) filtering and quasi-interpolation: a unified view. J. Sci. Comp. **67**, 237–261 (2016)
18. Mock, M.S., Lax, P.D.: The computation of discontinuous solutions of linear hyperbolic equations. Comm. Pure. Appl. Math. **31**, 423–430 (1978)
19. Picklo, M., Ryan, J.K.: Enhanced multi-resolution analysis for multi-dimensional data utilizing line filtering techniques. SIAM J. Sci. Comp. **44**, A2628–A2650 (2022)

20. Ryan, J.K.: Capitalizing on Superconvergence for more accurate multi-resolution discontinuous Galerkin methods. Comm. Appl. Math. Comp. **4** (2022)
21. Ryan, J.K., Li, X., Kirby, R.M., Vuik, C.: One-Sided Position-Dependent Smoothness-Increasing Accuracy-Conserving (SIAC) filtering over uniform and non-uniform meshes. J. Sci. Comp. **64**, 773–817 (2015)
22. Ryan, J.K., Shu, C.-W., Atkins, H.L.: Extension of a post-processing technique for discontinuous Galerkin Methods for hyperbolic equations with application to an Aeroacoustic problem. SIAM J. Sci. Comp. **26**, 821–843 (2004)
23. Sherwin, S.: Dispersion analysis of the continuous and discontinuous Galerkin formulations. Discontinuous Galerkin Methods, pp. 425–431. Springer, Berlin (2000)
24. Thomée, V.: High order local approximations to derivatives in the finite element method. Math. Comp. **31**, 652–660 (1977)
25. Vandeven, H.: Family of spectral filters for discontinuous problems. J. Sci. Comp. **6**, 159–192 (1991)
26. Wissink, B.W., Jacobs, G.B., Ryan, J.K., Don, W.S., van der Weide, E.T.A.: Shock regularization with smoothness-increasing accuracy-conserving Dirac-delta polynomial kernels. J. Sci. Comp. **77**, 579–596 (2018)
27. Zhang, M., Shu, C.-W.: Fourier analysis for discontinuous Galerkin and related methods. Chin. Sci. Bull. **54**, 1809–1816 (2009)
28. Zhang, Z.: Superconvergence of spectral collocation and p-version methods in one dimensional problems. Math. Comp. **74**, 1621–1636 (2005)

A Review of Cartesian Grid Active Flux Methods for Hyperbolic Conservation Laws

Erik Chudzik and Christiane Helzel

Abstract In 2011, Eymann and Roe introduced a new class of truly multidimensional finite volume methods. These so-called Active Flux methods use concepts, which are quite different from concepts that are typically used in numerical schemes for hyperbolic conservation laws. In particular the method is based on the use of point values as well as cell average values, it uses a continuous reconstruction and does not rely on the use of Riemann solvers. We will review the current state of the art of Cartesian grid Active Flux methods and present recent results of our group. These include a discussion of different evolution formulas for the update of the point values, results on the linear stability of the resulting methods as well as a discussion of limiters. Furthermore, we explore the use of Active Flux methods on Cartesian grids with adaptive mesh refinement.

Keywords Active flux methods · Hyperbolic conservation laws

1 Introduction

The Active Flux method, as introduced by Eymann and Roe [9, 10], is a fully discrete, truly multi-dimensional, third order accurate finite volume method with compact stencil in space and time. The method uses point values as well as cell average values of the conserved quantities as degrees of freedom. The point values, which are located along the grid cell boundary, and the cell average values of the previous time level provide a globally continuous, piecewise quadratic reconstruction, that preserves the cell averages. Point values are evolved in time using truly multidimensional evolution operators which may be derived from the partial differential equation.

The evolution of the cell average values is computed using a finite volume approach. A quadrature rule, typically Simpson's rule, is used to compute the numer-

E. Chudzik · C. Helzel (✉)
Institute of Mathematics, Heinrich-Heine University Düsseldorf, Düsseldorf, Germany
e-mail: christiane.helzel@hhu.de

E. Chudzik
e-mail: erik.chudtiu@hhu.de

E. Franck et al. (eds.), *Finite Volumes for Complex Applications X—Volume 1, Elliptic and Parabolic Problems*, Springer Proceedings in Mathematics & Statistics 432,
https://doi.org/10.1007/978-3-031-40864-9_6

ical fluxes. The point values at the previous, an intermediate and the new time level are used as nodes in the quadrature formula.

The fully discrete form and the local stencil in space and time lead to a method which is accurate even on coarse grids [19–22]. Another distinctive property of Active Flux methods is the use of a globally continuous reconstruction. As pointed out by Roe [18], methods based on piecewise continuous reconstructions typically introduce "one-dimensional physics" due to the use of Riemann solvers. Due to their high accuracy even on coarse grids and their ability to model the multi-dimensionality of the mathematical problem, Active Flux methods are interesting candidates for the computation of complex fluid mechanical processes. This motivates our current work on extending these methods for the approximation of multi-dimensional nonlinear systems.

Abgrall [1] pointed out that the Active Flux method allows to combine different writings of hyperbolic problems in a single numerical method. While the conservative form is used for the update of the cell average values, the primitive form or an entropy formulation might be used for the update of the point values.

Extensions of the Active Flux method to higher than third order have recently been explored for advection [19], acoustics [22] and general one-dimensional hyperbolic problems [2].

2 The Cartesian Grid Active Flux Method

In this section we present the main components of the Active Flux method for one- and two-dimensional hyperbolic problems. While the evolution of the cell average values is presented in a form valid for general hyperbolic problems in divergence form, we restrict our description of the evolution of the point values to the simplest case of linear advection. In Sect. 3 the evolution of the point values for further hyperbolic problems is discussed.

The Active Flux method is a truly multi-dimensional scheme, however the outline of the method can most easily be explained for one-dimensional conservation laws

$$\partial_t q(x,t) + \partial_x f(q(x,t)) = 0, \tag{1}$$

where $q : \mathbb{R} \times \mathbb{R}^+ \to \mathbb{R}^s$ is a vector of conserved quantities and $f : \mathbb{R}^s \to \mathbb{R}^s$ is a flux function.

The Active Flux method uses cell average values and point values of the conserved quantities, which are denoted by

$$Q_i^n \approx \frac{1}{\Delta x}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} q(x,t_n)dx, \quad Q_{i\pm\frac{1}{2}}^n \approx q(x_{i\pm\frac{1}{2}},t_n).$$

In order to describe a time step from t_n to t_{n+1}, we assume that for all i the quantities Q_i^n and $Q_{i\pm\frac{1}{2}}^n$ are at least third order accurate in space and time. The update of the cell average values is computed using a finite volume method

$$Q_i^{n+1} = Q_i^n - \frac{\Delta t}{\Delta x}\left(F_{i+\frac{1}{2}} - F_{i-\frac{1}{2}}\right),$$

where the numerical fluxes are computed using Simpson's rule, i.e.

$$F_{i\pm\frac{1}{2}} = \frac{1}{6}\left(f(Q^n_{i\pm\frac{1}{2}}) + 4f(Q^{n+\frac{1}{2}}_{i\pm\frac{1}{2}}) + f(Q^{n+1}_{i\pm\frac{1}{2}})\right) \approx \frac{1}{\Delta t}\int_{t_n}^{t_{n+1}} f(q(x_{i\pm\frac{1}{2}}, t))\, dt.$$

The crucial step of the Active Flux method is the computation of the point values $Q^{n+\frac{1}{2}}_{i\pm\frac{1}{2}}$ and $Q^{n+1}_{i\pm\frac{1}{2}}$ at the intermediate and new time level. In this section we describe the update of the point values for the simplest case of linear advection with advection speed $a \in \mathbb{R}$, i.e.

$$\partial_t q + a\partial_x q = 0.$$

The point values can be approximated using the well known exact evolution formula $q(x_{i\pm\frac{1}{2}}, t_n + \tau) = q(x_{i\pm\frac{1}{2}} - a\cdot\tau, t_n)$, $\tau \in \{\frac{\Delta t}{2}, \Delta t\}$. Using the point values and the cell average values, a continuous, piecewise quadratic reconstruction q^{rec} can be computed, which preserves the cell averages and agrees with the point values at the grid cell interfaces. In grid cell i the reconstruction can be expressed in the form

$$q_i^{rec}(\xi) = Q^n_{i-\frac{1}{2}}(3\xi^2 - 4\xi + 1) + Q_i^n(6\xi - 6\xi^2) + Q^n_{i+\frac{1}{2}}(3\xi^2 - 2\xi), \tag{2}$$

with $0 \le \xi \le 1$. Thus for advection the approximation of the point values has for $k \in \{\frac{1}{2}, 1\}$, $\tau = k\Delta t$ the form

$$Q^{n+k}_{i+\frac{1}{2}} = \begin{cases} q_i^{rec}(1 - a\tau/\Delta x) : a > 0 \\ q_{i+1}^{rec}(-a\tau/\Delta x) \ \ : a < 0. \end{cases}$$

For one-dimensional advection with advection speed $a > 0$ we observe

$$\begin{aligned}
\frac{a}{\Delta t}\int_{t_n}^{t_{n+1}} q(x_{i+\frac{1}{2}}, t)\, dt &= \frac{a}{\Delta t}\int_{t_n}^{t_{n+1}} q(x_{i+\frac{1}{2}} - a(t - t_n), t_n)\, dt \\
&= \frac{1}{\Delta t}\int_{x_{i+\frac{1}{2}} - a\Delta t}^{x_{i+\frac{1}{2}}} q(x, t_n)\, dx \\
&\approx \frac{\Delta x}{\Delta t}\int_{1 - a\frac{\Delta t}{\Delta x}}^{1} q_i^{rec}(\xi)\, d\xi \\
&= \frac{a}{6}\left(q_i^{rec}\left(1 - \frac{a\Delta t}{\Delta x}\right) + 4q_i^{rec}\left(1 - \frac{a\Delta t}{2\Delta x}\right) + q_i^{rec}(1)\right) \\
&=: F_{i+\frac{1}{2}}.
\end{aligned}$$

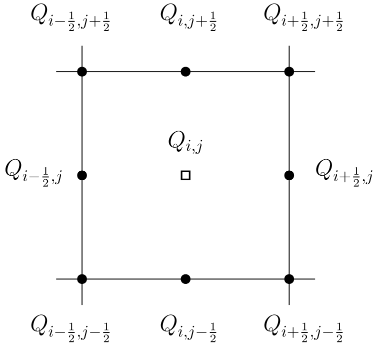

$Q_{i-\frac{1}{2},j+\frac{1}{2}}$
$Q_{i,j+\frac{1}{2}}$
$Q_{i+\frac{1}{2},j+\frac{1}{2}}$
$Q_{i,j}$
$Q_{i-\frac{1}{2},j}$
$Q_{i+\frac{1}{2},j}$
$Q_{i-\frac{1}{2},j-\frac{1}{2}}$
$Q_{i,j-\frac{1}{2}}$
$Q_{i+\frac{1}{2},j-\frac{1}{2}}$

Fig. 1 Degrees of freedom used for two-dimensional Cartesian grid Active Flux method

An approximation is only introduced by replacing the exact solution $q(x, t_n)$ by the third order accurate piecewise quadratic function q^{rec}. The application of Simpson's rule provides the exact integral. We will see below that the two-dimensional case is slightly different and that this difference leads to both a reduced stability and additional difficulties of the limiting process.

In the two-dimensional case, i.e. for the approximation of

$$\partial_t q + \partial_x f(q) + \partial_y g(q) = 0,$$

with $q : \mathbb{R} \times \mathbb{R} \times \mathbb{R}^+ \to \mathbb{R}^s$, $f, g : \mathbb{R}^s \to \mathbb{R}^s$, we use a Cartesian grid. The degrees of freedom of the method are again cell average values and point values of the conserved quantities. The point values are located along the grid cell boundary as illustrated in Fig. 1.

A piecewise quadratic and globally continuous reconstruction, which preserves the cell average values, can be described using appropriate basis functions. See [4, 13] for details. The update of the cell average values is performed by a finite volume update of the form

$$Q_{ij}^{n+1} = Q_{ij}^n - \frac{\Delta t}{\Delta x}\left(F_{i+\frac{1}{2},j} - F_{i-\frac{1}{2},j}\right) - \frac{\Delta t}{\Delta y}\left(G_{i,j+\frac{1}{2}} - G_{i,j-\frac{1}{2}}\right), \tag{3}$$

where the numerical flux is computed using the two-dimensional Simpson's rule. The flux across a vertical grid cell interface has the form

$$\begin{aligned} F_{i+\frac{1}{2},j} := \frac{1}{36}\big(& f(Q^n_{i+\frac{1}{2},j-\frac{1}{2}}) + 4f(Q^n_{i+\frac{1}{2},j}) + f(Q^n_{i+\frac{1}{2},j+\frac{1}{2}}) \\ &+4f(Q^{n+\frac{1}{2}}_{i+\frac{1}{2},j-\frac{1}{2}}) + 16f(Q^{n+\frac{1}{2}}_{i+\frac{1}{2},j}) + 4f(Q^{n+\frac{1}{2}}_{i+\frac{1}{2},j+\frac{1}{2}}) \\ &+f(Q^{n+1}_{i+\frac{1}{2},j-\frac{1}{2}}) + 4f(Q^{n+1}_{i+\frac{1}{2},j}) + f(Q^{n+1}_{i+\frac{1}{2},j+\frac{1}{2}})\big). \end{aligned} \tag{4}$$

Analogously, the flux at a horizontal grid cell interface can be computed. The crucial part of the Active Flux method is the update of the point values at the intermediate and new time level. For advection with advection speeds $a, b \in \mathbb{R}$, i.e. for scalar equations of the form

$$\partial_t q + a\partial_x q + b\partial_y q = 0, \tag{5}$$

we can again use the exact evolution formula $q(x, y, t_n + \tau) = q(x - a \cdot \tau, y - b \cdot \tau, t_n)$, $\tau \in \{\frac{\Delta t}{2}, \Delta t\}$. In order to use this exact evolution formula, we need a reconstruction of the numerical solution at time t_n. For each two-dimensional Cartesian grid cell a quadratic reconstruction of the form

$$q_{ij}^{rec}(\xi, \eta) = \sum_{k=1}^{9} c_k N_k(\xi, \eta) \tag{6}$$

with coefficients $c_1, \dots, c_9$ and basis functions $N_1, \dots, N_9$ given in Table 3 of [13] is used. The reconstruction interpolates the point values depicted in Fig. 1 and preserves the cell average.

For the advection equation, the flux computation can now be interpreted as an integration of the reconstructed function over a rhomboid as indicated in Fig. 2 (left). An important difference to the one-dimensional case is that in the multi dimensional case Simpson's rule is no longer exact for the reconstructed function. The reason is that the rhomboid will in general lie in two neighbouring grid cells and the integral over a piecewise quadratic function can in general not be computed exactly with Simpson's rule. Thus, we introduce two approximations when we compute numerical fluxes. The first approximation is introduced by our piecewise quadratic reconstruction and the second approximation is introduced by using Simpson's rule. This has consequences for the linear stability of the Active Flux method. Furthermore, it makes limiting of the Active Flux method more difficult in the multi-dimensional case. Both aspects will now be discussed in more detail.

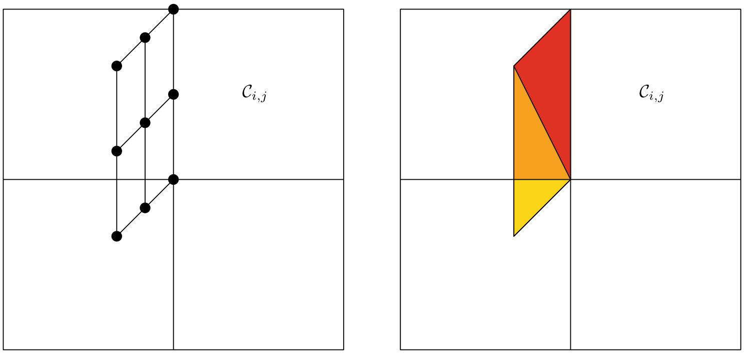


Fig. 2 Flux computation for advective transport using (left) Simpson's rule and (right) exact integration

2.1 Linear Stability of the Active Flux Method

In [7], we studied the linear stability of the Cartesian grid Active Flux method. Here we briefly present our results for the two-dimensional advection equation (5) with double periodic boundary conditions on an equidistant grid with $\Delta x = \Delta y$. Under these conditions we write the method in the form

$$\mathbf{Q}^{n+1} = A\mathbf{Q}^n, \tag{7}$$

where $\mathbf{Q}^n \in \mathbb{R}^{4m^2}$ consists of all degrees of freedom at the time level t_n, i.e. all point value degrees of freedom and all cell average values, and $A \in \mathbb{R}^{4m^2 \times 4m^2}$ describes the update of the Active Flux method during one time step.

A linear method of the form (7) is Lax-Richtmyer stable, iff $\|A^n\|$ is bounded independently of n. Here $\|\cdot\|$ is the spectral norm. The method is stable if all eigenvalues λ of A satisfy the condition $|\lambda| \leq 1$ and in addition if $|\lambda| = 1$ the geometric and algebraic multiplicity need to match. See [8] for more details.

In Fig. 3 (first row) we show the eigenvalues for the Active Flux method for the advection equation (5) using $a = b$ and different time step restrictions. While this is not a general stability proof, we observe that on this particular grid there exist eigenvalues with magnitude larger than one for time steps satisfying CFL $= 0.9$, i.e. the computation on this grid is unstable. For CFL $= 0.75$ the magnitude of the eigenvalues is bounded by one and the computation is stable.

In [7] we systematically investigated different advection speeds and always obtained stable results as long as the CFL number is bounded by 0.75. For the special cases $b = 0$ or $a = 0$ the method is stable as long as the time step is bounded by CFL ≤ 1. In practical computations we typically use CFL ≤ 0.7.

In the second row of Fig. 3 we show the eigenvalues for the Active Flux method with exact integration for the flux computation. In this case the method is stable for time steps satisfying CFL ≤ 1.

In [13] we showed the eigenvalues for the one-dimensional Active Flux method applied to the advection equation. In this case the magnitude of the eigenvalues is bounded by one as long as the time steps satisfy CFL ≤ 1.

2.2 Bound Preserving Reconstructions for Active Flux Methods

The continuous, piecewise quadratic Active Flux reconstruction might introduce new extrema. This can lead to unphysical oscillations, in particular near discontinuities or shock waves.

In [3, 13] several bound preserving reconstructions for the one-dimensional Active Flux method have been presented. Possible reconstructions presented in [13] include a hyperbolic reconstruction as well as a piecewise polynomial reconstruction, con-

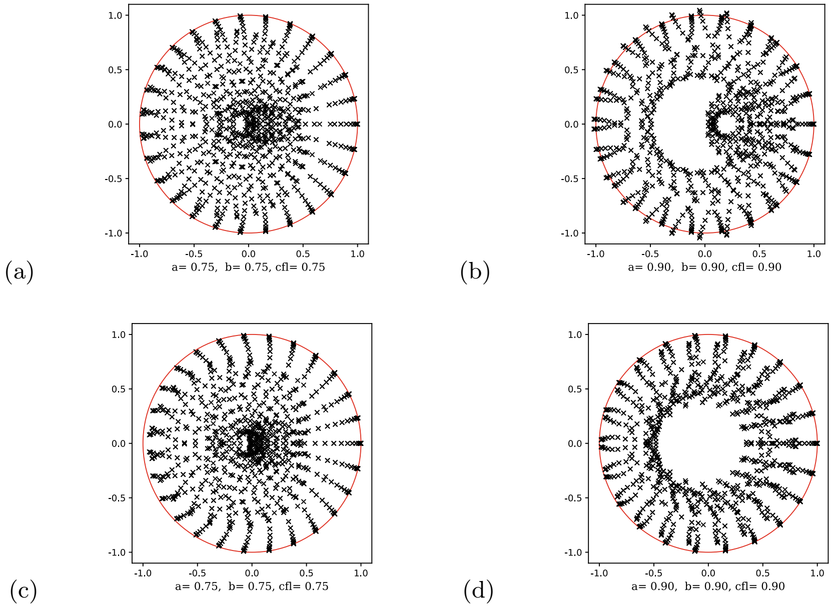


Fig. 3 Eigenvalues for the matrix A describing one time step of the Cartesian grid Active Flux method with $a = b$, $\Delta x = \Delta y = 1/20$, CFL $= 0.75$ and CFL $= 0.9$. (First row) flux computation using Simpson's rule and (second row) flux computation using exact integration

sisting of a constant part and a quadratic part. Both of these limited reconstructions preserve the cell average and interpolate the point values at the grid cell boundary. In [3, Theorem 6] the so-called power law limiting was introduced, which has the form

$$p_N(x) = Q^n_{i-\frac{1}{2}} + (Q^n_{i+\frac{1}{2}} - Q^n_{i-\frac{1}{2}}) \left(\frac{x - x_i + \Delta x/2}{\Delta x} \right)^N \quad x_{i-\frac{1}{2}} \leq x \leq x_{i+\frac{1}{2}},$$

with $N = \frac{Q^n_{i+\frac{1}{2}} - Q^n_i}{Q^n_i - Q^n_{i-\frac{1}{2}}}$. This reconstruction is monotone, preserves the mean and interpolates the point values. While all of those reconstructions efficiently limited oscillations, as illustrated by several simulations that can be found in [3, 13], the new cell average values might not be exactly bound preserving even for linear advection. The reason is that Simpson's rule is not necessarily exact for those reconstructions. Furthermore, there is no straight forward way to extend these reconstructions to the multi-dimensional case.

These shortcomings motivated us to introduce bound preserving limited reconstructions based on previous work of Zhang and Shu [7, 23]. In the two-dimensional case, let

$$M_{ij} := \max_{(x,y)\in C_{ij}} q_{ij}^{rec}(x,y), \quad m_{ij} := \min_{(x,y)\in C_{ij}} q_{ij}^{rec}(x,y),$$

with q_{ij}^{rec} as in (6), denote the minimum and the maximum of the Active Flux reconstruction in grid cell C_{ij}. Furthermore, $\bar{M}_{ij}$ and $\bar{m}_{ij}$ denote the maxima and minima of all the point values of q used for the Active Flux reconstruction in grid cell C_{ij}, i.e.

$$\begin{aligned}\bar{M}_{ij} &:= \max\big\{Q_{i-\frac{1}{2},j-\frac{1}{2}}, Q_{i-\frac{1}{2},j}, Q_{i-\frac{1}{2},j+\frac{1}{2}}, Q_{i,j+\frac{1}{2}},\\ &\qquad Q_{i+\frac{1}{2},j+\frac{1}{2}}, Q_{i+\frac{1}{2},j}, Q_{i+\frac{1}{2},j-\frac{1}{2}}, Q_{i,j-\frac{1}{2}}\big\}\\ \bar{m}_{ij} &:= \min\big\{Q_{i-\frac{1}{2},j-\frac{1}{2}}, Q_{i-\frac{1}{2},j}, Q_{i-\frac{1}{2},j+\frac{1}{2}}, Q_{i,j+\frac{1}{2}},\\ &\qquad Q_{i+\frac{1}{2},j+\frac{1}{2}}, Q_{i+\frac{1}{2},j}, Q_{i+\frac{1}{2},j-\frac{1}{2}}, Q_{i,j-\frac{1}{2}}\big\}.\end{aligned}$$

For $Q_{ij}^n \in [\bar{m}_{ij}, \bar{M}_{ij}]$ the limited reconstruction has the form

$$\tilde{q}_{ij}^{rec} = \theta(q_{ij}^{rec}(x,y) - Q_{ij}^n) + Q_{ij}^n, \text{ with } \theta = \min\left\{\left|\frac{\bar{M}_{ij} - Q_{ij}^n}{M_{ij} - Q_{ij}^n}\right|, \left|\frac{\bar{m}_{ij} - Q_{ij}^n}{m_{ij} - Q_{ij}^n}\right|, 1\right\}. \tag{8}$$

If $Q_{ij}^n \notin [\bar{m}_{ij}, \bar{M}_{ij}]$, we do not apply limiting in order not to decrease the accuracy near local extrema.

In each grid cell, this limited reconstruction is a single quadratic function, which preserves the cell average and satisfies $\bar{m}_{ij} \le q_{ij}^{rec}(x,y) \le \bar{M}_{ij}$ for all $(x,y) \in C_{ij}$. Although the reconstruction is no longer globally continuous, we can still apply the exact evolution operators. When computing the numerical fluxes, we however need to decide which point values are used at the old time level. We use the point values from the limited reconstruction in upwind direction. For smooth solutions applying the limiter does not decrease the order of convergence. The application of this limiting strategy in the one- and three-dimensional case is straight forward. We obtain the following result.

Theorem 1 *The one-dimensional Active Flux method for advective transport with bound preserving piecewise quadratic reconstruction does not produce new minima or maxima in the new cell average values.*

Proof In the one-dimensional case, the flux computation using Simpson's rule is exact for a reconstruction, which consists of a single parabula in each grid cell. Thus, the new cell average values agree with exact averages over parts of the bound preserving reconstruction and thus are bound preserving.

Remark 1 The result does not extend to the two- and three-dimensional case, since the flux computation using Simpson's rule is no longer exact.

For the two-dimensional advection equation the numerical flux can easily be computed exactly by integrating over the triangular regions indicated in the right plot of

Fig. 2. This would lead to a bound preserving approximation of the cell average values. However, such an approach can not be extended to more general hyperbolic problems. Therefore, we are currently exploring alternative approaches which preserve prescribed bounds of the solution. In practical applications the preservation of positivity is often an important requirement for numerical schemes. For example, when approximating the Euler equations of gas dynamics, pressure and density need to remain positiv. Hu et al. [14] proposed to combine a high-order accurate flux with the low order bound preserving Lax-Friedrichs flux approximation. More precisely, the idea is to write the finite volume update of the cell averages (3) in the form

$$\begin{aligned} Q_{ij}^{n+1} &= \frac{1}{4}\left(Q_{ij}^n + 4\frac{\Delta t}{\Delta x}F_{i-\frac{1}{2},j}\right) + \frac{1}{4}\left(Q_{ij}^n - 4\frac{\Delta t}{\Delta x}F_{i+\frac{1}{2},j}\right) \\ &\quad + \frac{1}{4}\left(Q_{ij}^n + 4\frac{\Delta t}{\Delta y}G_{i,j-\frac{1}{2}}\right) + \frac{1}{4}\left(Q_{ij} - 4\frac{\Delta t}{\Delta y}G_{i,j+\frac{1}{2}}\right). \end{aligned}$$

If one of the four terms on the right hand side becomes negative, then the corresponding numerical flux is replaced by a convex combination of the Lax-Friedrichs flux and the Active Flux flux.

We used this new approach to approximate solutions of the advection equation (5) with advection speeds $a = 0.5$, $b = 0.25$ on a grid with 128×128 mesh cells discretising the unit square. The initial values consist of a Gaussian hump and a square shaped discontinuity.

By using the bound preserving reconstruction (8) and the limiter of Hu et al. [14], no new minima are introduced. Numerical results are shown in Fig. 4.

By using the standard form of the Active Flux method with bound preserving reconstruction (8) and Simpson's rule to compute numerical fluxes, we observed undershoots or overshoots in the cell average values of size $\approx 10^{-3}$.

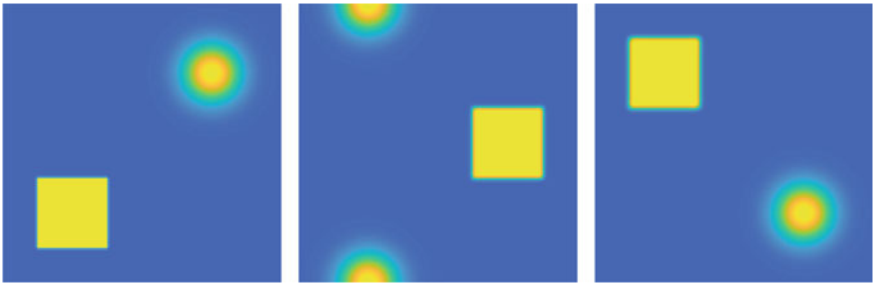


Fig. 4 Approximation of the advection equation using the bound preserving approach. The solution is shown at $t = 0$, $t = 0.5$ and $t = 1$

3 Multi-dimensional Approximation of Point Values

In this section we present different exact and approximative truly multi-dimensional evolution formulas that have successfully been used for the update of the point values in two-dimensional Active Flux methods.

3.1 Advective Transport in a Divergence Free Velocity Field

We consider a scalar hyperbolic problem of the form

$$\partial_t q + \partial_x \left(a(x, y, t)q(x, y, t)\right) + \partial_y \left(b(x, y, t)q(x, y, t)\right) = 0, \tag{9}$$

with given functions a and b that satisfy $\partial_x a(x, y, t) + \partial_y b(x, y, t) = 0$.

The numerical flux functions of the Active Flux method require point values of q at the intermediate and new time level, which can be computed using the characteristic curves

$$\begin{aligned} x'(t) &= a(x(t), y(t), t) \\ y'(t) &= b(x(t), y(t), t) \end{aligned}$$

of the partial differential equation. Starting with the position of the point value, the ode system for the characteristics is solved backwards in time for a half or a full time step. The point values of the conserved quantity q are obtained by evaluating the Active Flux reconstruction (6) at the foot point of the associated characteristic curve. Together with Kiechle and Chudzik, see [15], we used this approach to obtain a third order accurate method for the Vlasov-Poisson system.

3.2 Burgers' Equation

For the two-dimensional Burgers' equation

$$\partial_t q(x, y, t) + \partial_x \left(\frac{1}{2}q^2\right) + \partial_y \left(\frac{1}{2}q^2\right) = 0, \tag{10}$$

with $q : \mathbb{R}^2 \times \mathbb{R}^+ \to \mathbb{R}$, approximative evolution formulas for the update of the point values have been presented in [7].

For smooth solutions (10) can equivalently be written in the advective form

$$\partial_t q + q\partial_x q + q\partial_y q = 0,$$

which suggests the implicitly defined solution

$$q(x, y, t) = q(x - q(x, y, t)t, y - q(x, y, t)t, 0).$$

Starting with an appropriate initial guess, we iteratively compute

$$\begin{aligned}
\left(Q^{n+\frac{1}{2}}_{i-\frac{1}{2},j}\right)^{\ell} &= q^{rec}\left(x_{i-\frac{1}{2}} - \left(Q^{n+\frac{1}{2}}_{i-\frac{1}{2},j}\right)^{\ell-1}\frac{\Delta t}{2}, y_j - \left(Q^{n+\frac{1}{2}}_{i-\frac{1}{2},j}\right)^{\ell-1}\frac{\Delta t}{2}\right)\\
\left(Q^{n+\frac{1}{2}}_{i-\frac{1}{2},j-\frac{1}{2}}\right)^{\ell} &= q^{rec}\left(x_{i-\frac{1}{2}} - \left(Q^{n+\frac{1}{2}}_{i-\frac{1}{2},j-\frac{1}{2}}\right)^{\ell-1}\frac{\Delta t}{2}, y_{j-\frac{1}{2}} - \left(Q^{n+\frac{1}{2}}_{i-\frac{1}{2},j-\frac{1}{2}}\right)^{\ell-1}\frac{\Delta t}{2}\right)\\
\left(Q^{n+\frac{1}{2}}_{i,j-\frac{1}{2}}\right)^{\ell} &= q^{rec}\left(x_i - \left(Q^{n+\frac{1}{2}}_{i,j-\frac{1}{2}}\right)^{\ell-1}\frac{\Delta t}{2}, y_{j-\frac{1}{2}} - \left(Q^{n+\frac{1}{2}}_{i,j-\frac{1}{2}}\right)^{\ell-1}\frac{\Delta t}{2}\right)
\end{aligned}$$

and analogously for the full time step. The initial guess is computed from neighbouring cell average values, i.e. we set

$$\begin{aligned}
\left(Q^{n+\frac{1}{2}}_{i-\frac{1}{2},j}\right)^{0} &= \left(Q^{n+1}_{i-\frac{1}{2},j}\right)^{0} = \frac{1}{2}\left(Q^n_{i-1,j} + Q^n_{i,j}\right)\\
\left(Q^{n+\frac{1}{2}}_{i-\frac{1}{2},j-\frac{1}{2}}\right)^{0} &= \left(Q^{n+1}_{i-\frac{1}{2},j-\frac{1}{2}}\right)^{0} = \frac{1}{4}\left(Q^n_{i-1,j} + Q^n_{i,j} + Q^n_{i-1,j-1} + Q^n_{i,j-1}\right)\\
\left(Q^{n+\frac{1}{2}}_{i,j-\frac{1}{2}}\right)^{0} &= \left(Q^{n+1}_{i,j-\frac{1}{2}}\right)^{0} = \frac{1}{2}\left(Q^n_{i,j-1} + Q^n_{i,j}\right).
\end{aligned}$$

Using Taylor series expansion, it can be shown that each iteration improves the accuracy by one order, see [7]. Since the initial guess is a first order accurate approximation, it is enough to iterate twice.

Once the point values at the intermediate and new time level have been computed the numerical fluxes can be obtained using (4) and the cell average values can be evolved using (3). Numerical results are shown in Sect. 4.

3.3 Linear Acoustic Equations

For the acoustic equations

$$\partial_t p + c\nabla \cdot \mathbf{u} = 0 \tag{11}$$

$$\partial_t \mathbf{u} + c\nabla p = 0, \tag{12}$$

exact evolution operators for the update of the point values have been proposed by Eymann and Roe [10] for irrotational flow problems and by Fan and Roe [11] and Barsukow et al. [4] for the general case.

A form, which is particularly useful for the implementation, was given in [4]:

$$p(\mathbf{x},t) = \partial_r \left(r M_r\{p(\mathbf{x},0)\}\right)|_{r=ct} - \frac{1}{ct}\partial_r \left(r^2 M_r\{\mathbf{n}\cdot\mathbf{u}(\mathbf{x},0)\}\right)|_{r=ct} \tag{13}$$

$$\begin{aligned}\mathbf{u}(\mathbf{x},t) = \mathbf{u}(\mathbf{x},0) - \frac{1}{ct}\partial_r \left(r^2 M_r\{\mathbf{n}p(\mathbf{x},0)\}\right)|_{r=ct} \\ + \int_0^{ct} \frac{1}{r}\partial_r \left(\frac{1}{r}\partial_r \left(r^3 M_r\{(\mathbf{n}\cdot\mathbf{u}(\mathbf{x},0))\mathbf{n}\right) - r M_r\{\mathbf{u}(\mathbf{x},0)\}\right) dr.\end{aligned} \tag{14}$$

Here M_r denotes the spherical mean over a disc with radius r. In this representation of the exact solution all derivative terms are expressed as derivatives in radial direction. This avoids to work with delta functions.

Together with Maria Lukáčová we currently investigate Active Flux methods for linear hyperbolic systems using the method of bicharacteristics in order to compute the evolution of the point values. This is based on earlier work of Lukáčová-Medvid'ová et al. [16, 17]. For the two-dimensional linear acoustic equations we use a third order accurate approximative evolution operators of the form

$$p(P) \approx \frac{1}{\pi}\int_0^{2\pi} \left(p(Q(\theta)) - u(Q(\theta))\cos(\theta) - v(Q(\theta))\sin(\theta)\right) d\theta - p(P') \tag{15}$$

$$\begin{aligned}u(P) \approx \frac{1}{\pi}\int_0^{2\pi} \Big(- p(Q(\theta))\cos(\theta) + u(Q(\theta))\left(2\cos^2(\theta) - \frac{1}{2}\right) \\ +2v(Q(\theta))\sin(\theta)\cos(\theta)\Big) d\theta\end{aligned} \tag{16}$$

$$\begin{aligned}v(P) \approx \frac{1}{\pi}\int_0^{2\pi} \Big(- p(Q(\theta))\sin(\theta) + 2u(Q(\theta))\sin(\theta)\cos(\theta) \\ +v(Q(\theta))\left(2\sin^2(\theta) - \frac{1}{2}\right)\Big) d\theta,\end{aligned} \tag{17}$$

where $P := (\bar{x}, \bar{y}, t_n + \tau)$ is the point at which we want to compute the solution, $P' := (\bar{x}, \bar{y}, t_n)$ and $Q(\theta) := (\bar{x} + c\cdot\tau\cos(\theta), \bar{y} + c\cdot\tau\sin(\theta), t_n)$. The resulting formulas are easier to compute than the exact evolution formulas. For acoustics we observe an accuracy that compares well with the accuracy of the Active Flux method with exact evolution operator. Test calculations will be shown in Sect. 4. This motivates us to further explore the use of the method of bicharacteristics for further linear hyperbolic systeme.

4 Cartesian Grid Active Flux Methods with Adaptive Mesh Refinement

Together with Donna Calhoun [6], we developed Active Flux methods for Cartesian grids with adaptive mesh refinement. We implemented the method in ForestClaw [5], which is a software for parallel adaptive mesh refinement of patch-based solvers. It turns out that the Active Flux method is well suited for the use on adaptively refined grids. Here we briefly review our approach.

A typical ForestClaw grid structure is shown in Fig. 5. Each patch typically consists of 8×8 or 16×16 grid cells and two additional rows and columns of ghost cells. Ghost cells are used to implement the transfer of data between grid patches. Along the physical boundary they are also used to implement boundary conditions. The following transfer operators are needed

1. A transfer from fine grids to coarse grids is needed if grid patches are coarsened. The same approach is used to compute ghost cell values of coarse grids from neighbouring fine grids.
   This transfer can easily be implemented, since the fine mesh contains all the information needed on the coarse grid. The cell average values on the coarse grid are obtained by averaging cell average values of the underlying fine grids. The appropriate point values are copied from the fine grids.
2. A transfer from a coarse grid to a fine grid is needed if a patch is marked for refinement. The same approach is used to compute ghost cell values of a fine mesh from a neighbouring coarse patch.
   We use the degrees of freedom of the coarse grid to compute the Active Flux reconstruction. From this reconstruction we compute the cell average values and the point values on the fine grid.
3. For neighbouring grid patches of the same refinement level the ghost cell information can be copied from the neighbouring grid.

In addition to local refinement and coarsening, our adaptively refined Active Flux method allows sub-cycling in time, i.e. while a single time step is taken on the coarse grid, several smaller time steps may be taken on the finer grids. In order to allow sub-cycling we need two rows and columns of ghost cells. Without sub-cycling one row and column of ghost cells would be sufficient due to the local stencil of the method.

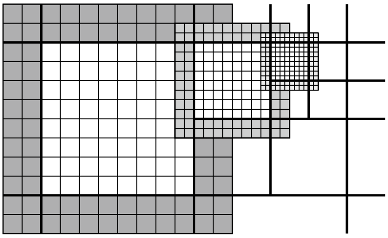


Fig. 5 ForestClaw grid structure with Cartesian grid patches of three different refinement levels [6]

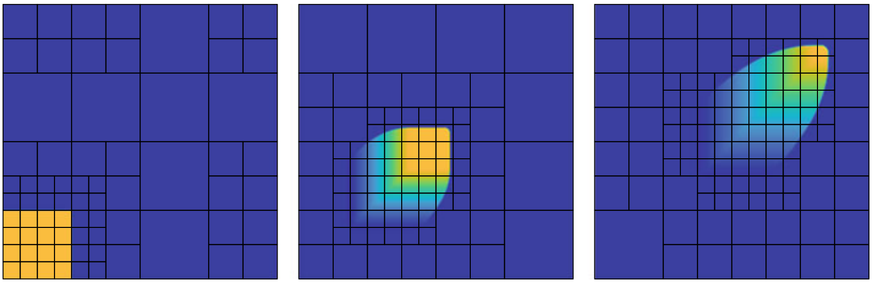


Fig. 6 Approximation of Burgers equation using the Active Flux method with adaptive mesh refinement and bound preserving limiter

All the details can be found in [6]. Numerical convergence studies, documented in [6], confirm third order accuracy.

In our first test computation we consider solutions of Burgers' equation with piecewise constant initial values as shown in the left plot of Fig. 6. The initial values are

$$q(x, y, 0) = \begin{cases} 2 : 0 \leq x, y \leq 1/4 \\ 1 : \quad \text{otherwise.} \end{cases}$$

At later times shock waves and rarefaction waves develop as shown in the next two plots. Here we used adaptively refined patches of Cartesian grids on three different levels.

In the second test computation we consider the approximation of a stationary vortex modeled by the two-dimensional acoustic equations (11) and (12). This test problem was proposed by Barsukow et al. [4]. It is similar to the Gresho vortex problem [12] but with $\nabla p = 0$ beside $\nabla \cdot \mathbf{u} = 0$ to obtain stationary solutions of the acoustic equations. The solution has the form

$$p(r) = 0, \quad \mathbf{u}(r) = \mathbf{n} \begin{cases} 5r & : \ 0 \leq r \leq 0.2 \\ 2 - 5r & : 0.2 < r \leq 0.4 \\ 0 & : \quad r > 0.4 \end{cases}$$

with $r = \sqrt{x^2 + y^2}$, $\mathbf{u} = (u, v)^T$, $\mathbf{n}(\theta) = (-\sin(\theta), \cos(\theta))^T$ and $\theta \in [0, 2\pi)$. We start with a discretisation of this profile on the adaptively refined grid and compute solutions at time $t = 100$. Here the Active Flux method with the approximative evolution operator (15)–(17) was used to update the point values. Figure 7 shows $|\mathbf{u}|$ at the final time both as two-dimensional plot that also shows the grid patches of the adaptively refined mesh as well as as scatter plot. While the method does not exactly preserves the steady state, the numerical results are very accurate.

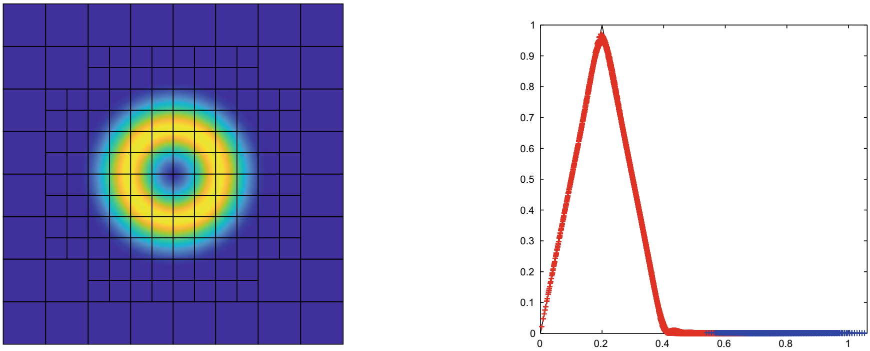


Fig. 7 Approximation of the vortex problem using the Active Flux method with adaptive mesh refinement. The point values are evolved in time using the approximate evolution operator (11)–(12). We show $|\mathbf{u}|$ at time $t = 100$ in a two-dimensional plot, which also indicates the grid patches, and in a scatter plot. The solid line in the right plot shows the exact solution

In our final test we perform a convergence study for the Lukácǒvá acoustic test problem [16], for which the solution has the form

$$\begin{aligned} p(x, y, t) &= -\frac{1}{c}\cos(2\pi ct)\,(\sin(2\pi x) + \sin(2\pi y)) \\ u(x, y, t) &= \frac{1}{c}\sin(2\pi ct)\cos(2\pi x) \\ v(x, y, t) &= \frac{1}{c}\sin(2\pi t)\cos(2\pi y). \end{aligned}$$

We consider this problem on the domain $[-1, 1] \times [-1, 1]$ with periodicity condition and $c = 1$. Starting with the data at time $t = 0$ we compute solutions at time $t = 1$ using Active Flux methods based on the exact evolution operator (13), (14) as well as the approximative evolution operator (15)–(17). The results of a convergence study, comparing the numerical solution with the exact solution, are documented in Tables 1 and 2. The Active Flux method with exact evolution operator is only slightly more accurate. However, the currently used approximative evolution operator, derived by

Table 1 Error at time $t = 1$ measured in the $\|\cdot\|_1$-norm and EOC for the Lukácǒvá test problem using exact evolution with CFL = 0.275

Level	Error		EOC	
	p	u, v	p	u, v
2	3.328532e-04	2.628078e-05	–	–
3	4.180608e-05	3.276039e-06	2.9931	3.0040
4	5.232828e-06	4.564793e-07	2.9981	2.8433

Table 2 Error at time $t = 1$ measured in the $\| \cdot \|_1$-norm and EOC for the Lukáčová test problem using the EG2 evolution operator with CFL = 0.275

$m_x = m_y$	Error			EOC		
	p	u	v	p	u	v
64	3.417769e-04	3.818359e-05	3.797061e-05	–	–	–
128	4.286948e-05	4.622912e-06	4.596061e-06	2.9950	3.0461	3.0464
256	5.365616e-06	5.793141e-07	5.746239e-07	2.9981	2.9964	2.9997

using the method of bicharacteristics, reduces the stability of the resulting Active Flux method.

5 Conclusions

We have presented a brief overview of the Active Flux method, an evolving finite volume method for hyperbolic problems.

An important component of the Active Flux method is an exact or approximative evolution operator for the update of the point values. Different truly multidimensional evolution operators have been discussed.

Even for scalar problem, where exact or approximated evolution operators are easily available, there is an additional difficulty by going from the one to the multidimensional case due to the fact that Simpson's rule, which is used for the flux computation, is not exact when integrating a piecewise quadratic function. We have shown that this reduces the stability of the Active Flux method and makes the limiting process more difficult.

Finally, we illustrated that Active Flux methods can easily be used on adaptively refined Cartesian grid patches.

References

1. Abgrall, R.: A combination of residual distribution and the active flux formulation or a new class of schemes that can combine several writings of the same hyperbolic problem: application to the 1d Euler equations. Commun. Appl. Math. Comput. **5**, 370–402 (2023)
2. Abgrall, R., Barsukow, W.: Extensions of active flux to arbitrary order of accuracy (2022). arXiv: 2208.14476
3. Barsukow, W.: The active flux method for nonlinear problems. J. Sci. Comput. **86**, 3 (2021)
4. 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)

5. Calhoun, D., Burstedde, C.: ForestClaw: a parallel algorithm for patch-based adaptive mesh refinement on a forest of quadtrees (2017). arXiv: 1703.03116
6. Calhoun, D., Chudzik, E., Helzel, C.: The Cartesian grid active flux method with adaptive mesh refinement. J. Sci. Comput. **94** (2023)
7. Chudzik, E., Helzel, C., Kerkmann, D.: The Cartesian grid active flux method: linear stability and bound preserving limiting. Appl. Math. Comput. **393**, 125501 (2021)
8. van Drosselaer, J., Kraaijevanger, J., Spijker, M.: Linear stability analysis in the numerical solution of initial value problems. Acta Numer. **2**, 199–237 (1993)
9. Eymann, T.A., Roe, P.L.: Active flux schemes for systems. AIAA 2011-3840 (2011)
10. Eymann, T.A., Roe, P.L.: Multidimensional active flux schemes. In: AIAA Conference Paper (2013)
11. Fan, D., Roe, P.L.: Investigations of a new scheme for wave propagation. In: AIAA Aviation Forum (2015)
12. Gresho, P.M., Chan, S.T.: On the theory of semi-implicit projection methods for viscous incompressible flow and its implementation via a finite element method that also introduces a nearly consistent mass matrix: part 2: Implementation. Int. J. Numer. Methods Fluids **11**, 621–659 (1990)
13. Helzel, C., Kerkmann, D., Scandurra, L.: A new ADER method inspired by the active flux method. J. Sci. Comput. **80**, 1463–1497 (2019)
14. Hu, X.Y., Adams, N.A., Shu, C.-W.: Positivity-preserving method for high-order conservative schemes solving compressible Euler equations. J. Comput. Phys. **242**, 169–180 (2013)
15. Kiechle, Y.-F., Chudzik, E., Helzel, C.: An active flux method for the Vlasov-Poisson system. accepted for publication in FVCA 2023
16. Lukáčová-Medvid'ová, M., Morton, K.W., Warnecke, G.: Evolution Galerkin methods for hyperbolic systems in two space dimensions. Math. Comput. **69**, 1355–1384 (2000)
17. Lukáčová-Medvid'ová, M., Saibertová, J., Warnecke, G.: Finite volume evolution Galerkin methods for nonlinear hyperbolic systems. J. Comput. Phys. **183**, 533–562 (2002)
18. Roe, P.: Is discontinuous reconstruction really a good idea? J. Sci. Comput. **73**, 1094–1114 (2017)
19. Roe, P.: A simple explanation of superconvergence for discontinuous Galerkin solutions to $u_t + u_x = 0$. Commun. Comput. Phys. **21**, 905–912 (2017)
20. Roe, P.L., Maeng, J., Fan, D.: Comparing active flux and discontinuous Galerkin methods for compressible flow. In: 2018 AIAA Aerospace Science Meeting
21. Roe, P.: Designing CFD methods for bandwidth—a physical approach. Comput. Fluids **214**, 104774 (2021)
22. Samani, I., Roe, P.: Acoustics on a coarse grid. In: AIAA Scitech 2023 Forum
23. Zhang, X., Shu, C.-W.: Maximum-principle-satisfying and positivity-preserving high-order schemes for conservation laws: survey and new developments. Proc. R. Soc. A **467**, 2752–2776 (2011)

Moving-Mesh Finite-Volume Methods for Hyperbolic Interface Dynamics

Christian Rohde

Abstract The numerical discretization of continuum-mechanical free boundary value problems for hyperbolic conservation laws becomes challenging when the dynamics of the interface depend sensitively on smaller-scale effects. A proper tracking of the interface and an efficient solution of the conservation laws in the bulk domains can be realized by a heterogeneous multi-scale ansatz combined with recently introduced moving-mesh concepts for finite-volume methods. To illustrate the approach we focus on two applications: the tracking of phase boundaries in compressible liquid-vapour flow and dimensionally mixed models for two-phase flow in fractured porous media. In the first case phase transition effects lead to non-standard interface dynamics. In the latter case the coupling conditions for the bulk domains involve the solution of evolution equations in the fractures which are represented as hypersurfaces.

Keywords Hyperbolic conservation laws · Moving-mesh methods · Two-phase flow · Fracturing porous media

1 Introduction

Free boundary value problems are ubiquitous in the mathematical modelling on the continuum-mechanical level. Applications include the dynamics of multi-phasic systems, self-gravitating fluids, fluid-solid interactions and so on. They also occur artificially when mixed-dimensional models are employed that are derived from fully dimensional but geometrically extreme settings. This applies to fracture propagation or the dynamics of thin fluid layers. We address in this note free boundary value problems that are governed by hyperbolic conservation laws in the bulk domains. Moreover we are interested in complex interfacial motions that can not directly writ-

C. Rohde (✉)
Institute of Applied Analysis and Numerical Simulation and Stuttgart Center for Simulation Science (SC SimTech), University of Stuttgart, Stuttgart, Germany
e-mail: Christian.Rohde@mathematik.uni-stuttgart.de
URL: https://www.ians.uni-stuttgart.de

E. Franck et al. (eds.), *Finite Volumes for Complex Applications X—Volume 1, Elliptic and Parabolic Problems*, Springer Proceedings in Mathematics & Statistics 432,
https://doi.org/10.1007/978-3-031-40864-9_7

ten in e.g. weak-solution frameworks but require to solve an additional transmission problem, which typically involves smaller-scale physics.

We start with a more precise description of the class of free-boundary value problems we address here. Let some domain $D \subset \mathbb{R}^d$ and $T > 0$ be given. By $\mathcal{U} \subset \mathbb{R}^m$ we denote the state space for the unknowns. The functions $\boldsymbol{f}_1, \ldots, \boldsymbol{f}_d : \mathcal{U} \to \mathbb{R}^m$ are smooth flux vectors which are summarized in $\mathbf{F} := (\boldsymbol{f}_1 | \cdots | \boldsymbol{f}_d)$ and supposed to satisfy the hyperbolicity condition

$$\operatorname{spec}\bigl(D\boldsymbol{f}_1(\boldsymbol{u})n_1 + \cdots + D\boldsymbol{f}_d(\boldsymbol{u})n_d\bigr) \subset \mathbb{R} \tag{1.1}$$

for all $\boldsymbol{u} \in \mathcal{U}$ and $\mathbf{n} = (n_1, \ldots, n_d)^T \in \mathcal{S}^{d-1}$.

For each $t \in [0, T]$ we search a co-dimension-1 manifold $\Gamma = \Gamma(t)$ and a function $\boldsymbol{u} = \boldsymbol{u}(\cdot, t) : D_{\pm}(t) \to \mathcal{U} \subset \mathbb{R}^m$ that satisfy an initial-boundary value problem for the system of first-order hyperbolic conservation laws given by

$$\boldsymbol{u}_t + \boldsymbol{f}_1(\boldsymbol{u})_{x_1} + \cdots + \boldsymbol{f}_d(\boldsymbol{u})_{x_d} = \mathbf{0} \text{ in } D_{\pm}(t). \tag{1.2}$$

The interface $\Gamma(t)$ evolves from some initially given manifold Γ_0 that defines the partition $D = D_{0,-} \cap \Gamma_0 \cap D_{0,+}$. For $t \in (0, T]$ the manifold Γ is also assumed to separate the domain D into the disjunct bulk domains $D_{\pm}(t)$ according to

$$D = D_-(t) \cap \Gamma(t) \cap D_+(t). \tag{1.3}$$

Besides appropriate initial/boundary conditions for the bulk unknowns, wellposedness of (1.2) requires to prescribe an evolution law for the interface in each $\boldsymbol{\xi} \in \Gamma(t)$ by its speed $\sigma = \sigma(\boldsymbol{\xi}, t) \in \mathbb{R}$ in normal direction $\mathbf{n} = \mathbf{n}(\boldsymbol{\xi}, t) \in \mathcal{S}^{d-1}$. W.l.o.g. we assume here that $\mathbf{n}$ points into $D_+(t)$. The interfacial dynamics is typically determined from transmission conditions posed across the interface Γ. In an abstract setting and with some appropriate operator $\mathcal{B}$ they take the form

$$\mathcal{B}[\boldsymbol{u}_{\pm}, \sigma; \mathbf{n}](\boldsymbol{\xi}, t) = \mathbf{0}. \tag{1.4}$$

In (1.4) we used the traces $\boldsymbol{u}_{\pm}(\mathbf{x}, t) = \lim_{\varepsilon \to 0} \boldsymbol{u}(\mathbf{x} \pm \varepsilon \mathbf{n}, t)$. A sample configuration for the evolution of the interface and the bulk domains is sketched in Fig. 1.

For hyperbolic conservation laws (1.2), the obvious purely algebraic choice of the Rankine-Hugoniot conditions

$$\mathcal{B}[\boldsymbol{u}_{\pm}, \sigma; \mathbf{n}] = -\sigma [\![\boldsymbol{u}]\!] + [\![\boldsymbol{f}_1(\boldsymbol{u})n_1 + \cdots + \boldsymbol{f}_d(\boldsymbol{u})n_d]\!] \tag{1.5}$$

ensures the conservation of the unknown's components across $\Gamma(t)$. Note that we used the notation $[\![\mathbf{w}]\!] = \mathbf{w}_+ - \mathbf{w}_-$ for the trace difference of some field $\mathbf{w}$ on D. The Rankine-Hugoniot conditions are a necessary condition for $\boldsymbol{u}$ to be a weak solution of (1.2) on the entire domain D. In this case there are a multitude of

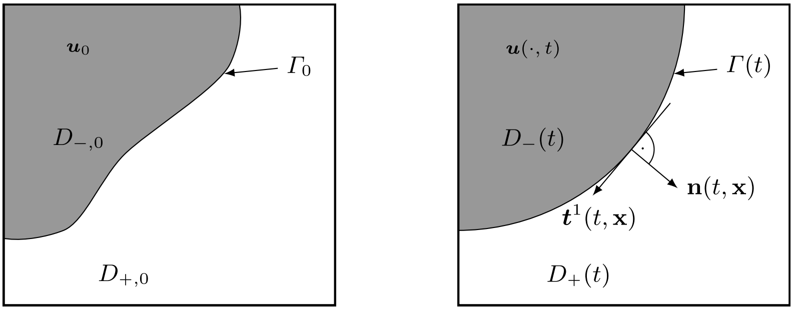

$\boldsymbol{u}_0$
Γ_0
$D_{-,0}$
$D_{+,0}$
$\boldsymbol{u}(\cdot,t)$
$\Gamma(t)$
$D_-(t)$
$\mathbf{n}(t,\mathbf{x})$
$\boldsymbol{t}^1(t,\mathbf{x})$
$D_+(t)$

Fig. 1 Sketch of the initial configuration and a snapshot for $t > 0$ ($d = 2$)

finite-volume methods that work properly on mesh topologies that do not support an explicit representation of the interface $\Gamma(t)$. We may refer to [4, 13] for more recent overviews only.

Rather, in this note we are interested in free boundary value problems of type (1.2), (1.4) that involve substantially more *complex interfacial dynamics*. Multiphase/multicomponent flows or flows in heterogeneous porous media provide suchlike scenarios. The numerical solution approach might then require an explicit tracking of the interface. To face the corresponding discretization challenge we have developed a versatile moving-mesh method that can be combined with standard finite-volume method: the moving-mesh finite-volume (MMFV) method. The moving-mesh algorithm resolves a time-varying $(d-1)$-dimensional manifold directly within the d-dimensional mesh, which means that the interface is represented in each discrete point of time by a subset of moving-mesh cell-surfaces. The underlying mesh is a conforming simplicial partition that fulfills the Delaunay property. It includes as core novelty local re-meshing algorithms that keep the (initially given) quality of the mesh. Open-source implementations of the moving-mesh algorithms in 3D are available via [2, 5]. Moving mesh methods are frequently used in fluid mechanics. However, we focus on multi-dimensional methods that include the approximation of interfaces like in e.g. [11, 21, 22, 25]. Up to our knowledge, none of these methods combines the following properties:

- the interface is explicitly preserved as a $(d-1)$-dimensional manifold in the moving mesh for all times,
- the MMFV method ensures by re-meshing that the quality of the initial mesh is preserved in time,
- the re-meshing routines are (in generic cases) restricted to local changes close to the interface,
- the MMFV method allows the approximation of interfacial motion with normal speeds different to fluid velocities or interface tip evolution.

After a short overview on the MMFV method in Sect. 2 we will report on two specific applications with complex interfaces. In Sect. 3 we consider compressible liquid-vapour flow with phase transitions. In this case the interfacial speed is not

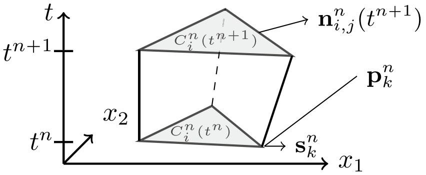


Fig. 2 Affine deformation of a volume $C_i^n(t)$ with vertex $\mathbf{p}_k^n$ in $[t^n, t^{n+1}]$ $(d = 2)$

determined from an algebraic system of equations like (1.5) but requires to solve a micro-scale molecular dynamics system. Different from this set-up is the application in Sect. 4 on two-phase flow in fractured porous media. The evolution in the two bulk porous media domains is then coupled by a transmission condition (1.4), that involves an evolution equation on the interface $\Gamma(t)$ itself. These sections base on the material in [5, 9, 19].

2 A Moving-Mesh Finite-Volume Method for Interface Tracking

Finite-volume methods on moving meshes in one space dimension have been first introduced for hyperbolic conservation laws by Harten and Hyman in [16]. For multiple space dimensions the development started with e.g. [12, 24] leading to a by now too large to cite amount of literature on moving-mesh methods for hyperbolic conservation laws. Most of the works refer to the design of error-optimizing meshes (see e.g. [17] for an overview) but much less has been done concerning the explicit tracking of interfaces as part of the solution of free-boundary value problems like (1.2), (1.4). This section is devoted to the design of such a moving-mesh finite-volume (MMFV) method. Precisely, the method will capture the discrete interface in each point of time as a family of mesh facets of a Delaunay-type unstructured mesh. In the discrete bulk domains, the partial differential equations (1.2) are solved by (time-explicit) finite-volume schemes leading to a piecewise constant approximation of the solution at each time point. The crucial challenge for the design of a corresponding moving mesh is then to prevent the formation of small-size cells that would deteriorate the time-step via the CFL-condition and render the entire approach inefficient. We achieve the preservation of the mesh quality by aligning the discrete interface tracking with a local re-meshing procedure.

In the remainder of the section we review our MMFV approach that roots for the one-dimensional and two-dimensional setting in [8, 9], respectively, and has just recently been extended to three space dimensions [2]. To introduce this MMFV method we set for the sake of simplicity $D = \mathbb{R}^d$, and let for $N \in \mathbb{N}$ a time partition $t^0 = 0 < t^1 < \cdots < t^N = T$ of $[0, T]$ be given. For $n \in \{0, \ldots, N-1\}$ let $I^n \subset \mathbb{N}$ be some index set. Define $\Delta t^n := t^{n+1} - t^n$ and $t^{n+\frac{1}{2}} := (t^{n+1} + t^n)/2$. A *fixed mesh* $\overline{\mathcal{T}}^n$ on $\mathbb{R}^d$ is then a set $\overline{\mathcal{T}}^n := \{C_i^n \mid C_i^n$ closed $d-$simplex, $i \in I^n\}$ such that the volumes C_i^n cover $\mathbb{R}^d$, and such that we have either for the $(d-1)$-dimensional

Hausdorff measure $\mathcal{H}_{d-1}(C_i^n \cap C_j^n) = 0$ or, if $\mathcal{H}_{d-1}(C_i^n \cap C_j^n)_{d-1} \neq 0$, that the set $S_{i,j}^n := C_i^n \cap C_j^n$ is a common facet of C_i^n and C_j^n. We define pair index sets for facets by $E^n = \{(i,j) \in I^n \times I^n \mid \mathcal{H}_{d-1}(S_{i,j}^n) \neq 0\}$, and the mesh width h^n as the maximum of all edges' length. For each $(i,j) \in E^n$ we define $\mathbf{n}_{i,j}^n \in \mathcal{S}^{d-1}$ as the outer unit vector of $S_{i,j}^n$ w.r.t. C_i^n. For $i \in I^n$, the index set of all neighbors of C_i^n is given as $N^n(i) = \{j \in I^n \mid \mathcal{H}_{d-1}(S_{i,j}^n) > 0\}$. Furthermore, we collect the vertices of the mesh in the set $\{\mathbf{p}_k^n \mid k \in K^n\}$.

To define a moving mesh we move vertices of the fixed mesh $\overline{\mathcal{T}}^n$. Therefore let the *shift* of a vertex $\mathbf{p}_k^n$ in $[t^n, t^{n+1}]$ be given by vectors $\mathbf{s}_k^n \in \mathbb{R}^d$ satisfying the geometric CFL condition

$$\Delta t^n |\mathbf{s}_k^n| \leq \frac{1}{2} h^n \quad \forall k \in K^n. \tag{2.1}$$

Then, by a slight abuse of notation, the *vertex motion function* is defined by

$$\mathbf{p}_k^n(t) := \mathbf{p}_k^n(t^n) + (t - t^n)\mathbf{s}_k^n. \tag{2.2}$$

Based on these notions we proceed to

Definition 1 (*Moving mesh for $n \in \mathbb{N}$*) Let the fixed mesh $\overline{\mathcal{T}}^n$ on $\mathbb{R}^d$ be given. For $i \in I^n$, let $\mathcal{L}(C_i^n)$ be the space of affine mappings from C_i^n to $\mathbb{R}^d$ and define

$$\Phi_i^n : \begin{cases} [t^n, t^{n+1}] \to \mathcal{L}(C_i^n), \\ \quad t \qquad \mapsto \Phi_i^n(t) = \Phi_i^{n,t}, \end{cases} \tag{2.3}$$

with

$$\Phi_i^{n,t}(\mathbf{x}) := \mathbf{x} + (t - t^n)\Big(\mathbf{s}_{k_{d+1}}^n + \sum_{l=1,\dots,d} \lambda_l(\mathbf{x})(\mathbf{s}_{k_l}^n - \mathbf{s}_{k_{d+1}}^n)\Big) \qquad (\mathbf{x} \in C_i^n). \tag{2.4}$$

Here $(\lambda_1(\mathbf{x}), \dots, \lambda_{d+1}(\mathbf{x}))^T$ denotes the barycentric coordinate of $\mathbf{x}$ in K_i^n, and $\mathbf{s}_{k_1}^n, \dots, \mathbf{s}_{k_{d+1}}^n$ are the shifts of the vertices $\mathbf{p}_{k_1}^n, \dots, \mathbf{p}_{k_{d+1}}^n$ of C_i^n.

We call $\mathcal{T} = \mathcal{T}(t)$ a *moving mesh in* $[t^n, t^{n+1}]$, if

$$\mathcal{T}(t) = \{\Phi_i^{n,t}(C_i^n)\}_{i \in I} \ \forall t \in [t^n, t^{n+1}]. \tag{2.5}$$

Note that we have $\mathcal{T}(t^n) = \overline{\mathcal{T}}^n$. The geometric CFL condition (2.1) guarantees that $\mathcal{T}(t)$ defines for all $t \in [t^n, t^{n+1}]$ a fixed mesh on $\mathbb{R}^d$ that has the same mesh topology as $\overline{\mathcal{T}}^n$, i.e., the same index sets $I = I^n$, $N(i) = N^n(i)$, $E = E^n$, $K = K^n$ can be used as a start. In particular we obtain a fixed mesh $\overline{\mathcal{T}}^{n+1}$ on $\mathbb{R}^d$ with volumes $C_i^{n+1} = \Phi_i^{n,t^{n+1}}(C_i^n)$. This motivates to define the (time-dependent) volumes $C_i^n(t)$ and the time-dependent facets $S_{i,j}^n(t)$ of the moving mesh $\mathcal{T} = \mathcal{T}(t)$ by

$$C_i^n(t) := \Phi_i^{n,t}(C_i^n) \text{ and } S_{i,j}^n(t) := \Phi_i^{n,t}(S_{i,j}^n) \ \forall t \in [t^n, t^{n+1}]. \tag{2.6}$$

Analogously, the vector $\mathbf{n}^n_{i,j}(t)$ denotes the outer normal of $S^n_{i,j}(t)$, see Fig. 2 for some illustration. By $\mathbf{s}^n_{i,j}$ we denote the (constant) shift vector for the center of gravity of $S^n_{i,j}(t)$.

To define a finite-volume scheme on $\mathcal{T}$ we introduce the numerical flux $\mathbf{g}^n_{i,j} = \mathbf{g}^n_{i,j}(\boldsymbol{u}, \boldsymbol{v})$ for (1.2) and the geometrical flux $\mathbf{h}^n_{i,j} = \mathbf{h}^n_{i,j}(\boldsymbol{u}, \boldsymbol{v})$ as Lipschitz-continuous functions that satisfy for all $\boldsymbol{u}, \boldsymbol{v} \in \mathcal{U}$ and all $(i, j) \in I \times N(i)$ the consistency and conservation properties

$$
\begin{aligned}
(i)\ \ & \mathbf{g}^n_{i,j}(\boldsymbol{u}, \boldsymbol{u}) = \mathbf{F}(\boldsymbol{u}) \cdot \mathbf{n}^n_{i,j}(t^{n+\frac{1}{2}}),\\
(ii)\ \ & \mathbf{h}^n_{i,j}(\boldsymbol{u}, \boldsymbol{u}) = -\left(\mathbf{n}^n_{i,j}(t^{n+\frac{1}{2}}) \cdot \mathbf{s}^n_{i,j}\right) \boldsymbol{u},\\
(iii)\ \ & \mathbf{g}^n_{i,j}(\boldsymbol{u}, \boldsymbol{v}) + \mathbf{h}^n_{i,j}(\boldsymbol{u}, \boldsymbol{v}) = -\Big(\mathbf{g}^n_{j,i}(\boldsymbol{v}, \boldsymbol{u}) + \mathbf{h}^n_{j,i}(\boldsymbol{v}, \boldsymbol{u})\Big).
\end{aligned}
$$

Note that the " $\cdot$ " in (i) has to be understood componentwise.

For cell averages $\{\boldsymbol{u}^n_i \in \mathcal{U}\}_{i \in I}$ given, the mapping

$$
\text{FVS} : \left(\{\boldsymbol{u}^n_i\}_{i \in I}, \mathcal{T}\ |_{[t^n, t^{n+1}]}\right) \mapsto \{\boldsymbol{u}^{n+1}_i\}_{i \in I}
$$

is called *finite volume step* for (1.2), if the values $\boldsymbol{u}^{n+1}_i$ are computed from

$$
\begin{aligned}
&\left|C^n_i(t^{n+1})\right| \boldsymbol{u}^{n+1}_i\\
&= \left|C^n_i(t^n)\right| \boldsymbol{u}^n_i - \Delta t^n \sum_{j \in N(i)} \left|S^n_{i,j}\left(t^{n+\frac{1}{2}}\right)\right| \Big(\mathbf{g}^n_{i,j}(\boldsymbol{u}^n_i, \boldsymbol{u}^n_j) + \mathbf{h}^n_{i,j}(\boldsymbol{u}^n_i, \boldsymbol{u}^n_j)\Big). \qquad (2.7)
\end{aligned}
$$

Remark 1 The finite volume step (2.7) involves the numerical flux $\mathbf{g}^n_{i,j}$ and the geometrical flux $\mathbf{h}^n_{i,j}$. To motivate in particular the form of the latter flux we integrate (1.2) over the space-time cell $C^n_i(t)$ (see (2.6) and Fig. 2). Application of the Reynolds transport theorem for the time derivative and the Gauß theorem for the spatial derivatives implies

$$
\begin{aligned}
&\int_{C^n_i(t^{n+1})} \boldsymbol{u}(\mathbf{x}, t^{n+1})\, d\mathbf{x}\\
&= \int_{C^n_i(t^n)} \boldsymbol{u}(\mathbf{x}, t^n)\, d\mathbf{x} - \left(\int_{t^n}^{t^{n+1}} \int_{\partial C^n_i(t)} \left(\mathbf{F}(\boldsymbol{u}(\boldsymbol{\xi}, t)) - \boldsymbol{u}(\boldsymbol{\xi}, t)\mathbf{s}^n_i(\boldsymbol{\xi}, t)^T\right) \cdot \mathbf{n}\, d\boldsymbol{\xi}\, dt\right).
\end{aligned}
$$

Here $\mathbf{s}^n_i(\cdot, t) : \partial C^n_i(t) \to \mathbb{R}^d$ denotes the shift of a point $\boldsymbol{\xi}$ from $\partial C^n_i(t)$, which is computed as time derivative of (2.4).

Using the finite volume step the MMFV scheme on a moving mesh $\mathcal{T}$ to fixed meshes $\overline{\mathcal{T}}^0, \ldots, \overline{\mathcal{T}}^{N-1}$ on the time intervals $[t^0, t^1]$,…,$[t^{N-1}, t^N]$, respectively, is summarized in the following algorithm.

Algorithm 1 (*MMFV method on a given moving mesh*)

```
Require: $\boldsymbol{u}_0, T, \mathcal{T}$
$t^0 = 0, n = 0$
$\{\boldsymbol{u}_i^0\}_{i\in I} = \left\{\frac{1}{|C_i^0(0)|}\int_{C_i^0(0)} \boldsymbol{u}_0\, d\mathbf{x}\right\}_{i\in I}$        ▷ Initial values
while $t^n < T$ do
    $\{\boldsymbol{u}_i^{n+1}\}_{i\in I} = \mathrm{FVS}(\{\boldsymbol{u}_i^n\}_{i\in I}, \mathcal{T}\,|_{[t^n,t^{n+1}]})$
    $t^{n+1} = t^n + \Delta t^n, n = n + 1$
```

Of course the method can only be stable if the wave speeds from (1.1) are captured by a CFL condition which we suppose to be satisfied. We define the approximate solution computed within Algorithm 1 as the piecewise constant function

$$\boldsymbol{u}_h(\mathbf{x}, t) = \boldsymbol{u}_i^n \qquad \text{if } t \in [t^n, t^{n+1}) \text{ and } \mathbf{x} \in C_i^n(t).$$

We note that the finite-volume method in Algorithm 1 is conservative by its construction and the evaluation of the geometrical quantities at $t^{n+\frac{1}{2}}$. The method is (formally) first-order accurate but higher-order order accuracy can be achieved by standard means using appropriate multiple evaluation of numerical/geometrical fluxes and corresponding time-stepping methods. However, the overall accuracy is limited by the approximation for the free boundary which will be discussed as the next step.

To care about the approximation of the free boundary $\Gamma = \Gamma(t)$, we combine the MMFV method from Algorithm 1 with a special tracking of the interface. The basic idea of the final method is to track a discrete interface that consists of a set of connected facets of the moving mesh. This enables us to model the flux across the interface by (special) numerical fluxes, and paves the way to solve even evolution equations on the interface.

First, we let for some $n \in \{0, \dots, N-1\}$ a discrete interface Γ_h^n be given as a connected set of facets. Denote by $\mathcal{E} \subset E$ the set of index pairs (i, j), such that $S_{i,j}^n \in \Gamma_h^n$ holds (, which implies $(i, j) \in \mathcal{E} \Rightarrow (j, i) \in \mathcal{E}$). Furthermore let $\mathcal{K} \subset K$ be the index set of all vertices $\mathbf{p}_k^n$ on Γ_h^n, i.e., $k \in \mathcal{K}$. This defines implicitly discrete bulk domains $D_{h,\pm}^n$ with corresponding volume index sets $I_\pm$.

Taking into account (1.4) the interfacial dynamics in the interval $[t^n, t^{n+1}]$ requires to determine first facet speeds $\sigma_{i,j}^n$ for $(i, j) \in \mathcal{E}$. For the interface tracking, they have to be computed using information from the smaller scales. In the framework of hyperbolic conservation laws it is likely that this information can be gained for some interface facet $S_{i,j}^n$ from a (possibly approximate) solution of a $\mathbf{n}_{i,j}^n$-rotated Riemann problem for (1.2) with initial states $\boldsymbol{u} = \boldsymbol{u}_i^n$ and $\boldsymbol{v} = \boldsymbol{u}_j^n$. Therefore, we assume an *interface solver*

$$\mathcal{R}(\mathbf{n}_{i,j}^n; \boldsymbol{u}, \boldsymbol{v}) = (R_1(\mathbf{n}_{i,j}^n; \boldsymbol{u}, \boldsymbol{v}), R_2(\mathbf{n}_{i,j}^n; \boldsymbol{u}, \boldsymbol{v}), R_3(\mathbf{n}_{i,j}^n; \boldsymbol{u}, \boldsymbol{v}))^T \tag{2.8}$$

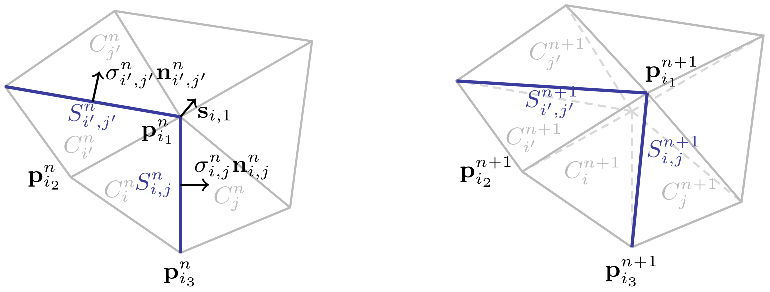


Fig. 3 Motion of the (blue) discrete interface Γ^n_h for $d = 2$: cell volumes C^n_i, C^n_j, $C^n_{i'}$, $C^n_{j'}$ and facets $S^n_{i,j}$, $S^n_{i',j'}$ of the interface Γ^n_h at time $t = t^n$ (left figure) and corresponding geometry at time $t = t^{n+1}$

to be given. Here $R_1 \in \mathbb{R}$ provides the speed of the interface in the Riemann solution and $R_2 \in \mathcal{U}$ and $R_3 \in \mathcal{U}$ stand for the respective trace states at the interface. We define then for all $(i, j) \in \mathcal{E}$ the facet speeds from

$$\sigma^n_{i,j} := R_1(\mathbf{n}^n_{i,j}; \boldsymbol{u}^n_i, \boldsymbol{u}^n_j)$$

The (up to now prescribed) shift $\mathbf{s}^n_k$ of a vertex $\mathbf{p}^n_k$ for $k \in \mathcal{K}$ in $[t^n, t^{n+1}]$ is computed from

$$\mathbf{s}^n_k := \sum_{(i,j)\in\mathcal{E},\, i\in I_-,\, \mathbf{p}^n_k\in S^n_{i,j}} \left(\sigma^n_{i,j}\,\left|S^n_{i,j}\right|\mathbf{n}^n_{i,j}\right) \Big/ \sum_{(i,j)\in\mathcal{E},\, i\in I_-,\, \mathbf{p}^n_k\in S^n_{i,j}} \left|S^n_{i,j}\right|, \tag{2.9}$$

and by $\mathbf{s}^n_k := \mathbf{0}$ for $k \in K \setminus \mathcal{K}$. Thus we obtain from (2.2) the vertex motion function $\mathbf{p}^n_k(t)$ for all $k \in K$ and a *moving mesh $\mathcal{T}$ with interface tracking* from Definition 1. Note that the discrete interface Γ^{n+1}_h is given by the union $\bigcup_{(i,j)\in\mathcal{E}} S^{n+1}_{i,j}$. A sketch of the interface motion is displayed in Fig. 3.

To define a MMFV method with interface tracking it remains to make precise the choice of the numerical and geometrical fluxes. We use our numerical and geometrical fluxes for $(i, j) \in E \setminus \mathcal{E}$ as before in the finite-volume step (2.7) but insert the Godunov flux at interface facets with index pairs $(i, j) \in \mathcal{E}$, i.e.,

$$\begin{aligned} \mathbf{g}^n_{i,j}(\boldsymbol{u}, \boldsymbol{v}) &= \boldsymbol{F}(R_2(\mathbf{n}^n_{i,j}; \boldsymbol{u}, \boldsymbol{v})) \cdot \mathbf{n}^n_{i,j}(t^{n+\frac{1}{2}}), \\ \mathbf{h}^n_{i,j}(\boldsymbol{u}, \boldsymbol{v}) &= -\sigma^n_{i,j}(\boldsymbol{u}, \boldsymbol{v}) R_2(\mathbf{n}^n_{i,j}; \boldsymbol{u}, \boldsymbol{v}). \end{aligned} \tag{2.10}$$

It is the choice (2.10) that integrates smaller-scale information via the adjacent state $R_2(\mathbf{n}^n_{i,j}; \boldsymbol{u}, \boldsymbol{v})$ computed by the interface solver $\mathcal{R}$ in (2.8). The interface solver $\mathcal{R}$ is evaluated only once for each facet and the state $R_3(\mathbf{n}^n_{i,j}; \boldsymbol{u}, \boldsymbol{v})$ is then used to calculate the flux associated with the geometrically identically facet $S^n_{j,i}$.

Now all tools for the tracking algorithm are given, which is summarized in

Algorithm 2 (*MMFV method with interface tracking*)

```
Require: $\boldsymbol{u}_0, T, \overline{\mathcal{T}}^0$ with $\Gamma_h^0$
$t^0 = 0, n = 0$
$\{\boldsymbol{u}_i^0\}_{i\in I} = \left\{\frac{1}{|C_i^0(0)|}\int_{C_i^0(0)} \boldsymbol{u}_0 \, d\mathbf{x}\right\}_{i\in I}$    ▷ Initial values
while $t^n < T$ do
    Compute $\mathcal{R}$ for all $(i, j) \in \mathcal{E}$.    ▷ Compute interface motion
    Compute moving mesh with interface tracking $\mathcal{T}\,|_{[t^n,t^{n+1}]}$ in $[t^n, t^{n+1}]$.
    $\Gamma_h^{n+1} = \bigcup_{(i,j)\in\mathcal{E}} S_{i,j}^{n+1}$
    $\{\boldsymbol{u}_i^{n+1}\}_{i\in I} = \mathrm{FVS}(\{\boldsymbol{u}_i^n\}_{i\in I}, \mathcal{T}\,|_{[t^n,t^{n+1}]})$
    $t^{n+1} = t^n + \Delta t^n, n = n + 1$
```

Algorithm 2 provides an approximation for the free-boundary value problem (1.2), (1.4). In [9] we have proven that Algorithm 2 can keep planar traveling-wave solutions exactly, preserves mass and satisfies (using entropy-dissipative numerical fluxes in the bulk regions) a discrete entropy inequality. A full convergence analysis for $d = 1$ can be found in [8].

However, Algorithm 2 will not terminate if the geometrical CFL condition (2.1) fails. In this case the point motion functions leads to a vertex distribution at t^{n+1} that cannot be meshed with the given (time-invariant) mesh topology. Let us assume that the geometrical CFL condition is always satisfied (like we assume it for the hyperbolic CFL conditions). Both can be achieved by reducing the time step. But even if mesh topology can be kept the sole movement of points from the discrete interface might lead to very small volumes in the bulk domains and (in view of curvature changes) very small or very big distances of interface vertices. As a remedy an additional post-processing of the mesh is necessary for any time step. The *re-meshing operator* $\mathrm{RE} = \mathrm{RE}(\mathcal{T}(t^{n+1}))$ is supposed to construct the fixed mesh $\overline{\mathcal{T}}^{n+1}$ such that

- the approximate interface Γ_h^n is preserved in $\overline{\mathcal{T}}^{n+1}$ (up to a given tolerance),
- the mesh parameter h^{n+1} of $\overline{\mathcal{T}}^{n+1}$ is above a critical threshold,
- and the distance of interface vertices of $\overline{\mathcal{T}}^{n+1}$ remains in a prescribed interval.

To achieve these criteria based on given mesh parameters, additional vertices might be placed or deleted on Γ_h^n and in a small neighborhood around the interface. The resulting vertices set will then be meshed. We will not define this re-meshing operator RE here but note that moving-mesh concepts that realize the constraints mentioned above in $d \in \{1, 2, 3\}$ space dimensions have been introduced in [2, 5]. They ensure that the fixed meshes at any time step preserve an initially given Delaunay property. These are essentially the new contributions for our final MMFV method.

Algorithm 3 (*MMFV method with interface tracking and re-meshing*)

```
Require: $\boldsymbol{u}_0, T, \overline{\mathcal{T}}^0$ with $\Gamma_h^0$, mesh parameters
1: $t^0 = 0, n = 0$
2: $\{\boldsymbol{u}_i^0\}_{i \in I} = \left\{ \frac{1}{|C_i^0(0)|} \int_{C_i^0(0)} \boldsymbol{u}_0 \, d\mathbf{x} \right\}_{i \in I^0}$    ▷ Initial values
3: while $t^n < T$ do
4:     Compute $\mathcal{R}$ for all $(i, j) \in \mathcal{E}^n$.    ▷ Compute interface motion
5:     Compute moving mesh with interface tracking $\mathcal{T}\ |_{[t^n, t^{n+1}]}$ in $[t^n, t^{n+1}]$.
6:     $\{\hat{\boldsymbol{u}}_i^{n+1}\}_{i \in I^n} = \mathrm{FVS}(\{\boldsymbol{u}_i^n\}_{i \in I^n}, \mathcal{T}\ |_{[t^n, t^{n+1}]})$
7:     $\overline{\mathcal{T}}^{n+1} \leftarrow \mathrm{RM}(\mathcal{T}(t^{n+1}))$    ▷ Re-meshing step
8:     Determine $\Gamma_h^{n+1}$ from $\Gamma_h^n(t^{n+1})$ and $\overline{\mathcal{T}}^{n+1}$.
9:     Projection $\{\boldsymbol{u}_i^{n+1}\}_{i \in I^{n+1}} \leftarrow \{\hat{\boldsymbol{u}}_i^{n+1}\}_{i \in I^n}$
10:    $t^{n+1} = t^n + \Delta t^n, n = n + 1$
```

Within Algorithm 3, step 7, the topology of the mesh can change due to the re-meshing process. This is different as in Algorithms 1, 2 and implies that all index sets can change and must be augmented with the upper index n. For the same reason a projection step is needed for the approximate finite-volume solution $\boldsymbol{u}_h$ and the determinations of the approximate free boundary Γ_h^n at $t = t^n$.

In the sequel we provide results obtained with Algorithm 3 for two different applications.

3 Phase-Boundary Dynamics in Compressible Liquid-Vapour Flow

As first instance of a free boundary value problem for the conservation law system (1.2) we consider the Euler equations for compressible flow given by

$$\begin{array}{rcl} \varrho_t & + & \nabla \cdot (\varrho \boldsymbol{v}) = 0, \\ (\varrho \boldsymbol{v})_t & + & \nabla \cdot (\varrho \boldsymbol{v} \otimes \boldsymbol{v}) + \nabla p = \mathbf{0}, \\ E_t & + & \nabla \cdot ((E + p)\boldsymbol{v}) = 0 \end{array} \quad \text{in } D_\pm(t) \text{ for } t \in (0, T). \tag{3.1}$$

The unknowns of the system (3.1) are the fluid density $\varrho = \varrho(\mathbf{x}, t) > 0$, the fluid momentum density $\varrho \boldsymbol{v} = (\varrho \boldsymbol{v})(\mathbf{x}, t) \in \mathbb{R}^d$, the total energy density $E = E(\mathbf{x}, t) \in \mathbb{R}$, and the interface $\Gamma(t)$. The total energy density satisfies $E = \varrho \varepsilon + \frac{1}{2} \varrho |\boldsymbol{v}|^2$, with ε denoting the specific internal energy.

To close the system, the pressure p, the specific internal energy ε, and the temperature T are connected by equations of state (EOS). We express the pressure by $p(\varrho, T) = \varrho^2 \psi_\varrho(\varrho, T)$, with ψ being the specific Helmholtz free energy. The specific internal energy results from $\varepsilon(\varrho, T) = \psi(\varrho, T) - T \psi_T(\varrho, T)$ and the entropy S from $S = S(\varrho, T) = -\partial_T \psi(\varrho, T)$. The system (3.1) is hyperbolic [10], if the EOS satisfy

$$p_\varrho(\varrho, S) > 0, \quad T(\varrho, S) > 0. \tag{3.2}$$

Here, we choose the Helmholtz free energy from [15] that is consistent with a Lennard-Jones potential on the molecular scale (see below), and governs in particular noble gases. Using the admissibility criterion (3.2) we deduce that (for small enough temperatures) the state space splits in two separate regions in $(0, \infty) \times \mathbb{R}^{d+1}$, that are identified with a vapour (+) and a liquid (–) fluid phase. The domains $D_\pm = D_\pm(t)$ are then the subsets of D occupied by the ±-phase. The unknown interface $\Gamma(t)$ forms the joint boundary of these two sets.

Thus (3.1) fits to the general class of free boundary value problems with hyperbolic dynamics from Sect. 1. It remains to fix the transition operator $\mathcal{B}$ from (1.4) to determine the motion of the interface $\Gamma(t)$. It is well known that the Rankine-Hugoniot conditions (1.5) do not suffice to ensure well-posedness of the free-boundary value problem but have to be augmented by a further condition (see for [23] for a review in the isothermal case). Various suggestions of additional algebraic relations have been made but up to now there is no choice that would ensure the unique solution of the Riemann problem in the entire state space.

Therefore there is no Riemann-problem based interface solver $\mathcal{R}$ as required in (2.8) for the MMFV method. As a remedy we have suggested in [19] a molecular-dynamics based interface solver $\mathcal{R}_{\mathrm{md}}$. $\mathcal{R}_{\mathrm{md}}$ determines for the continuum mechanical states $\boldsymbol{u}, \boldsymbol{v}$ in (2.8) a corresponding initial particle state in a special linked-cell approach [3]; using a prescribed number of particles. Then, molecular-dynamics simulations are run and (ensemble-averaged) results provide an estimate for the facet speed and the trace states using the Irving-Kirkwood formulas. In this way we obtained the globally applicable interface solver $\mathcal{R}_{\mathrm{md}}$. We conclude the section with a simulation of the MMFV method from [19].

Example 1 (*Condensating droplet in 2D*) We consider the domain $D = (-1.5, 1.5)^2$. Initially it is separated in the liquid droplet domain $D_{0,-}(0) = \{\mathbf{x} = (x_1, x_2) \in \mathbb{R}^2 \mid |\mathbf{x}|_2^2 < 0.15\}$ and the vapour complement $D_{0,+}$ in D. The initial data in primitive variables are chosen as

$$(\varrho, (v_1, v_2), T)(\mathbf{x}, t) = \begin{cases} (0.62, (0, 0), 0.8) & \mathbf{x} \in D_{0,-}, \\ (0.06, (0, 0), 0.9) & \mathbf{x} \in D_{0,+} \text{ and } x_1 \geq -0.5, \\ (0.06, (0.5, 0), 0.9) & \text{otherwise.} \end{cases}$$

At the left boundary, at $x_1 = -1.5$, we prescribe inflow boundary conditions by setting ghost cell values corresponding to $(\varrho, (v_1, v_2), T) = (0.06, (0.5, 0.0), 0.9)$. Otherwise, we use outflow boundary conditions. The setting induces a vapour wave traveling to the right and colliding with the droplet. The evolution of the density and velocity fields using the MMFV method with the Lax-Friedrichs flux in the bulk domains is displayed in Fig. 4. Since the droplet is initially not in equilibrium a relaxing condensation process is started. When the vapour wave reaches the droplet it gets deformed and convected through the domain.

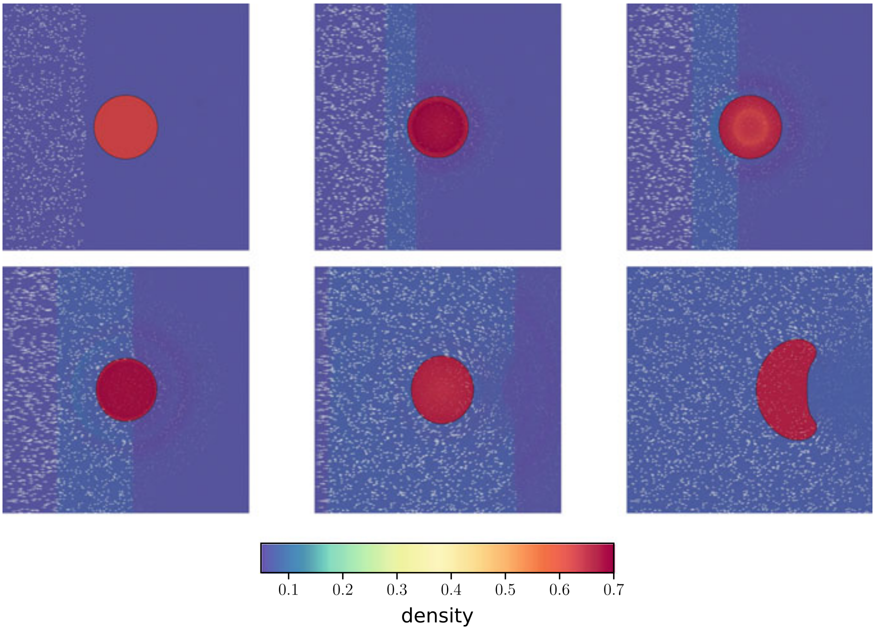


Fig. 4 Density (color) and velocity (arrows) for $t = 0.0, 0.2, 0.3, 0.5, 1.25, 5.0$ (upper left to lower right). The pictures display the impingement of a density wave on an initially static droplet with lower temperature. The results are taken from [19]

4 Fracture Propagation in Porous Media

Mixed-dimensional modelling for flow in porous media is frequently used when dealing with extreme geometries like fractures. For fractures with e.g. disc-like structures in 3D, the originally fully dimensional fractures are reduced to co-dimension-1 structures dissecting the surrounding fully dimensional bulk region. This discrete fracture network approach has been pioneered for Darcy flow in [20] and has been since then extended to various fields. In this section we refer mainly to the works [6, 7] on two-phase flows in fracturing porous media which assume the fractures to be filled with debris such that they can be also considered as a porous medium.

To introduce the model we denote for $t \in [0, T]$ by $\Gamma = \Gamma(t)$ a hypersurface (i.e., a fracture as free boundary), that forms together with the surrounding bulk domain $D = D(t)$ the entire domain $D \subset \mathbb{R}^d$. The function $\omega = \omega(\boldsymbol{\xi}, t) > 0$ is the corresponding fracture aperture function in $\boldsymbol{\xi} \in \Gamma(t)$, see Fig. 5 for a sketch of the geometrical setting.

The bulk domain $D(t)$ and the fracture $\Gamma(t)$ are filled with two immiscible and incompressible fluids. The flow dynamics is then described in $D(t)$ by the wetting fluid saturation $S = S(\cdot, t) : D(t) \to [0, 1]$, the global pressure $P = P(\cdot, t)$:

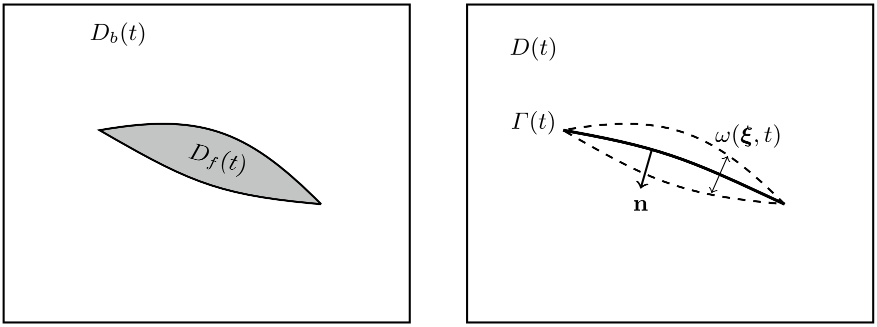


Fig. 5 Dimensional reduction of fully dimensional fracture domain $D_f(t)$ to a hypersurface $\Gamma(t)$ as fracture in the domain $D(t)$

$D(t) \to \mathbb{R}$, and the total velocity $\boldsymbol{v} = \boldsymbol{v}(\cdot, t) : D(t) \to \mathbb{R}^d$. If we neglect gravitational forces, capillarity effects and set the porosities equal to 1, the unknowns obey

$$\left.\begin{aligned} S_t + \nabla \cdot (f(S)\boldsymbol{v}) &= q_w, \\ \boldsymbol{v} + \lambda(S)\mathbf{K}\nabla P &= \mathbf{0}, \\ \nabla \cdot \boldsymbol{v} &= q_w + q_{nw} \end{aligned}\right\} \quad \text{in } D(t). \tag{4.1}$$

Here, the function $f = f(S)$ is the (nonlinear) fractional flux function, $\lambda = \lambda(S)$ is the mobility, $\mathbf{K} = \mathbf{K}(\mathbf{x}, t)$ is the positive-definite bulk permeability matrix, and $q_{n/nw} = q_{n/nw}(\mathbf{x}, t)$ indicate fluid source terms. On the interface $\Gamma(t)$ we search for the reduced quantities $S_\Gamma = S_\Gamma(\cdot, t) : \Gamma(t) \to [0, 1]$, $P_\Gamma(\cdot, t) : \Gamma(t) \to \mathbb{R}$ and $\boldsymbol{v}_\Gamma(\cdot, t) : \Gamma(t) \to \mathbb{R}^{d-1}$ that satisfy

$$\left.\begin{aligned} (\omega S_\Gamma)_t + \nabla_{\boldsymbol{\xi}} \cdot (\omega f(S_\Gamma)\boldsymbol{v}_\Gamma) &= [\![f(S)\boldsymbol{v} \cdot \mathbf{n}]\!] + \omega q_w^\Gamma, \\ \omega \boldsymbol{v}_\Gamma + \lambda(S_\Gamma)\mathbf{K}^\Gamma \nabla_{\boldsymbol{\xi}}(\omega P_\Gamma) &= \mathbf{0}, \\ \nabla_{\boldsymbol{\xi}} \cdot (\omega \boldsymbol{v}_\Gamma) &= [\![\boldsymbol{v} \cdot \mathbf{n}]\!] + \omega(q_w^\Gamma + q_{nw}^\Gamma) \end{aligned}\right\} \quad \text{in } \Gamma(t). \tag{4.2}$$

The notations with upper Γ-index refer to the same physical but possibly dimensionally reduced quantities as in (4.2). The differential operators with lower index $\boldsymbol{\xi}$ are the differential operators on the fracture manifold. The systems (4.1), (4.2) are closed at the hypersurface $\Gamma(t)$ by the mass conservation and momentum balance relations

$$\begin{aligned} f(S_\pm) &= f^\Gamma(S_\Gamma), \\ \boldsymbol{v}_\pm \cdot \mathbf{n}_\pm &= -\lambda(S_\Gamma)K_{\mathbf{n}}^\Gamma \left(\frac{P_\Gamma - P_\pm}{\omega/2} + \frac{P_\Gamma - P_+ - P_-}{\omega/4} \right). \end{aligned} \tag{4.3}$$

The number $K_{\mathbf{n}}^\Gamma > 0$ is the projection of $\mathbf{K}$ in the normal direction. By fixing a normal vector $\mathbf{n} \in \mathcal{S}^{d-1}$ we can naturally determine domains $D_\pm(t)$ around $\Gamma(t)$ (see Fig.

5). This defines then the trace evaluations $S_{\pm}$, $\boldsymbol{v}_{\pm}$ as well as the brackets $[\![\cdot]\!]$, like in Sect. 3. The entire discrete fracture network model (4.1)–(4.3) has to be augmented by initial conditions and boundary conditions for the saturations, pressures, and total velocities in bulk and fracture domains. The initial fracture hypersurface Γ_0 has to be prescribed.

The fracture interface $\Gamma(t)$ moves at the tips in tangential direction (fracture growth) and not in normal direction as the phase boundary in Sect. 3. Since we exclude closing of the fracture by assuming $\omega(\cdot, t) > 0$ for all $t \in [0, T]$, we have $\Gamma(t_1) \subseteq \Gamma(t_2)$ for $t_1 < t_2$, meaning that $\Gamma(t_1)$ remains static for $t > t_1$. To ensure in particular mass conservation coupling to the two-phase dynamics in the fracture is then essential, see $(4.3)_1$. The new feature of the mathematical model (4.1)–(4.3) concerning the evolution of the fracture $\Gamma = \Gamma(t)$ is its growth by dynamical fracturing. Possible driving forces for fracturing are the hydromechanical pressure or a suitable stress distribution in the solid skeleton. The resulting growth of the tip has to be computed from a smaller-scale model that relies on energetic considerations like the Griffith criterion. In the sequel we assume that the motion of the fracture is given and refer to [6] for the coupling with a smaller-scale fracture model.

Remark 2 (i) For $S \equiv S_\Gamma \equiv 1$, the model (4.1)–(4.3) reduces to the single-phase case as described in [20]. For $\omega = const.$, we reconstruct the models as in e.g. [1, 14], but obtain different coupling conditions. For unsaturated flow we refer also to the more recent work [18].

(ii) The systems (4.1) in $D(t)$ and (4.2) in $\Gamma(t)$ are of mixed hyperbolic-elliptic type. However, the dynamics is driven by the nonlinear hyperbolic equation for the saturation such that it fits to the general conservation law set-up in Sect. 1.

A mixed-dimensional MMFV method for (4.1)–(4.3) has been introduced in [7] which relies on the moving-mesh approach from [5]. The latter realizes in particular the growth of $\Gamma(t)$ at its tip(s) by adding new vertices and re-meshing such that the moving mesh is a conforming triangulation in every discrete time step. Nevertheless, an interface solver $\mathcal{R}$ in the sense of (2.8) is also needed. $(4.1)_1$ is a scalar conservation law that depends on the total velocity as parameter. Thus, the solution of the Riemann problem produces two waves. The wave of interest for $\mathcal{R}$ is associated with the fracture. Therefore the speed of the wave of interest is zero but the computation of the adjacent states, which are not equal to the end states of the Riemann problem, involves the solution of (4.2) for the dynamics within the fracture. We point out that the geometrical flux $\mathbf{h}_{ij}^n$ vanishes since all facets S_{ij}^n of $\Gamma(t^n)$ that do not include the tip(s) do not move within $[t^n, t^{n+1}]$. The numerical flux g_{ij}^n is chosen as the exact flux, evaluated in the adjacent states, see (2.10). The (prescribed) movement of the tip (a vertex of the mesh) is represented as the point motion function which induces a movement of the mesh locally around the tip vertex. The fluxes can be taken again as in (2.10) evaluating the bulk saturations in the (Godunov) flux.

We conclude with an illustration of the resulting MMFV method.

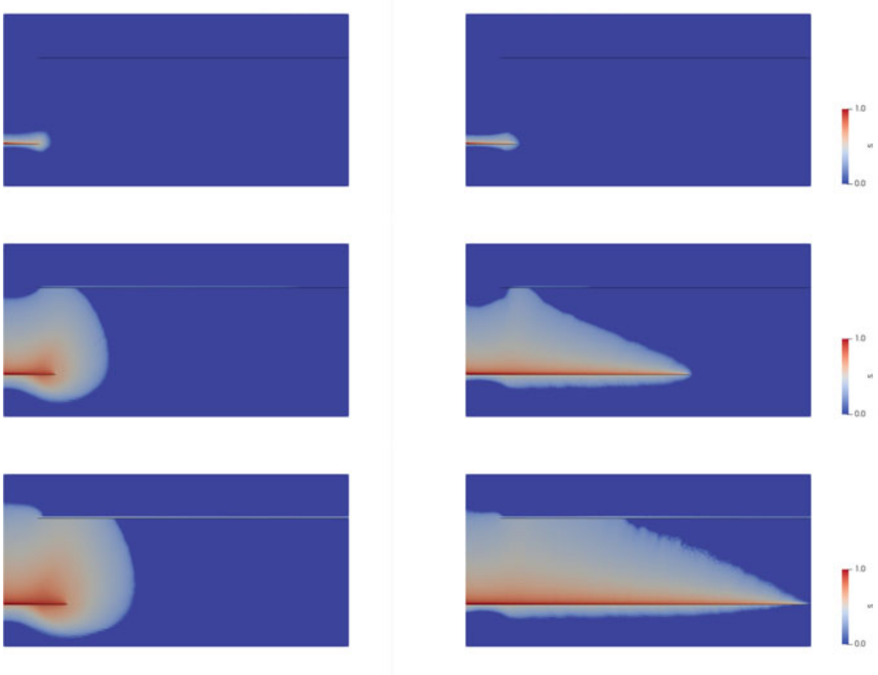


Fig. 6 Saturation fields for a domain with two horizontal fractures at $t = 0.1, 1.1, 1.5$. The left hand side pictures correspond to a tip speed $s = 0.1$ of the lower fracture $\Gamma_2(t)$ and the right hand side pictures to the speed $s = 1.0$. The results are taken from [7]

Example 2 (*Two-phase flow fronts in fracturing domain*) Let $D = (0, 2) \times (0, 1)$. We consider two horizontally aligned fractures $\Gamma_{1/2}(t)$ with constant aperture $\omega = 0.01$ such that the upper fracture is defined by $\Gamma_1 = (0.2, 2) \times \{0.75\}$ and is kept static. The tip of the lower fracture $\Gamma_2 = \Gamma_2(t)$ moves with speed $s \in \{0.1, 1.0\}$ from left to right reaching the right boundary at $t = 1.8, 18$, respectively. The initial saturation in D is zero and no-flux boundary conditions are prescribed on $\partial D \setminus (\Gamma_1 \cap \Gamma_2(t))$. The lower fracture acts as inlet for the wetting fluid whereas for the upper fracture outlet conditions are prescribed.

The numerical solutions at different time steps can be observed from Fig. 6. Discontinuous saturation fronts travel through the fracture Γ_2 and invade into the bulk domain. From $t = 1.5$ on the front has reached Γ_1 which acts as a barrier for further vertical expansion transporting the wetting-phase to the right boundary.

Acknowledgements Funded by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC 2075 - 390740016. We acknowledge the support by the Stuttgart Center for Simulation Science (SimTech).

References

1. Ahmed, E., Jaffré, J., Roberts, J.E.: A reduced fracture model for two-phase flow with different rock types. Math. Comput. Simul. **137**, 49–70 (2017)
2. Alkämper, M., Magiera, J., Rohde, C.: An interface preserving moving mesh in multiple space dimensions (2021). https://arxiv.org/abs/2112.11956
3. Allen, M., Tildesley, D.: Computer Simulation of Liquids, 2nd edn. Oxford University Press Inc, Oxford (2017)
4. Barth, T., Herbin, R., Ohlberger, M.: Finite volume methods: foundation and analysis. In: Stein, E., de Borst, R., Hughes, T. (eds.), Encyclopedia of Computational Mechanics, chapter 5, pp. 1–60. Wiley (2018)
5. Burbulla, S., Dedner, A., Hörl, M., Rohde, C.: Dune-mmesh: The DUNE grid module for moving interfaces. J. Open Source Softw. **7**(74), 3959 (2022)
6. Burbulla, S., Formaggia, L., Rohde, C., Scotti, A.: Modeling fracture propagation in poro-elastic media combining phase-field and discrete fracture models. Comput. Methods Appl. Mech. Eng. **403** (2023)
7. Burbulla, S., Rohde, C.: A finite-volume moving-mesh method for two-phase flow in fracturing porous media. J. Comput. Phys. **458**, paper no. 111031 (2022)
8. Chalons, C., Engel, P., Rohde, C.: A conservative and convergent scheme for undercompressive shock waves. SIAM J. Numer. Anal. **52**(1), 554–579 (2014)
9. Chalons, C., Rohde, C., Wiebe, M.: A finite volume method for undercompressive shock waves in two space dimensions. ESAIM Math. Model. Numer. Anal. **51**(5), 1987–2015 (2017)
10. Dafermos, C.: Hyperbolic Conservation Laws in Continuum Physics, 3rd edn. Springer, Cham (2010)
11. Dai, M., Schmidt, D.P.: Adaptive tetrahedral meshing in free-surface flow. J. Comput. Phys. **208**(1), 228–252 (2005)
12. Falcovitz, J., Alfandary, G., Hanoch, G.: A two-dimensional conservation laws scheme for compressible flows with moving boundaries. J. Comput. Phys. **138**(1), 83–102 (1997)
13. Feireisl, E., Lukáčová-Medvid'ová, M., Mizerová, H., She, B.: Numerical Analysis of Compressible Fluid Flows. Springer, Cham (2021)
14. 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)
15. Gross, J., Sadowski, G.: Perturbed-Chain SAFT: An equation of state based on a perturbation theory for chain molecules. Ind. Eng. Chem. Res. **40**(4), 1244–1260 (2001)
16. Harten, A., Hyman, J.M.: Self-adjusting grid methods for one-dimensional hyperbolic conservation laws. J. Comput. Phys. **50**(2), 235–269 (1983)
17. Huang, W., Russell, R.D.: Adaptive Moving Mesh Methods. Springer, New York (2011)
18. List, F., Kumar, K., Pop, I.S., Radu, F.A.: Rigorous upscaling of unsaturated flow in fractured porous media. SIAM J. Math. Anal. **52**(1), 239–276 (2020)
19. Magiera, J., Rohde, C.: A molecular-continuum multiscale model for inviscid liquid-vapor flow with sharp interfaces. J. Comput. Phys. **469**, paper no. 111551 (2022)
20. 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)
21. Perot, B., Nallapati, R.: A moving unstructured staggered mesh method for the simulation of incompressible free-surface flows. J. Comput. Phys. **184**(1), 192–214 (2003)
22. Quan, S., Schmidt, D.P.: A moving mesh interface tracking method for 3d incompressible two-phase flows. J. Comput. Phys. **221**(2), 761–780 (2007)
23. Rohde, C.: Fully resolved compressible two-phase flow: modelling, analytical and numerical issues. In: Bulíček, M., Feireisl, E., Pokorný, M. (eds.), New Trends and Results in Mathematical Description of Fluid Flows, Nečas Center Series, chapter 4, pp. 115–181. Birkhäuser (2018)

24. Tang, H., Tang, T.: Adaptive mesh methods for one- and two-dimensional hyperbolic conservation laws. SIAM J. Numer. Anal. **41**, 487–515 (2003)
25. Tukovic, Z., Jasak, H.: Simulation of free-rising bubble with soluble surfactant using moving mesh finite volume/area method. In: Proceedings of 6th International Conference on CFD in Oil and Gas, Metallurgical and Process Industries (2008)

Mixed Dimensional Modeling with Overlapping Continua on Cartesian Grids for Complex Applications

Malgorzata Peszynska, Tyler Fara, Madison Phelps, and Nachuan Zhang

Abstract We outline a strategy for prototyping computational models for complex applications on domains with disparate volumes. Instead of mixed dimensional approaches, we exploit the idea of overlapping continua and Cartesian grids. This strategy allows to build prototypes and explore model sensitivities in a robust simulation framework. As specific examples we consider thermal conduction in human tissue and hypothermia, Richards-Darcy coupled system for soil-root flow, and mixed use traffic simulation on and off paved paths in urban environment.

Keywords Cartesian grids · Mixed dimensional modeling · Overlapping continua

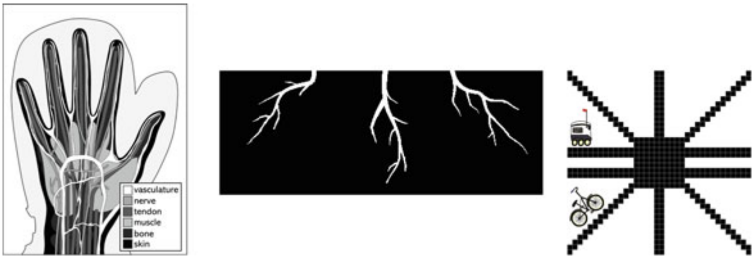
1 Introduction

In this paper we outline a strategy for rapid construction of prototype computational schemes for complex physical problems on domains with disparate volume. Instead of body fitting and unstructured grids we use overlapping continua, immersed boundary and fictitious domain techniques [13, 18] within an algorithmic framework based on Cartesian grids and semi-implicit iterative schemes known from multiphysics applications and frameworks [1, 3]. This approach is robust and simple and allows to focus on application-specific challenges.

We consider systems which involve some form of mixed-size domain and couplings for models involving some discretizations of PDEs. Recently, there is abundance of theoretical and applications oriented work on mixed dimensional setting for a variety of applications: blood flow, hydro–mechanical coupling in fracture networks, and more. The theoretical basis: detailed delicate functional analytic setting of the PDEs and of the underlying finite element approximations involves distributed

M. Peszynska (✉) · T. Fara · M. Phelps · N. Zhang
Oregon State University, Corvallis, OR 97331, USA
e-mail: mpesz@math.oregonstate.edu
URL: http://www.math.oregonstate.edu/ mpesz/

E. Franck et al. (eds.), *Finite Volumes for Complex Applications X—Volume 1, Elliptic and Parabolic Problems*, Springer Proceedings in Mathematics & Statistics 432,
https://doi.org/10.1007/978-3-031-40864-9_8



(a) (b) (c)

Fig. 1 Cartoons of three domains considered in this paper for the applications to **a** hypothermia modeling [2], **b** vegetation-soil systems [22] **c** mixed-use traffic network along some paths (in black) and off-path (white)

line sources or hybrid coupled continua–network models; see, e.g., [12, 28] and earlier applications oriented work including [7, 9, 11].

Our approach is to start from images to define the geometry of the computational domain $\Omega \subset \mathbb{R}^d$ which we project on a Cartesian grid. The lower dimensional domains are the unions of some grid/voxel cells embedded within the original Ω; see Fig. 1. Next we apply the concept of overlapping continua [6, 27] or the homogenized models [10] to handle the mixed dimensional coupling. For discretization, we use the well known practical computational framework of lowest order mixed finite element methods implemented as cell centered finite differences (CCFD), and we develop model approximations and enhancements suitable for the selected phenomena.

This strategy allows to build prototype models, propose their extensions and enhancements, and investigate their robustness, flexibility, and sensitivity with case studies which guide how to extend, modify, and interpret the models, while paying attention to numerical discretization (h, τ), and implementation (solver). This paper illustrates the simplicity and potential of this methodology which we recommend as a stepping stone towards refinement strategies beyond the applications we discuss.

We select three disparate applications dubbed H, RS, and T which involve coupled variables defined on model- and variable-specific domains which may overlap, with mixed dimensional setting; see Fig. 1. The computational models solved on these domains are discretizations of PDEs which communicate on the intersection of the domains, or are "aware" of the variables defined on these other domains. The models H, RS and T are on modeling hypothermia problems in tissue, water uptake in large root-soil systems, and mixed-use traffic flow on a campus domain. From solver point of view, case H is semilinear, and case RS and T require care in the treatment of advective terms.

The outline of this paper is as follows. In Sect. 2 we recall the basic algorithm for a scalar parabolic–hyperbolic PDE implemented as CCFD. In Sect. 3 we introduce the H (hypothermia) model, an extension of the linear bioheat flow model [10, 15, 19] for which we exploit fictitious domain approach as well as introduce the nonlinear feature of vasoconstriction. In Sect. 4 we employ an extension of $d = 1$ Richards-

Darcy models for RS (root-soil) systems from [4, 26] to mixed dimensional setting in $d \geq 2$. In Sect. 5 we develop model T for mixed-species traffic flow on a campus network involving several species whose trajectories may or may not be confined to pathways, and which may or may not be aware of other traffic participants.

While due to the scope we cannot cover all the details, the models H, RS, and T can help in further construction of more refined models, also for other complex applications. Such refined models can take advantage of sensitivity studies carried out with the prototype models.

2 Notation and Flow Models

In the paper we apply the following common notation. First, χ_S denotes the characteristic function of a set S, $|S|$ its measure, and $\overline{S}$ its closure. For a function f on S, $\langle f \rangle_S$ denotes its average on S. We work in spatial domains which are open bounded sets $\Omega \subset \mathbb{R}^d, d = 1, 2, 3$, possibly partitioned into some material subdomains, open sets Ω_m. The case $d \leq 2$ is natural for Sect. 5, and the problems in Sects. 3 and 4 work with $d \leq 3$. For the boundary $\partial\Omega$ or that of any of the subdomains $\partial\Omega_m$, η is a unit normal vector pointing outward, and we assume that these boundaries are sufficiently smooth. We will consider Dirichlet conditions on $\partial\Omega_D$ and flux conditions on $\partial\Omega_N$.

We will denote by $0 \leq t \leq T$ the time variable, with T the final time of the simulation. In time–discrete models, t^n is the time step, and $t^0 = 0$ denotes the initial time. With uniform time stepping we have $t^n = n\tau$.

2.1 CCFD as the P0-RT$_{[0]}$ Mixed Finite Element Scheme

We consider first a generic scalar parabolic quasi-linear homogeneous PDE

$$c\partial_t u - \nabla \cdot (D\nabla u) + \nabla \cdot (f(u)) + \mathcal{C}(u) = 0, \;\; x \in \Omega, t > 0 \tag{1}$$

which is supplemented by some boundary and initial conditions. Assume D is a symmetric uniformly positive definite tensor, c is uniformly nonnegative, f is sufficiently smooth, and $\mathcal{C}(u) = c_b(x, u, t)u$ is the Helmholtz term. (See (H) model below). For nonlinear implicit parabolic problem, $D = D(u)$, and $c\partial_t u$ derives from some $\partial_t C(u)$ (See (RS) model below). When $c = 0, \mathcal{C} = 0$, (1) is a steady flow problem, an elliptic PDE (See (T) model below).

For spatial discretization, we assume that $\Omega = \Omega_h = \bigcup_{(ij)\in\mathcal{T}_h} \omega_{ij}$ on some underlying background Cartesian grid of rectangular cells ω_{ij} of center x_{ij} whose edges $\mathcal{E}_h$ align with $\partial\Omega$ and with the material interfaces, with $\max_{ij} |\omega_{ij}| = h^2$. The cells are identified with voxels (image pixels); some $\omega_{ij} \notin \Omega$ are "key-outs" outside the union $\mathcal{T}_h$ of indices as in Fig. 1. The unknown u is approximated by piecewise constants

$u_{ij} \approx u(x_{ij})$; we collect $u_h = (u_{ij})_{(ij)\in\mathcal{T}_h}$. The flux components q_h defined over the edges $\mathcal{E}_h$ are from the mixed finite element $RT_{[0]}$ space. This discretization is well known [20] for its locally conservative properties and easiness for modeling nonlinear multi-physics applications. Since Dirichlet conditions are satisfied weakly [20], fictitious domain strategies [13] are their natural extension.

Applying implicit-explicit time stepping, we seek u_h^n which solves

$$\mathcal{M}(u_h^n) = (M_h + \tau_n A_h)u_h^n + \tau\mathcal{C}(u_h^n) + \tau\nabla_h \cdot F_h(u_h^{n-1}) = Mu_h^{n-1}. \qquad (2)$$

Here M_h is the mass matrix, and A_h is the stiffness matrix incorporating the boundary conditions; $\mathcal{C}$ applies pointwise to each degree of freedom of u_h^n. Further, F_h is the numerical flux defined based on f, and $\nabla_h\cdot$ is the discrete counterpart of $\nabla\cdot$ on the grid $\mathcal{T}_h$. For scalar problems we use Godunov flux; when u is a vector, we require a Riemann solver [8, 17]. For nontrivial F_h, we apply operator splitting: we solve the advection portion $c\partial_t u + \nabla \cdot f(u) = 0$ explicitly, followed by an implicit diffusion-reaction step solved directly or by iteration, time-lagging the nonlinearity and/or the coupling. After u_h^n is found, we retrieve the fluxes q_h^n. In what follows we drop the reference to h, n.

2.2 Coupled Overlapping Continua Models

We consider next a system based on (1)

$$c_1\partial_t u_1 - \nabla \cdot (D_1\nabla u_1) + c(u_1 - u_2) = f_1,\ \ x \in \Omega_1, t > 0 \qquad (3a)$$

$$c_2\partial_t u_2 - \nabla \cdot (D_2\nabla u_2) + c(u_2 - u_1) = f_2,\ \ x \in \Omega_2, t > 0 \qquad (3b)$$

The model (3a, 3b) when $\Omega_1 = \Omega_2$ is known as the Barenblatt model for multiscale flow in porous media [6, 27] postulated for the flow dynamics in composite multiscale media with vastly different storage c_j, D_j. See also [16] for illustrations.

In this paper Ω_1 is a continuum, and Ω_2 is “almost” one-dimensional, $\Omega_2 \subset \Omega_1$ with $0 < |\Omega_2| << |\Omega_1|$ in the same $\mathbb{R}^d$ measure, and $c = c(x) = C\chi_{\Omega_1\cap\Omega_2}(x)$, $C \geq 0$, only active on $\Omega_1 \cap \Omega_2$. Our scheme for (3a, 3b) extends directly those in Sect. 2.1, and the coupling can be resolved by iteration or directly. This alternative to mixed dimensional setting allows an easy proof–of–concept complex nonlinear dynamics.

We consider three models which we call in shorthand by H, RS, and T. In H model of hypothermia we have constant coefficients except possibly nonlinear c_b to model vasoconstriction. For the root-soil problem RS, $c_2 = 0$, and the model for u_1 is the nonlinear Richards equation, where $D_1 = D_1(u_1)$ and $c_1 = c_1(u_1)$ feature degenerate behavior. We also consider a model T of overlapping continua involving mixed traffic participants utilizing different trajectories. In that model $f(x, u) = v(x, u)\bar{f}(u)$, and $v(x, u)$ is computed solving an auxiliary elliptic problem similar to (1) with $c = 0, \mathcal{C} = 0$.

3 Hypothermia Model: Perfusion and Vasoconstriction

We aim to build a qualitatively realistic model for the temperature in the complex tissue domain Ω such as in Fig. 1a connected to the rest of the body Ω_{body} (not shown). When exposed to low external temperatures, the body thermoregulation system attempts various strategies to prevent hypothermia and tissue damage. While these mechanisms are complex and not completely understood by physiologists, our prototype model based on overlapping continua and fictitious domain concepts could eventually aid, e.g., in developing patient-specific hypothermia therapy or cryo–preservation strategies which require rapid turnaround time from an image to simulation.

3.1 Blood Perfusion Model

The body Ω is composed of $\Omega_{\text{vessel}}, \Omega_{\text{capillaries}}, \Omega_{\text{tissue}}$. Heat transport is by convection in $\Omega_{\text{vessel}}, \Omega_{\text{capillaries}}$ and by conduction in Ω_{tissue}. Metabolic sources and products and energy are exchanged between $\Omega_{\text{capillaries}}$ and Ω_{tissue}; the latter process, called perfusion, makes heat conduction in $\Omega_{\text{capillaries}} \cup \Omega_{\text{tissue}}$ similar to flow in porous medium; see, e.g., recent developments in [23].

Literature offers various models for thermal perfusion that simplify or enhance the overlapping continua model (3a, 3b) on $\Omega = \Omega_{\text{tissue}}$, with $\theta|_{\Omega_{\text{capillaries}}} \approx \theta^* = \text{const}$, the arterial temperature. In particular, the so-called Pennes model from [19] postulates $\mathcal{C}(\theta) = c_b(\theta - \theta^*)$ in

$$c\partial_t \theta - \nabla \cdot (k \nabla \theta) + \mathcal{C}(\theta) = 0, x \in \Omega, t > 0. \tag{4}$$

Here c and k are the volumetric heat capacity and thermal conductivity, and $c_b = c_{bh} v_b = \text{const}$, where c_{bh} and v_b are blood volumetric heat capacity and perfusion rate. Here and below we ignore heat sources. In [10], $c_b(\theta - \theta^*)$ is derived by homogenization techniques in $\Omega_{\text{capillaries}}, \Omega_{\text{tissue}}$ made of ϵ-size unit cells Y with a Robin boundary condition imposed on the interface $\Gamma = Y \cap \partial\Omega_{\text{capillaries}} \cap \partial\Omega_{\text{tissue}}$, and v_b is shown to be proportional to $|\Gamma|$. In [7], Ω_{vessel} is treated as a 1D domain, with convection in Ω_{vessel} coupled to that in $\Omega_{\text{capillaries}} \cup \Omega_{\text{tissue}} \subset \mathbb{R}^3$. Several multi-equation models use a similar approach, e.g. [14, 25].

3.2 Fictitious Domain and Immersed Boundary Approaches

We aim to apply (4) in a realistic setting involving Ω_{vessel} and a complicated external boundary $\partial\Omega$. For the former, we include a penalty term $c_{\text{vessel}}(\theta - \theta^*)$ on Ω_{vessel} to enforce $\theta|_{\Omega_{\text{vessel}}} \approx \theta^*$; for the latter, we proceed similarly on some $\tilde{\Omega} \setminus \Omega$ with

$\tilde{\Omega}$ of simpler shape than Ω. These treatments resemble the Immersed Boundary [18] and fictitious domain [13] approaches. We define $\mathcal{C}(\theta) = C(x)(\theta - \theta^*)$, with C equal $c_b, c_{\text{vessel}}, c_D$ on each $x \in \Omega \setminus \Omega_{\text{vessel}}$, Ω_{vessel}, $\tilde{\Omega} \setminus \Omega$, and $\theta^* = \theta^*(x)$ equal $\theta_{\text{body}}, \theta_{\text{vessel}}, \theta_D$, respectively. We postulate an extension of (4) to $\tilde{\Omega}$ as follows

$$\tilde{c}\partial_t\tilde{\theta} - \nabla \cdot (\tilde{k}\nabla\tilde{\theta}) + \mathcal{C}(\tilde{\theta}) = 0, x \in \tilde{\Omega}, t > 0, \tag{5}$$

The coefficients c, k can be extended to $\tilde{\Omega}$ in any convenient way as long as (5) is well-posed. The model (5) is next approximated by the mixed finite elements/CCFD setting as described in Sect. 2.1.

3.3 Examples: Model Adaptivity for Hypothermia

We illustrate the effect of the blood perfusion term $c_b(\theta - \theta^*)$ and fictitious domain term $c_D(\theta - \theta^*)$ in a hypothermia example based on geometry in Fig. 1a aiming for a physically motivated scenario to simulate the onset of frostbite in the human hand exposed to cold air over $t \le T = 1200$ s. We extend $\Omega = \Omega_{\text{hand}} \subset \tilde{\Omega}$ by Ω_{air} so that $\tilde{\Omega}$ is a large enough rectangular domain. The tissue in Ω_{hand} has heterogeneous properties including large blood vessels. We identify $\partial\Omega_{\text{wrist}} = \partial\Omega \cap \{x : x_2 = 0\}$ as part of $\partial\Omega_D$ made of $\partial\Omega_{\text{wrist}} \cup \partial\Omega_{\text{nonwrist}}$.

We start with a $d = 1$ slice of the 2d domain where the effect of c_b, c_D is easily quantified. Convergence studies (not shown) show expected order $O(\tau + h^2)$.

Example 1 Consider $\Omega = (0, L)$, $L = 0.15$ m, and $\Omega \subset \tilde{\Omega} = (-0.5, 0.15)$ with $\Omega_{\text{vessel}} = \emptyset$. Let $c = 3 \times 10^6$, $k = 0.3$, $\theta^* = 37$. Set $\partial\Omega_{\text{nonwrist}} = \{0\}$, $\partial\Omega_{\text{wrist}} = \{0.15\}$ and $\theta_{\text{nonwrist}} = -40$, $\theta_{\text{wrist}} = 37$. Let $\theta(x, 0) = \theta_{\text{init}}(x) = -40\chi_{[-0.05,0]} + (\frac{1540}{3}x - 40)\chi_{[0,0.15]} = \lim_{t\to\infty} \theta(x, t)$. To simulate, we let $\tau = 1$, and $h = 10^{-4}$, with $M = 2000$ cells. We try $c_b \in \{0, 10, 10^2, 10^3\}$ and $c_D \in \{0, 10^0, 10^1, \dots, 10^8\}$.

Figure 2 illustrates how c_b and c_D work together. Large c_b drives $\theta|_\Omega$ towards θ^*, and $\theta|_\Omega$ is essentially steady state at $t = T$. In turn, a large c_D more strictly enforces $\theta|_{\tilde{\Omega}\setminus\Omega} = \theta^D$, though one must keep the condition number $\kappa(\mathcal{M}_h)$ of (2) reasonable. For example, $c_b = 10^3$, $c_D = 10^7$, gives $\theta(0, t^N) = -39.42$, which reasonably approximates the Dirichlet condition at $\partial\Omega_{\text{nonwrist}}$.

We consider next the 2d example, where the hand is protected by mitten material in Ω_{mitten}, case (M), or not, case (N). Other notation is adapted easily.

Example 2 We use a uniform grid 100×133 over $\tilde{\Omega}$ shown in Fig. 3. The coefficients c, k, c_b, θ^* are given in Table 1. We use $\theta|_{\partial\Omega_{\text{nonwrist}}} = \theta^*_{\text{air}} = -40$ °C, $\theta|_{\partial\Omega_{\text{wrist}}} = \theta^*_{\text{body}} = 37$ °C, and let $\theta(x, 0) = \chi_\Omega\theta_{\text{body}} + \chi_{\tilde{\Omega}\setminus\Omega}\theta_{\text{air}}$.

The model exhibits qualitatively intuitive behavior, as shown in Fig. 3. Next we post-process the results to analyze the extent of hypothermia and possible frostbite, i.e. $\theta(x, t) < 0$ °C. Frostbite is avoided in case (M) at least until $t = 1200$, but occurs in case (N), affecting about 11% of cells in Ω_{hand}.

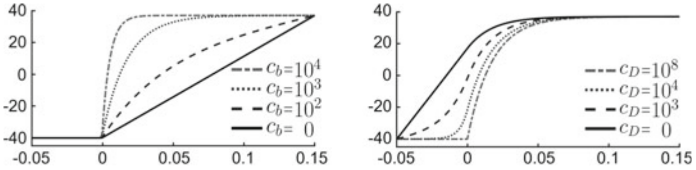


Fig. 2 Simulation results for Example 1. Left: near steady-state solutions ($t = 1200$); various c_b and $c_D = 10^7$. Right: near steady-state solutions; $c_b = 10^3$ and various c_D

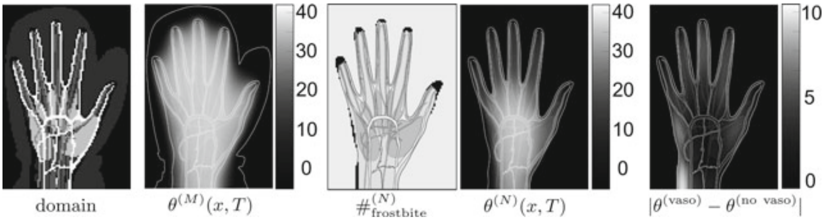


Fig. 3 Simulation results for Examples 2 and 4. Cases are denoted by superscripts

Table 1 Material parameters for Examples 2 and 4

Domain	Material	c (J K^{-1} m^{-3})	k (W m^{-1} K^{-1})	c_b (W m^{-3})	θ^* (°C)
Ω_1	Bone	2.7×10^6	0.31	10^{-3}	37
Ω_2	Muscle	3.7×10^6	0.49	10^{-3}	37
Ω_3	Nerve	4.1×10^6	0.49	10^{-3}	37
Ω_4	Skin	3.7×10^6	0.37	10^{-3}	37
Ω_5	Tendon	3.8×10^6	0.47	10^{-3}	37
Ω_{vessels}	Blood	3.8×10^6	0.52	10^5 or 1	37
Ω_{mitten}	Goose down	1512	0.16	0	37
$\tilde{\Omega} \setminus \Omega$	Air	858	0.024	10^5	–40

Next we address the quality of model adaptation. The term $\mathcal{C}(\theta)$ simulates perfusion in $\Omega_{\text{hand}} \setminus \Omega_{\text{vessels}}$ and acts as a penalty term in Ω_{vessels} and $\tilde{\Omega} \setminus \Omega$. While c_b has physiological/modeling meaning [10], c_{vessel} and c_D are chosen somewhat ad-hoc to ensure, respectively, $\theta|_{\Omega_{\text{vessel}}} \approx \theta^*|_{\Omega_{\text{vessel}}}$ and $\theta|_{\partial\Omega_{\text{nonwrist}}} \approx \theta^*|_{\Omega_{\text{air}}}$. We test the quality of the adapted model by checking if this first condition holds. We record $\theta_{\max} = \max_{ij} \theta^n_{ij}$ at $t^n = T$. When $c_{\text{vessel}} = 10^5$, case (N) has $\theta_{\max} = 35.7$. However, when $c_{\text{vessel}} = 1$, case (N) has $\theta_{\max} = 9.8$. Case (M) is similar.

3.4 Hypothermia with Vasoconstriction

With model sensitivity to $c_b, c_D, c_{\text{vessel}}$ reasonably understood, we extend (4) to account for defense against hypothermia with vasoconstriction, when decrease in the body temperature captured by thermoreceptors induces arterial smooth muscle constriction, reducing blood flow in extremities to retain heat near the core body. However, such action may cause permanent morphological changes, e.g. frostbite, in extremities.

To model vasoconstriction, we recall c_b in (4) depends on Γ, thus we hypothesize that c_b decreases when sensory data $\mathcal{S}(\theta) = \langle\theta\rangle_\Omega$ decreases, affecting the venous blood temperature $\approx \mathcal{S}(\theta)$ returning to the body. Without increased metabolism, the body core temperature θ_{body} may decrease, further decreasing $\mathcal{S}(\theta)$, c_b, and $\theta_{\partial\Omega_{\text{wrist}}}$. We postulate the following simple model

$$\theta_{\partial\Omega_{\text{wrist}}} = \theta^* = \theta_{\text{body}}, c_b = c_b(\theta_{\text{body}}); \frac{d\theta_{\text{body}}}{dt} + c_B(\theta_{\text{body}} - \mathcal{S}(\theta))_+ + \lambda = \mu, \quad (6)$$

with $\mu \geq 0$ representing metabolism and λ a Lagrange multiplier ensuring $\theta_{\text{body}} \leq 37$, mimicking thermoregulation. We solve (6) coupled to (4) semi-implicitly, requiring some $c_B \geq 0$, $\mu \geq 0$, and some model for $c_b = c_b(\theta_{\text{body}})$.

Example 3 We consider data as in Example 1, and consider (4), (6) first with (a) constant fixed $\theta_{\text{body}} = 37$, $\theta^* = 37$, $c_b = 10^3$. Next we set up the variable nonlinear model with $c_B = 10$. We also consider two vasoconstriction model variants (b) $c_b(\theta) = 10^3\chi_{\theta>10}$, and (c) $c_b(\theta) = 10^3\frac{\theta_{\text{body}}}{37}\chi_{\theta>10}$. For dramatic effects, we consider $\mu = 0$, study $\theta_{\text{body}}(t)$ and retrieve the position $x^* \in \Omega : \theta(x^*, t) = 0$ which indicates the extent of frostbite.

Simulations for Example 3 confirm intuition. Upon vasoconstriction (b-c), frostbite is more extensive than in (a), but not dramatically. There is not much difference in θ between (b) and (c), but there is difference in θ_{body} between (b-c).

Example 4 We consider data as in Example 2, but with $c_b = 10^{-3}$. We repeat the experiments for $t \leq T = 120$ s, varying c_b, c_D and the vasoconstriction model, with $\mathcal{S}(\theta) = \langle\theta\rangle$. We set $c_b = $ const (vaso model off), or variable $c_b(\theta) \sim \chi_{\theta>10}$ (vaso model on) to model an instantaneous response to θ.

The results in Table 2 demonstrate some model sensitivity to the parameters. There is little dependence on c_b if $c_b \leq 10^4$, perhaps because c_{vessel} is fixed. As expected, less frostbite occurs when vasoconstriction is not active; see also Fig. 3 for comparison of θ^{vaso} and $\theta^{\text{no vaso}}$.

Summary: We believe the CCFD implementation of the H model is robust with the fictitious domain and model variants. The results agree qualitatively with the intuition; the most crucial parameters are c_b and c_{vessel}.

Table 2 Sensitivity of the H model with vasoconstriction in Example 4 at $t = T = 120$ s to c_b, c_D, and the choice of vasoconstriction model (vaso or off). We define $\#_{\text{frostbite}}$ as the number of the cells ω_{ij} where $\theta_{ij}|_{t=T} < 0$, and $\#_{\text{off}}$ those with $\theta_{ij}|_{t=T} < 10$

c_D	c_b	vaso	$\#_{\text{frostbite}}$	$\#_{\text{off}}$	θ_{min}	θ_{max}	θ_{ave}
10^6	10^{-3}	Yes	42	165	–16.03	36.99	27.28
10^6	10^{-3}	No	42	165	–16.03	36.99	27.28
10^6	10^2	Yes	42	163	–16.04	36.99	27.29
10^6	10^2	No	42	163	–15.98	36.99	27.29
10^6	10^4	Yes	32	131	–15.37	36.99	28.18
10^6	10^4	No	22	124	–11.81	36.99	28.28
10^6	10^5	Yes	1	3	–0.02	36.98	32.76
10^6	10^5	No	0	2	8.38	36.98	32.78
10^7	10^4	Yes	6	40	–9.03	36.99	32.66
10^7	10^4	No	5	39	–6.65	36.99	32.70
10^7	10^5	Yes	0	2	0.87	37	34.46
10^7	10^5	No	0	2	8.70	37	34.47

4 Root–Soil Flow Model in $d \geq 1$

Consider a complex domain shown in Fig. 1b representing a plant root embedded in soil, adapted from [22]. Our long-term goal is to simulate water flow in this root-soil system combined with other coupled phenomena including surface and above-surface models plus energy equation. Therefore, even though roots have a much smaller volume than reasonable soil volumes, we aim to build a general flexible physically meaningful RS model in $d \geq 1$.

Our RS model extends the $d = 1$ overlapping continua root-soil models from [4, 26], a nonlinear Richards-Darcy generalization of (3a, 3b). We consider $\Omega_r \subset \Omega_s \subset \mathbb{R}^d$, with $d \geq 1$, both with positive $\mathbb{R}^d$ measures $|\Omega_r| << |\Omega_s|$, and denote soil/root variables by subscripts $_s/_r$. We maintain the overlapping continua feature and do not resolve the root's micro-structure. We use Richards equation in unsaturated soil domain Ω_s for (incompressible) water flow in Ω_s coupled to saturated flow in Ω_r as follows

$$\phi_s \frac{\partial S_s}{\partial t} + \nabla \cdot q_s = -c(P_s - P_r), \;\; x \in \Omega_s, t > 0, \tag{7a}$$

$$\nabla \cdot q_r = \;\; c(P_s - P_r), \;\; x \in \Omega_r, t > 0, \tag{7b}$$

$$q_s = -\frac{K_s}{\mu} k(S_s)\left(\nabla P_s - \rho g \nabla D\right), \qquad q_r = -\frac{K_r}{\mu}\left(\nabla P_r - \rho g \nabla D\right), \tag{7c}$$

$$P_s(x, 0) = P_{s,init}, \;\; x \in \Omega_s, \tag{7d}$$

$$P_s(x, t) = -P_c(S_{s,D}), \;\; x \in \partial\Omega_{s,D}; \;\; q_s \cdot \eta = q_{s,N}, \;\; x \in \partial\Omega_{s,N}, \tag{7e}$$

$$P_r(x, t) = P_{r,D}, \;\; x \in \partial\Omega_{r,D}; \;\; q_r \cdot \eta = q_{r,N}, \;\; x \in \partial\Omega_{r,N}. \tag{7f}$$

Here S_s is the saturation (volume fraction) of water, P_m and q_m are the pressure and flux of the fluid in Ω_m. We have $P_s = -P_c(S_s)$, where P_c is capillary pressure. In addition, ϕ and K are porosity and permeability of the medium, $k(S)$ is the relative permeability, μ and ρ are viscosity and density of the fluid, g is the gravitational acceleration, D is the depth under the soil's surface, and $c = c_{rs}\chi_{\Omega_r}$ is the exchange term between root and soil, with $c_{rs} \geq 0$. Initial and boundary data are as stated in (7d)–(7f).

The model (7a) is derived when $d = 1$ in [4, 26] based on several modeling assumptions. Model in [4] emphasizes the geometrical complexity of the root-soil exchange, and works in potentials $\frac{P_m}{g}$ as variables. The idea is to set up the fluid exchange between root and soil in a simple manner and avoid discretization at the scale of individual xylems through which the actual water transport takes place. To this end, the radial character of the flow to the roots, the low conductivity of the inner portion of the roots (endodermis), along with high permeability of epidermis and medium permeability of cortex, are assumed. In [26] this model is extended to allow subroots and root network with stochastic updates to this $d = 1$ model, and predict that dry and wet zones will develop in the soil as a result of water uptake by plant roots. Both [4, 26] provide formulas for c_{rs}.

4.1 Computational Model Specifics for d $\geq$ 1 and Challenges

First we consider $c_{rs} = 0$. The main well known difficulty when working with Richards equation in any dimension is the nonlinear degenerate parabolic character of the problem due to the behavior of $k(S)$, $P_c(S)$. We illustrate this behavior using the well known algebraic model based on experimental data known as van-Genuchten-Mualem model

$$P_c(S) = \frac{1}{\alpha}(S^{-\frac{1}{m}} - 1)^{\frac{1}{\nu}}, \quad k(S) = S^{\epsilon}\left[1 - \left(1 - S^{\frac{1}{m}}\right)^m\right]^2. \tag{8}$$

The set of parameters $\epsilon = \frac{1}{2}$, $m = 1 - \frac{1}{\nu}$, $\alpha = 10^{-4}$, and $\nu = 2.237$ characterizes fine soil; see this and others collected in [21]. A change of variables in (7a) reveals that an equivalent model reads

$$\partial_t S - \nabla \cdot (D_s(S)\nabla S) + \nabla \cdot A_s(S) = 0.$$

Here $D_s(S) \geq 0$ but degenerates to 0 when $S \downarrow 0$. Thus the model features a degenerate behavior, while the nonlinear advective term $A(S)$ when $D \neq$ const requires careful treatment of this hyperbolic term. Due to these difficulties, the use of lower order numerical scheme such as that in Sect. 2.1 is appropriate, and so is the use of upwinding. Furthermore, the choice of primary unknowns is delicate, since $P_c(S)$ is unbounded near $S \downarrow 0$, but $P_c^{-1}(p)$ is unbounded when $p \downarrow 0$. These features prompted various analyses of numerical schemes and delicate discussion of nonlinear solvers; see, e.g., [5, 20, 21, 24]. In particular, [5, 24] prove $O(h+\tau)$ error estimates to the mixed finite element method, but some of this work requires regularization and strong assumptions on smoothness. In turn, [21] evaluates the fully implicit CCFD schemes for nontransformed, nonregularized models with strong heterogeneities and locally large fluxes, and test variants of averaging, implicit or semi-implicit solutions, and different choices of primary unknowns.

When $c_{rs} > 0$, the computational model for RS system features an additional difficulty, since most likely $|\Omega_r| << |\Omega_s|$. With the approach we outlined in Sect. 2.1, this is, however, not an issue. We are able to simulate successfully the flow of water in the overlapping continua system; this is illustrated by our examples below.

4.2 Examples for RS Model

Example 5 We consider $\Omega_r = \Omega_s = (0, L)$ and $L = 1$ m, other data in Table 3, and $k(S)$, $P_c(S)$ given by (8). To model evaporation and precipitation at the surface of the ground, we impose Neumann boundary conditions at $x = 0$, and at $x = L$:

$$q_s(0,t) = q_{s,\text{top}}, \quad q_r(0,t) = q_{r,\text{top}}, \quad q_s(1,t) = 0, \quad q_r(1,t) = 0, \tag{9}$$

$$q_{s,\text{top}} = \begin{cases} -4\times 10^{-7}, & \text{if } t \in (0, 1.2] \cup (3.5, 4] \ \ [\text{days}], \\ \ \ 2\times 10^{-7}, & \text{if } t \in (1.5, 2] \ \ [\text{days}], \\ \ \ 0, & \text{otherwise}, \end{cases}$$

$$q_{r,\text{top}} = \begin{cases} -10^{-8}, & \text{if } t \in (0, 1.2] \cup (1.5, 2] \cup (3.5, 4] \ \ [\text{days}], \\ \ \ 0, & \text{otherwise}. \end{cases}$$

The simulation results in Fig. 4 show response of the system to the infiltration through boundary (9) over 4 days.

Table 3 Parameter values for Example 5

Parameter	ϕ_s	$K_s[\text{m}^2]$	$K_r[\text{m}^2]$	$\mu[\text{Pa}\cdot\text{s}]$	$\rho[\text{kg/m}^3]$	$g[\text{m/s}^2]$	c	$\tau[\text{s}]$	$h[\text{m}]$
Value	0.4	10^{-11}	5×10^{-12}	10^{-3}	10^3	9.8066	10^{-10}	600	1/50

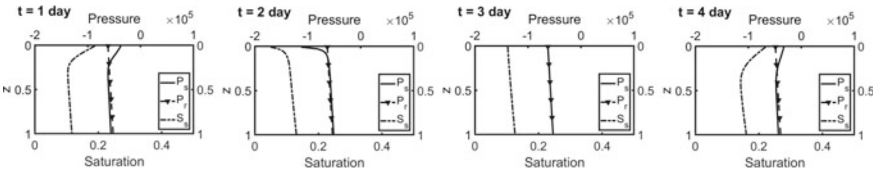
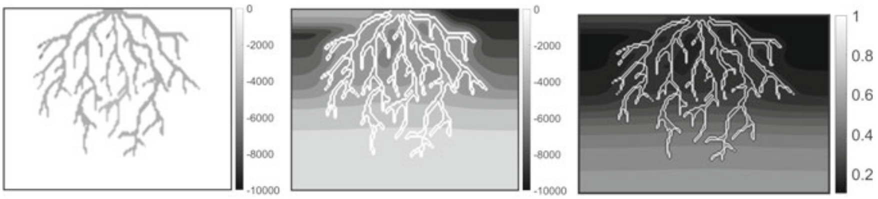


Fig. 4 Solutions P_s, S_s, and P_r in Example 5: S_s and P_s rise when there is precipitation, and drop due to the root's absorption of water from the soil under dry weather



$P_r|_{t=2400}$ $P_s|_{t=2400}$ $S_s|_{t=2400}$

Fig. 5 Numerical solution for P_r, P_s, and S_s, from Example 6. For P_s, S_s, the contour $\partial\Omega_r$ is superimposed over Ω_s to guide the eye. Vertical axis is aligned with gravity

Example 6 We continue with data from Example 5 in $\Omega_s = [0, 1] \times [0, 1]$, grid 120×80 over $t \leq T = 3600$ s. We simulate infiltration with the Richards–Darcy model, starting from initial (pressure-saturation) equilibrium (by setting $q_s = q_r = 0$, $P_s = P_r$, and $P_{s,\text{top}} = P_{r,\text{top}} = 0$), adding water only from the left top portion of $\partial\Omega_s$, with no flux conditions elsewhere. We use the coupling coefficient $c_{rs} = 10^{-8}$.

The plots in Fig. 5 show that the P_r is well equilibrated with P_s. The water from the left boundary eventually reaches the root. The solutions feature sharp fronts as usual for Richards equation. The system strongly depends on c_{rs}.

Summary: We find the RS model based on Richards-Darcy equations quite complex. As long as its Richards portion is well calibrated, the model is robust. In the coupled system, there is high sensitivity to the coefficient c_{rs}.

5 Mixed-Use Traffic Flow on Campus

Traffic flow models can have continuum (PDE) or discrete (individual based model) form. The former have gained interest since the LWR models [17] and present interesting hyperbolic structure of the underlying PDE, the (macroscopic) transport model $u_t + \nabla \cdot f(u) = 0$ (supplemented by inflow boundary conditions) for the average density u (of cars) with flux function $f(u)$, e.g., $f(u) = (1 - \frac{u}{u_{max}})uv$. However, these traffic models for cars on paved roadways are inherently one-dimensional, and since the spatial scale is much larger than the average length of of car, individualistic behavior is not preserved.In contrast, microsopic individualistic behavior models are important for emergency scenarios such as during tsunami or wildfires as well as for urban and architectural design considerations of possible congestion patterns. However, such models on network can become unmanageable, and efforts to manage their complexity involve upscaling to a model resembling "Darcy flow" [9].

Computational models of traffic scenarios start with an image such as in Fig. 1c which describes the paths Ω^P within a domain $\Omega \subset \mathbb{R}^2$. One possible model for traffic on this network augments the LWR flux with $-K\nabla u$, where K is a diffusion tensor and v captures the preferred direction of motion; both are based on the mean value and standard deviation of transition probabilities between sites at each time step [9]. However, this approach does not include coupling between the different species nor allow travel off the network.

The model we build accounts for the traffic on and off Ω^P as well as for interactions between the species. Our simulations might support the design, control, and monitoring of robot trajectories as well as of campus pathways.

5.1 Background and Model Development

We consider traffic with $M = 3$ species: human pedestrians, humans on bicycles, and autonomous delivery robots, numbered $m = 1, 2, \dots 3$. We let average densities $u = (u_1, \dots u_M)$ and flux function $f(u) = (f_1(u), \dots f_M(u))$. We set $f_m(x, t; u) = v_m(x, t; u)\bar{f}_m(u)$ to separate the trajectories $v_m(x, t; u)$ from traffic awareness modelled by $\bar{f}_m(u)$. While robots or bicycles can only travel on paved paths within some $\Omega^P \subset \Omega$, humans can alter their trajectory and veer off Ω^P when they are aware of a (possible) congestion.

Flow and trajectories: Each species m has a trajectory given by the velocity field $v_m = v_m(x)$, $x \in \Omega$. Most can only move on the paved paths, i.e. so that supp $v_m \subset \Omega^P$. Typically most species are aware of others and may alter their speed locally in time, but most cannot alter their trajectories. For some species, say m', we allow $v_m' = v_m'(x, t, (u_m)_m)$ with supp $v_{m'} \setminus \Omega^P \neq \emptyset$ depending on the current conditions.

To determine a trajectory v_m, assume that species m intends to get from some inlet point $x_m^{in} \in \partial\Omega$ to some outlet point $x_m^{out} \in \partial\Omega$. We solve a pseudo-potential problem $\nabla \cdot v_m = 0$ for ψ_m so that $v_m = -K_m \nabla \psi_m$. We set $\partial\Omega_D = \Gamma_{in} \cup \Gamma_{out}$ where each Γ_{in}, Γ_{out} is a small region of $\partial\Omega$ around the inlet and outlet, respectively, and require the Dirichlet condition $\psi|_{\Gamma_{in}} = 1$, and $\psi|_{\Gamma_{out}} = 0$, and we require Neumann condition $v_m \cdot n|_{\partial\Omega_N} = 0$ on $\partial\Omega_N = \partial\Omega \setminus \partial\Omega_D$. Solving for v_m is a well-posed problem with a scheme for "flow" from Sect. 2.1.

It remains to set up the species specific "mask" K to determine these paths preferentially and allow, e.g., to set up alternative pathways to stairs, or set up ad-hoc detours. In the simplest scenarios we simply set $K|_{\Omega^P} = \kappa_m$ and $K|_{\Omega\setminus\Omega^P} = 0$, and κ_m can be calibrated to an average speed of species m. Also, the species "aware" of traffic are allowed to alter their trajectory with $K_m = K_m(x, y, (u_m)_m)$ built heuristically. For example, if at some point $(x, y) \in \Omega^P$ we have a traffic congestion in some area Ω^C, we would set $K_m|_{\Omega^C} = 0$ and $K_m|_{\Omega\setminus\Omega^C} = \kappa_m$, and recalculate v_m to avoid Ω^C.

The transport model: With $f_m = v_m \bar{f}_m$ and a given v_m, $\bar{f}_m(u)$ can be a linear or nonlinear flux function, e.g., the LWR model.

Scheme: To approximate the flow and transport solutions, we follow Sect. 2.1 and [8, 17] and apply Godunov's method, under CFL condition. For $M > 1$ the situation is more complicated but for the simple case studies we develop the Godunov method suffices. For the simulations in $d = 1$, we report grid refinement studies; we also confirm grid error for the nonlinear problems to be $O(\sqrt{h})$ (not shown).

5.2 Examples of Simulations with T Model

To build up our intuition, we illustrate the coupled dynamics of the $M = 2$ species with $\Omega = (0, 2)$, assuming they travel in the same direction. The pedestrian density is u_1, and that of robots is u_2. We construct the flux functions heuristically to model the interaction between the species, and in particular "traffic awareness". Species 1 (humans) do not feature awareness of other species and $\bar{f}_1(u) = u_1(1 - u_1)$ (LWR model), but $\bar{f}_2(u) = v_2(u_1)u_2$ where the robots slow down considerably when they notice humans, with

$$v_2(u_1) = \begin{cases} 0.5, & u_1 < 0.5 \\ -4u_1 + 2.5, & 0.5 \leq u_1 \leq 0.6 \\ 0.1, & u_1 > 0.6. \end{cases} \tag{10}$$

Example 7 We prescribe $u_1^{init}(x) = \chi_{[0.6,0.7]}(x)$ and $u_2^{init}(x) = 0.1\chi_{[0.4,0.8]}$ to illustrate the traffic flow and "traffic awareness": the robots start ahead, alongside and behind human pedestrians on the 1d network; see Fig. 6.

We see that u_1 feature a rarefaction typical in LWR traffic flow models [17]. Also, the robots follow linear advection away from pedestrians, but develop two traveling waves for robots ahead and behind the humans according to (10). The snapshot at $t = 0.75$ shows that $u_2|_{[0.5,0.65]}$ matches linear advection, but when $u_1 > 0.6$, the robots slow down to $v_2(u_1) = 0.1$ causing a large spike.

Example 8 Now we consider the campus network with Ω^P shown in Fig. 1c. We simulate mixed dimensional network traffic on Ω^P with the occasional traffic off Ω^P. We design trajectories of the species as shown in Fig. 7, and we simulate the transport as shown in Fig. 8. In one variant, (a) humans stay on pavement. In another, (b) humans venture off the path to avoid congestion.

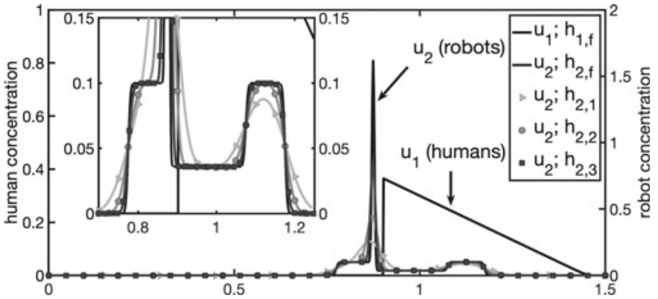


Fig. 6 Illustration for Example 7: plots at at time $t = 0.75$ including grid refinement for u_2 with $h_{2,1} = 5 \cdot 10^{-3}$, $h_{2,2} = 1.25 \cdot 10^{-3}$, and $h_{2,3} = 3.125 \cdot 10^{-4}$ with finest grid $h_{i,f} = 7.8125 \cdot 10^{-5}$ for $i = 1, 2$

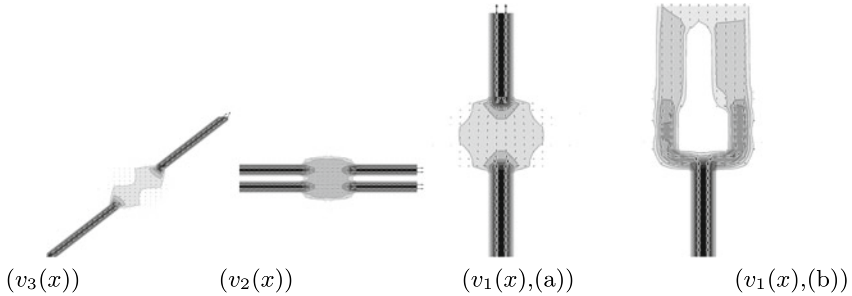
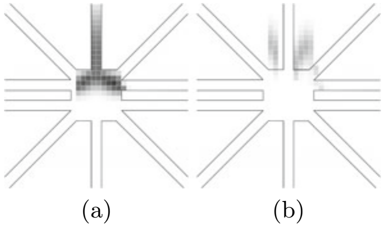


Fig. 7 Trajectories v_m in Example 8 for u_3 (bicycles), u_2 (robots) and u_1 (humans) on **a** paved paths and **b** on the grass due to congestion

Fig. 8 Results of Example 8 at time $t = 0.5$ show the average human concentration u_1 in case (**a**). In case (**b**) due to congestion by u_2 and u_3 the pedestrians alter v_1 off the network Ω^P to avoid the bottleneck



Summary: The computational framework of overlapping continua allows to simulate complex traffic patterns involving species of very different trajectories. More work is needed to identify the data for such simulations and applications to the design of campus network and of, e.g., robot software.

6 Summary and Outlook

Our implementation for each of the applications H, RS, and T is efficient and based on a common robust CCFD framework which allows for approximate enforcement of Dirichlet conditions and easy accounting for complex domains. The model approximations and enhancements were easily introduced; we also identified critical model components and parameters. For H, it is the c_b coefficient, and its role in the vasoconstriction. For RS, it is by far also the exchange coefficient c_{rs}. For T, it is the local velocity, the mask K, and the interaction models between the species. These elements should be validated with data.

However, parameter identification based on experimental or imaging data remains quite challenging for coupled systems, where identifying proper relationships and models can be difficult. The prototype models we presented serve as a useful first step to identify the main features of dynamics.

Acknowledgements This research was partially supported by the grants NSF DMS-1522734, DMS-1912938, PI: Malgorzata Peszynska.

References

1. IPARS, A New Generation Framework for Petroleum Reservoir Simulation. https://csm.oden.utexas.edu/ipars/
2. Color Hand Anatomy Concept Banner Poster Card. https://www.istockphoto.com/
3. PETSC, Portable, Extensible Toolkit for Scientific Computation. https://petsc.org/release/
4. Arbogast, T., Obeyesekere, M., Wheeler, M.F.: Numerical methods for the simulation of flow in root-soil systems. SIAM J. Numer. Anal. **30**(6), 1677–1702 (1993)
5. Arbogast, T., Wheeler, M.F., Zhang, N.Y.: A nonlinear mixed finite element method for a degenerate parabolic equation arising in flow in porous media. SIAM J. Numer. Anal. **33**(4), 1669–1687 (1996)
6. Barenblatt, G.I., Zheltov, I.P., Kochina, I.N.: Basic concepts in the theory of seepage of homogeneous liquids in fissured rocks (strata). J. Appl. Math. Mech. **24**, 1286–1303 (1960)
7. D'Angelo, C., Quarteroni, A.: On the coupling of 1d and 3d diffusion-reaction equations: application to tissue perfusion problems. Math. Models Methods Appl. Sci. **18**(08), 1481–1504 (2008)
8. Dawson, C.: Godunov-mixed methods for advection-diffusion equations in multidimensions. SIAM J. Numer. Anal. **30**(5), 1315–1332 (1993)
9. Della Rossa, F., D'Angelo, C., Quarteroni, A., et al.: A distributed model of traffic flows on extended regions. Netw. Heterog. Media **5**(3), 525–544 (2010)
10. Deuflhard, P., Hochmuth, R.: Multiscale analysis of thermoregulation in the human microvascular system. Math. Methods Appl. Sci. **27**(8), 971–989 (2004)
11. 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**(4), 1089–1116 (2014)
12. Gjerde, I.G., Kumar, K., Nordbotten, J.M.: A mixed approach to the Poisson problem with line sources. SIAM J. Numer. Anal. **59**(2), 1117–1139 (2021)
13. Glowinski, R., Pan, T.W., Periaux, J.: A fictitious domain method for dirichlet problem and applications. Comput. Methods Appl. Mech. Eng. **111**(3–4), 283–303 (1994)
14. He, Z.Z., Liu, J.: A coupled continuum-discrete bioheat transfer model for vascularized tissue. Int. J. Heat Mass Transf. **107**, 544–556 (2017)
15. Hochmuth, R.: Homogenization for a non-local coupling model. Appl. Anal. **87**(12), 1311–1323 (2008)
16. Klein, V., Peszynska, M.: Adaptive double-diffusion model and comparison to a highly heterogeneous micro-model. J. Appl. Math. **2012**, 1–26 (2012)
17. LeVeque, R.J.: Finite volume methods for hyperbolic problems, vol. 31. Cambridge University Press (2002)
18. Mittal, R., Iaccarino, G.: Immersed boundary methods. Annu. Rev. Fluid Mech. **37**, 239–261 (2005)
19. Pennes, H.H.: Analysis of tissue and arterial blood temperatures in the resting human forearm. J. Appl. Physiol. **1**(2), 93–122 (1948)
20. Peszynska, M., Jenkins, E.W., Wheeler, M.F.: Boundary conditions for fully implicit two-phase flow. In: Recent Advances in Numerical Methods for Partial Differential Equations and Applications, vol. 306, pp. 85–106 (2002)
21. Peszynska, M., Yi, S.Y.: Numerical methods for unsaturated flow with dynamic capillary pressure in heterogeneous porous media. Int. J. Numer. Anal. Model. **5**, 126–149 (2008)
22. Polyashenko, O.: Set of black tree roots isolated on white background (2020). https://www.istockphoto.com/vector/web-gm1250518500-364760337

23. Qohar, U.N.A., Zanna Munthe-Kaas, A., Nordbotten, J.M., Hanson, E.A.: A nonlinear multiscale model for blood circulation in a realistic vascular system. R. Soc. Open Sci. **8**(12), 201,949 (2021)
24. Radu, F., Pop, I., Knabner, P.: Order of convergence estimates for an Euler implicit, mixed finite element discretization of Richards' equation. SIAM J. Numer. Anal. **42**(4), 1452–1478 (2004)
25. Reichold, J., Stampanoni, M., Keller, A.L., Buck, A., Jenny, P., Weber, B.: Vascular graph model to simulate the cerebral blood flow in realistic vascular networks. J. Cereb. Blood Flow Metab. **29**(8), 1429–1443 (2009)
26. Roose, T., Fowler, A.C.: A model for water uptake by plant roots. J. Theor. Biol. **228**(2), 155–171 (2004)
27. Showalter, R.E., Visarraga, D.B.: Double-diffusion models from a highly-heterogeneous medium. J. Math. Anal. Appl. **295**(1), 191–210 (2004)
28. Vidotto, E., Koch, T., Kooeppl, T., Helmig, R., Wohlmuth, B.: Hybrid models for simulating blood flow in microvascular networks. Multiscale Model. Simul. **17**(3), 1076–1102 (2019)

Elliptic and Parabolic Problems: Contributed Papers

PolyMAC: Staggered Finite Volume Methods on General Meshes for Incompressible Navier–Stokes Problems

Pierre-Loïc Bacq, Antoine Gerschenfeld, and Michael Ndjinga

Abstract We consider new finite volume methods of staggered type for incompressible Navier–Stokes equations. These methods, called PolyMAC, generalize the MAC scheme to general meshes. We briefly describe the different versions of PolyMAC and present guidelines for an efficient use. To do so, we define a benchmark of 3D Navier–Stokes problems based on the successive benchmarks established during previous FVCA conferences. Finally, we identify weaknesses of each PolyMAC version and suggest ways to improve the current performances.

Keywords Finite volumes · General meshes · Staggered scheme · Navier–Stokes

1 Introduction

We are interested in the discretisation of the incompressible Navier–Stokes equations: find $\vec{u}$ and p such that

$$\begin{aligned} \partial_t \vec{u} + (\vec{u} \cdot \nabla)\vec{u} - \nu \Delta \vec{u} + \nabla p &= \vec{f} \quad \text{in } \Omega \,, \\ \nabla \cdot \vec{u} &= 0 \quad \text{in } \Omega \,, \end{aligned} \tag{1}$$

where $\vec{u}$ is the velocity, p the pressure and $\nu > 0$ is the viscosity. In this instance, the domain $\Omega \subset \mathbb{R}^3$ is the unit cube in 3D.

In this paper, we compare three versions of PolyMAC, a Finite Volume (FV) discretisation scheme which generalizes the MAC [9] scheme to general meshes. Indeed, MAC-like schemes yield robust discretisations of fluid dynamics equations and prevent the emergence of spurious modes, but are restricted to Cartesian meshes. The comparison is made on a benchmark of Navier–Stokes problems defined on various meshes introduced during the different FVCA conferences [4, 7]. We estimate

P.-L. Bacq (✉) · A. Gerschenfeld · M. Ndjinga
CEA, DES/ISAS/DM2S/STMF, Université Paris-Saclay, 91191 Gif-sur-Yvette, France
e-mail: pierre-loic.bacq@cea.fr

E. Franck et al. (eds.), *Finite Volumes for Complex Applications X—Volume 1, Elliptic and Parabolic Problems*, Springer Proceedings in Mathematics & Statistics 432,
https://doi.org/10.1007/978-3-031-40864-9_9

the order of convergence of each PolyMAC method and the density of the resulting linear system and we propose guidelines for the use of PolyMAC. Note that all three schemes are implemented on the TRUST platform [5].

The remaining of this paper is organised as follows. In Sect. 2, we describe briefly the different versions of PolyMAC and how they differ. Section 3 presents the different meshes and the equations approximated in this paper, while the numerical results are detailed in Sect. 4. Finally, conclusions are drawn in Sect. 5.

2 PolyMAC

The key idea behind the MAC scheme, namely the staggering of the velocity and pressure unknowns, is transposed in the PolyMAC context by selecting as unknowns the normal component of the velocity at the faces and the pressure at the center of the cells. Depending on the version of PolyMAC considered, auxiliary unknowns are introduced, but the staggering of the main unknowns allows to keep the benefits of the MAC method.

2.1 PolyMAC I

PolyMAC I is a mimetic FV discretisation very similar to the one analysed in [3], except for the discretisation of the convection term, which we summarise here. First, we use the incompressibility constraint to rewrite the convection term as $\nabla \cdot (\vec{u} \otimes \vec{u})$ and we introduce the vorticity $\vec{\omega} = \nabla \times \vec{u}$ to rewrite the diffusion term as $\nabla \times \vec{\omega}$. Equation (1) becomes:

$$\begin{aligned} \partial_t \vec{u} + \nabla \cdot (\vec{u} \otimes \vec{u}) + \nu \nabla \times \vec{\omega} + \nabla p &= \vec{f} \quad \text{in } \Omega\,, \\ \nabla \times \vec{u} - \vec{\omega} &= 0 \quad \text{in } \Omega\,, \\ \nabla \cdot \vec{u} &= 0 \quad \text{in } \Omega\,. \end{aligned} \tag{2}$$

To discretize the system (2), we use the velocity at the faces, the vorticity on the edges and the pressure at the cells. First, from the velocity at the faces, we build a velocity at the elements, which we then convect with itself in order to build an approximation of the convection operator $\nabla \cdot (\vec{u} \otimes \vec{u})$ at the elements. This operator is then expressed in terms of the velocities at the faces and gives $C(\mathbf{u}_f^t)$ in a matrix form. Note that the convection operator can be chosen to be *upwind* or *centered*.

The resulting linear system can then be written as

$$\begin{pmatrix} \frac{M_{\mathbf{u}}}{\Delta t} + C(\mathbf{u}_f^t) & R & G \\ R^T & -\frac{1}{\nu} M_{\boldsymbol{\omega}} & 0 \\ G^T & 0 & 0 \end{pmatrix} \begin{pmatrix} \mathbf{u}_f^{t+\Delta t} \\ \nu \boldsymbol{\omega}_e^{t+\Delta t} \\ \mathbf{p}_c^{t+\Delta t} \end{pmatrix} = \begin{pmatrix} \frac{M_{\mathbf{u}}}{\Delta t} \mathbf{u}_f^t \\ 0 \\ 0 \end{pmatrix}, \tag{3}$$

where the unknowns are written in bold letters with a subscript indicating where the unknown is discretized (f at the faces, e on the edges and c at the cells) and with a superscript indicating the time step. Note that without the convection term $C(\mathbf{u}_f^t)$, the system matrix is symmetric. Furthermore, the Navier–Stokes equations are linearised at each time step by approximating the convective flow by the velocity solution at the previous time step, here $\mathbf{u}_f^t$.

2.2 PolyMAC II

Unfortunately, PolyMAC I presents a saddle point in the linear system (3). As a consequence, one must use a direct solver at each time step since the system matrix evolves during the simulation. PolyMAC I is therefore very costly in terms of memory and computational time. A second version of PolyMAC was then designed, with the aim of a simpler matrix in the top left block.

This was achieved by discretising Eq. (1) with the help of an auxiliary variable, here, the velocity vector at the cells. The momentum equation is first discretized on the auxiliary variables, i.e. at the cells and yields the second line in the matrix. Then, the convective and diffusive terms are interpolated from the cells to the faces and injected in the corresponding equations, which gives the first line of the matrix. The linear system takes the form

$$\begin{pmatrix} \frac{I}{\Delta t} & C_f(\mathbf{u}_e^t) + D_f(\mathbf{u}_e^t) & G_f^{MPFA} \\ 0 & \frac{I}{\Delta t} + C_c(\mathbf{u}_e^t) + D_c(\mathbf{u}_e^t) & G_c^{MPFA} \\ G^T & 0 & 0 \end{pmatrix} \begin{pmatrix} \mathbf{u}_f^{t+\Delta t} \\ \mathbf{u}_c^{t+\Delta t} \\ \mathbf{p}_c^{t+\Delta t} \end{pmatrix} = \begin{pmatrix} \frac{I}{\Delta t}\mathbf{u}_f^t \\ \frac{I}{\Delta t}\mathbf{u}_c^t \\ 0 \end{pmatrix}, \qquad (4)$$

and the objective is reached, as the top left block is now proportional to the identity: a SIMPLE-like method can be used to solve the saddle-point problem.[1]

For the equation at the cells, the convective term is computed as for PolyMAC I. For the diffusive term, the MPFA-O scheme [1] is used to discretize the gradient of the velocity, of which we take the discrete divergence. The gradient of the pressure is also discretized by a MPFA-O scheme.

Note that PolyMAC II is briefly described in [8].

2.3 PolyMAC III

PolyMAC II still presents difficulties: on some meshes, such as tetrahedral meshes, the MPFA-O scheme requires a large stencil and the numerical cost of the method becomes prohibitive. So, a third version of PolyMAC, PolyMAC III, was designed,

[1] Note that in the numerical experiments of this paper, we still use a direct solver, as the main goal is to compare the accuracy of each scheme.

which keeps a simpler top left block than PolyMAC I, while keeping the numerical cost under control.

PolyMAC III starts, as PolyMAC I, from the formulation of Eq. (2). The diffusive term is again computed *via* the vorticity. Here, the pressure gradient is computed by a mimetic method called HFV [6], which introduces auxiliary variables at the faces, in this instance, the pressure at the faces $\mathbf{p}_f$. The convection is computed as for PolyMAC I. The system matrix then has the following form

$$\begin{pmatrix} \frac{I}{\Delta t} + C(\mathbf{u}_f^t) & R & G_c^{HFV} & G_c^{HFV} \\ \tilde{R} & -\frac{1}{\nu} M_{\boldsymbol{\omega}} & 0 & 0 \\ G^T & 0 & 0 & 0 \\ 0 & 0 & P_c & P_f \end{pmatrix} \begin{pmatrix} \mathbf{u}_f^{t+\Delta t} \\ \boldsymbol{\omega}_e^{t+\Delta t} \\ \mathbf{p}_c^{t+\Delta t} \\ \mathbf{p}_f^{t+\Delta t} \end{pmatrix} = \begin{pmatrix} \frac{I}{\Delta t}\mathbf{u}_f^t \\ 0 \\ 0 \\ 0 \end{pmatrix} \tag{5}$$

where the last line in the matrix and the blocks P_i with $i \in \{c, f\}$ are defined by a continuity relation on the pressure gradient across the faces.

3 Test Problems

The different versions of PolyMAC are compared with the help of a series of $3D$ meshes defined in benchmarks established during previous cycles of the FVCA conferences [4, 7]. Those meshes are listed in Table 1 and some of them are illustrated on Fig. 1. One of the objective of this work is to characterize the convergence properties of PolyMAC and so, for each mesh, we consider several sizes.

Table 1 Numerical results obtained for the different problems with all three PolyMAC methods: from left to right, the order of convergence in velocity, the order of convergence in pressure and the sparsity ratio of the linear system

PolyMAC	Order convergence						Sparsity ratio		
	Velocity			Pressure					
	I	II	III	I	II	III	I	II	III
Hexa	1.08	**1.11**	**1.11**	**1.20**	**1.20**	**1.20**	0.06	0.08	**0.04**
Locally Refined (3D)	1.53	**1.71**	1.53	**1.26**	**1.26**	**1.26**	0.08	0.07	**0.04**
Kershaw (3D)	**2.67**	0.63	0.78	**3.45**	0.72	0.75	0.01	0.01	**0.00**
CheckerBoard	0.84	**1.23**	0.84	0.96	**1.26**	1.14	0.08	0.06	**0.04**
Voronoi	0.36	**1.02**	0.90	−1.77	**1.74**	1.68	0.16	0.13	**0.11**
Tetrahedra	**1.23**	/	1.20	**0.87**	/	0.78	0.02	/	**0.01**
Random	**1.05**	/	**1.05**	0.72	/	**0.90**	0.05	/	**0.03**

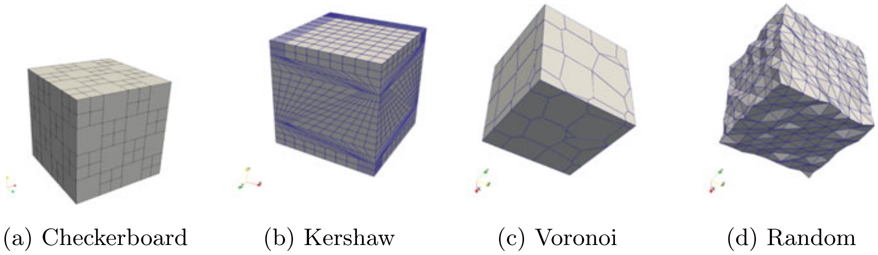

(a) Checkerboard (b) Kershaw (c) Voronoi (d) Random

Fig. 1 Representation of some meshes from the benchmark

We consider Eq. (1) and the Navier–Stokes problem described in [2] which models a rotating flow whose exact solution is given by

$$\begin{aligned}\vec{u} &= \begin{pmatrix} y-z \\ z-x \\ x-y \end{pmatrix} \\ p &= (x^2+y^2+z^2) - xy - xz - yz - \tfrac{1}{4}\,,\end{aligned} \tag{6}$$

and the right-hand-side is computed with the exact solution. The viscosity is equal to $\nu = 10^{-2}$ and the Reynolds is equal to 100. The numerical results are presented in the next section.

4 Numerical Results

In this section, we compare the performances of the three PolyMAC schemes on the problems described in Sect. 3. The discrete systems (3), (4) and (5) are solved with a direct solver, here, a LU-factorisation.

First, we evaluate the accuracy of each scheme. To do so, we draw the error in velocity and pressure with respect to the number of velocity and pressure elements respectively. These results are shown on Fig. 2. For those results, we estimate the orders of convergence for each scheme, as presented in Table 1. Second, we consider the sparsity of the linear systems resulting from each discretisation. To do so, we compute the sparsity ratio of the linear matrix associated to each problem also in Table 1. This ratio is equal to the number of non-zero elements (nnz) divided by the total number of elements in the matrix (n).

Looking at the velocity convergence on the left panel of Fig. 2, we observe that all schemes behave very similarly on the `Hexa` mesh. This is due to the fact that all schemes reduce to very similar discrete problems on such a mesh. Otherwise, conclusions are not so easily drawn. First, PolyMAC I seems more efficient on `Kershaw` (3D) meshes. For the `Voronoi` mesh, PolyMAC III seems the most accurate scheme, while PolyMAC II works better on the `CheckerBoard` and `Locally Refined`

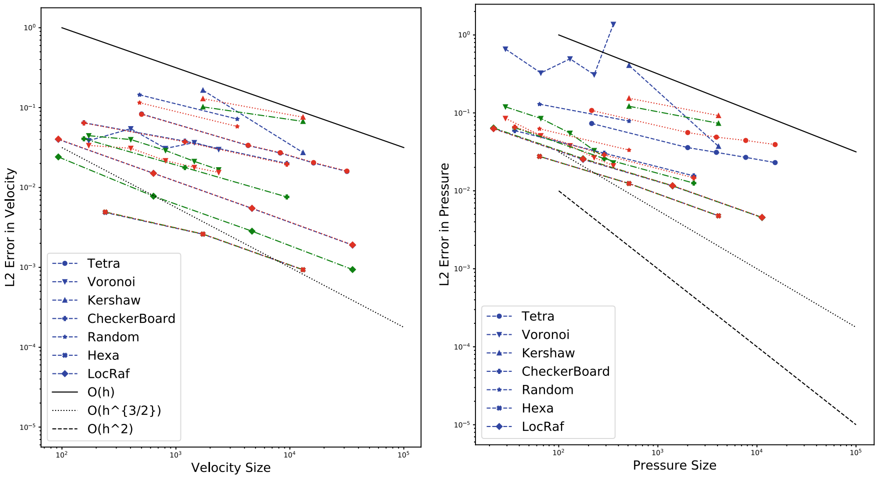


Fig. 2 Convergence curves of the different PolyMAC versions on 3D meshes—left: velocity convergence; right: pressure convergence. PolyMAC I is represented in blue, PolyMAC II in green and PolyMAC III in red. The different meshes are represented by different symbols detailed in the caption of each image. In black, we show different orders of convergence

(3D) meshes. On the other hand, PolyMAC I and PolyMAC III show similar performance on the `Tetrahedra` and `Random` meshes, for which we could not get any results with PolyMAC II because of memory restrictions.

However, the convergence orders in Table 1 paint a clearer picture: for each mesh, the highest order of convergence is highlighted in bold. We see clearly that for all meshes where there were no memory problem—except for the `Kershaw` (3D) meshes—PolyMAC II is the better choice.

Let us consider briefly the convergence in pressure, represented on the right panel of Fig. 2 and the estimation of the order of convergence is represented in Table 1. The order of convergence is generally around 1.0.[2] Moreover, the variation between schemes seems to be less pronounced than in the velocity case. Globally, PolyMAC II behaves more consistently than the other two versions, but the difference is small.

To get more insight into the relative complexity of each method as well as their memory requirements, we finally computed the sparsity ratio of the resulting linear systems. The latter is computed as the number of non-zero elements nnz divided by the total number of elements available in the matrix n^2 where n is the dimension of the matrix (multiplied by 100 to express the sparsity ratio as a percentage). This indicates how sparse the resulting matrix is. The sparser the matrix the lighter the memory requirements and the more efficient the solving stage will be. Results are

[2] Except for the PolyMAC I scheme on the `Voronoi` and `Kershaw` (3D) meshes. Note however that in the `Kershaw` (3D) case, we only have two data points and the value of the order of convergence needs to be considered carefully.

shown in the last three columns of Table 1. The first obvious observation is that the third scheme PolyMAC III yields the sparsest linear systems and will then have the most cost-efficient solving step.

Regarding PolyMAC I and PolyMAC II schemes, the situation is a little bit more complicated. On most cases, the density of the linear system is similar between the two. On the `Tetrahedra` and `Random` meshes the sparsity of the matrix explodes, which leads to too large memory requirements. This is due to the fact that the PolyMAC II method uses a larger stencil to approximate the unknowns, which leads to denser matrices on tetreahedra. However, for the `Voronoi` mesh, PolyMAC I leads to much denser matrices.

Those conclusions are globally confirmed by numerical tests we did on 2D meshes and which are not included in this paper to keep it short, but will be part of a future paper. We observed two main differences with the conclusions drawn for the 3D cases. On one hand, PolyMAC II was consistently the most accurate scheme with the highest order of convergence, except on a 2D version of the Kershaw mesh. On the other hand, PolyMAC II also showed significantly more sparsity than the other two schemes on 2D cases.

5 Conclusions

In this paper, we investigate the properties of three different schemes developed by the CEA to ultimately model multiphase compressible flows. The idea is to generalize the MAC scheme to general meshes in a finite volume framework, hence the name PolyMAC. We describe briefly how the three variants are built and we introduce the test problems and the different meshes we use to evaluate their performances.

We observe that PolyMAC II is globally the most accurate scheme, but it may suffer from a higher numerical cost due to a higher density of the linear system matrix. This has caused the numerical cost to explode on the `Tetrahedra` and `Random` meshes. Generally, we would not recommand to use PolyMAC II if the mesh is made up of tetrahedra. Concerning, the two other schemes, except on `Kershaw` (3D) mesh, where PolyMAC I is significantly better, PolyMAC I and PolyMAC III present a similar accuracy. This can be related to the fact that both PolyMAC I and PolyMAC III share the same diffusion scheme. PolyMAC I being numerically more costly, one would prefer PolyMAC III. Indeed, this latter scheme provides good results while being numerically viable.

Preliminary results have nonetheless shown that on complicated industrial meshes with very deformed cells, both PolyMAC II and PolyMAC III could be unstable, while PolyMAC I stayed robust. For these problematic cases, the values of the velocity diverged because of the convective term. There were no indication in the numerical experiments reported here to anticipate such behaviour. This could motivate the inclusion of more realistic test cases in academic benchmarks. To conclude, we recommend the use of PolyMAC III as default scheme and to switch to PolyMAC II if one requires more accuracy. If both PolyMAC III and PolyMAC II fail, PolyMAC I can offer a reliable alternative, even if it can be numerically expensive to solve the resulting linear system.

References

1. Aavatsmark, I.: An introduction to multipoint flux approximations for quadrilateral grids. Comput. Geosci. **6**, 405–432 (2002)
2. Angeli, P.E., Puscas, M.A., Fauchet, G., Cartalade, A.: FVCA8 benchmark for the Stokes and Navier-Stokes equations with the TrioCFD code-benchmark session. In: Cancès, C., Omnes, P. (eds.) Finite Volumes for Complex Applications VIII - Methods and Theoretical Aspects, pp. 181–202. Springer International Publishing, Cham (2017)
3. Beltman, R., Anthonissen, M., Koren, B.: Conservative polytopal mimetic discretization of the incompressible Navier-Stokes equations. J. Comput. Appl. Math. **340**, 443–473 (2018). https://doi.org/10.1016/j.cam.2018.02.007
4. Boyer, F., Omnes, P.: Benchmark proposal for the FVCA8 conference: finite volume methods for the Stokes and Navier-Stokes equations. In: Cancès, C., Omnes, P. (eds.) Finite Volumes for Complex Applications VIII - Methods and Theoretical Aspects, pp. 59–71. Springer International Publishing, Cham (2017)
5. CEA: TRUST platform (2022). https://github.com/cea-trust-platform
6. Droniou, J., Eymard, R., Gallouët, T., Guichard, C., Herbin, R.: The Gradient Discretisation Method. Springer International Pulishing AG (2018). https://link.springer.com/book/10.1007/978-3-319-79042-8
7. Eymard, R., Henry, G., Herbin, R., Hubert, F., Klöfkorn, R., Manzini, G.: 3D benchmark on discretization schemes for anisotropic diffusion problems on general grids. In: Fořt, J., Fürst, J., Halama, J., Herbin, R., Hubert, F. (eds.) Finite Volumes for Complex Applications VI Problems & Perspectives, pp. 895–930. Springer, Berlin (2011)
8. Gerschenfeld, A., Gorsse, Y.: Development of a robust multiphase flow solver on general meshes; Application to sodium boiling at the subchannel scale (Paper 36282). In: 19th International Topical Meeting on Nuclear Reactor Thermal Hydraulics (NURETH-19) (2022)
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

Finite Volume Approximations for Non-linear Parabolic Problems with Stochastic Forcing

Caroline Bauzet, Flore Nabet, Kerstin Schmitz, and Aleksandra Zimmermann

Abstract We propose a two-point flux approximation finite-volume scheme for a stochastic non-linear parabolic equation with a multiplicative noise. The time discretization is implicit except for the stochastic noise term in order to be compatible with stochastic integration in the sense of Itô. We show existence and uniqueness of solutions to the scheme and the appropriate measurability for stochastic integration follows from the uniqueness of approximate solutions.

Keywords Stochastic non-linear parabolic equation · Multiplicative Lipschitz noise · Finite-volume method · Upwind scheme · Diffusion-convection equation · Variational approach

1 Introduction

Let Λ be a bounded, open, connected and polygonal set of $\mathbb{R}^d$ with $1 \leq d \leq 3$. Moreover let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space endowed with a right-continuous, complete filtration $(\mathcal{F}_t)_{t\geq 0}$ and let $(W(t))_{t\geq 0}$ be a standard, one-dimensional Brownian motion with respect to $(\mathcal{F}_t)_{t\geq 0}$ on $(\Omega, \mathcal{A}, \mathbb{P})$.

C. Bauzet
LMA UMR 7031, CNRS, Centrale Marseille, Aix Marseille University, Marseille, France
e-mail: caroline.bauzet@univ-amu.fr

F. Nabet
CMAP, CNRS, Institut Polytechnique de Paris, École Polytechnique, 91120 Palaiseau, France
e-mail: flore.nabet@polytechnique.edu

K. Schmitz (✉)
Fakultät für Mathematik, Universität Duisburg-Essen, Essen, Germany
e-mail: kerstin.schmitz@uni-due.de

A. Zimmermann
Institut für Mathematik, TU Clausthal, Clausthal-Zellerfeld, Germany
e-mail: aleksandra.zimmermann@tu-clausthal.de

E. Franck et al. (eds.), *Finite Volumes for Complex Applications X—Volume 1, Elliptic and Parabolic Problems*, Springer Proceedings in Mathematics & Statistics 432,
https://doi.org/10.1007/978-3-031-40864-9_10

For $T > 0$, we consider the following non-linear parabolic problem forced by a multiplicative stochastic noise:

$$\begin{aligned} du - \Delta u\, dt + \operatorname{div}\big(\mathbf{v} f(u)\big)\, dt &= g(u)\, dW(t) + \beta(u)\, dt, && \text{in } \Omega \times (0, T) \times \Lambda; \\ u(0, .) &= u_0, && \text{in } \Omega \times \Lambda; \\ \nabla u \cdot \mathbf{n} &= 0, && \text{on } \Omega \times (0, T) \times \partial\Lambda; \end{aligned} \tag{1}$$

where div is the divergence operator with respect to the space variable and $\mathbf{n}$ denotes the unit normal vector to $\partial\Lambda$ outward to Λ. After setting L_f, L_β and L_g in $\mathbb{R}_+^*$, we assume the following hypotheses on the data:

H_1: $u_0 \in L^2(\Omega; H^1(\Lambda))$ is $\mathcal{F}_0$-measurable.
H_2: $f : \mathbb{R} \to \mathbb{R}$ is nondecreasing, L_f-Lipschitz continuous with $f(0) = 0$.
H_3: $g : \mathbb{R} \to \mathbb{R}$ is a L_g-Lipschitz continuous function.
H_4: $\beta : \mathbb{R} \to \mathbb{R}$ is L_β-Lipschitz continuous with $\beta(0) = 0$.
H_5: $\mathbf{v} \in \mathscr{C}^1([0, T] \times \overline{\Lambda}; \mathbb{R}^d)$ such that $\operatorname{div}(\mathbf{v}) = 0$ in $[0, T] \times \Lambda$ and $\mathbf{v} \cdot \mathbf{n} = 0$ on $[0, T] \times \partial\Lambda$.

Remark 1 Note that for recent results on viscous stochastic conservation laws in dimension 1, we also refer to [4].

1.1 Concept of Solution

We will be interested in the concept of solution as defined below, which we will call a variational solution:

Definition 1 A predictable stochastic process u is a variational solution to Problem (1) if it belongs to

$$L^2(\Omega; \mathscr{C}([0, T]; L^2(\Lambda))) \cap L^2(\Omega; L^2(0, T; H^1(\Lambda)))$$

and satisfies, for all $t \in [0, T]$, in $L^2(\Lambda)$, and $\mathbb{P}$-a.s. in Ω

$$\begin{aligned} & u(t) - u_0 - \int_0^t \Delta u(s)\, ds + \int_0^t \operatorname{div}\big(\mathbf{v}(s, .) f(u(s))\big)\, ds \\ & = \int_0^t g(u(s))\, dW(s) + \int_0^t \beta(u(s))\, ds. \end{aligned}$$

Existence, uniqueness and regularity of this variational solution is well-known in the literature, see, e.g., [2].

1.2 Outline

In this contribution, we propose a finite-volume approximation scheme for the solution of (1) in the sense of Definition 1. We show existence and uniqueness of solutions to the scheme. In Sect. 2, we introduce the notation for our finite-volume framework. In Sect. 3, we introduce our finite-volume scheme. The main result is contained in Sect. 4.

2 Admissible Finite-Volume Meshes and Notations

In order to perform a finite-volume approximation of the variational solution of Problem (1) on $[0, T] \times \Lambda$ we need first of all to set a choice for the temporal and spatial discretization. For the time-discretization, let $N \in \mathbb{N}^*$ be given. We define the fixed time step $\Delta t = \frac{T}{N}$ and divide the interval $[0, T]$ in $0 = t_0 < t_1 < \cdots < t_N = T$ equidistantly with $t_n = n\Delta t$ for all $n \in \{0, \dots, N\}$. For the space discretization, although we use the two-dimensional vocabulary, e.g., polygonal, edge, etc., what we present is also valid in space dimension $1 \leq d \leq 3$. We refer to [1] for a more general definition of finite-volume admissible meshes in dimension $1 \leq d \leq 3$.

Definition 2 (*Admissible finite-volume mesh*) An admissible finite-volume mesh $\mathcal{T}$ of Λ (see Fig. 1) is given by a family of open polygonal and convex subsets K, called *control volumes* of $\mathcal{T}$, satisfying the following properties:

- $\overline{\Lambda} = \bigcup_{K \in \mathcal{T}} \overline{K}$.
- If $K, L \in \mathcal{T}$ with $K \neq L$ then $\operatorname{int} K \cap \operatorname{int} L = \emptyset$.
- If $K, L \in \mathcal{T}$, with $K \neq L$ then either the $(d-1)$-dimensional Lebesgue measure of $\overline{K} \cap \overline{L}$ is 0 or $\overline{K} \cap \overline{L}$ is the edge if $d = 2$ (or the face if $d = 3$), denoted by $\sigma = K|L$, separating the control volumes K and L.
- To each control volume $K \in \mathcal{T}$, we associate a point $x_K \in \overline{K}$ (called the center of K) such that: If $K, L \in \mathcal{T}$ are two neighbouring control volumes the straight line between the centers x_K and x_L is orthogonal to $\sigma = K|L$.

Once an admissible finite-volume mesh $\mathcal{T}$ of Λ is fixed, we will use the following notations.

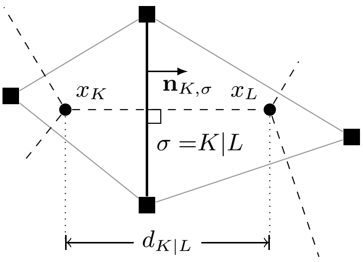


Fig. 1 Notations of the mesh $\mathcal{T}$ associated with $\Lambda \subseteq \mathbb{R}^2$

2.1 Notation

- $h = \text{size}(\mathcal{T}) = \sup\{\text{diam}(K) : K \in \mathcal{T}\}$, the mesh size.
- $d_h \in \mathbb{N}$ the number of control volumes $K \in \mathcal{T}$ with $h = \text{size}(\mathcal{T})$.
- $\mathcal{E}_{\text{int}} := \{\sigma : \sigma \nsubseteq \partial\Lambda\}$ is the set of interior edges (or faces) of the mesh $\mathcal{T}$.
- For $K \in \mathcal{T}$, $\mathcal{E}_K$ is the set of edges (or faces) of K, $\mathcal{E}_{K,\text{int}} = \mathcal{E}_K \cap \mathcal{E}_{\text{int}}$ and m_K is the d-dimensional Lebesgue measure of K.
- For $K \in \mathcal{T}$ and $\sigma \in \mathcal{E}_K$, $\mathbf{n}_{K,\sigma}$ is the unit normal vector to σ outward to K.
- Let $K, L \in \mathcal{T}$ be two neighbouring control volumes. For $\sigma = K|L \in \mathcal{E}_{\text{int}}$, let m_σ be the $(d-1)$-dimensional Lebesgue measure of σ and $d_{K|L}$ the distance between x_K and x_L.
- For any vector $u_h = (u_K)_{K\in\mathcal{T}} \in \mathbb{R}^{d_h}$, we define the L^2-norm on Λ by

$$\|u_h\|_{L^2(\Lambda)} = \left(\sum_{K\in\mathcal{T}} m_K|u_K|^2\right)^{\frac{1}{2}}.$$

In the sequel, we note $|x|$ the euclidean norm of $x \in \mathbb{R}^d$ with $d \geq 1$.

3 The Finite-Volume Scheme

Firstly, we define the vector $u_h^0 = (u_K^0)_{K\in\mathcal{T}} \in \mathbb{R}^{d_h}$ by the discretization of the initial condition u_0 of Problem (1) over each control volume:

$$u_K^0 := \frac{1}{m_K}\int_K u_0(x)\,dx, \quad \forall K \in \mathcal{T}. \tag{2}$$

The finite-volume scheme we propose reads, for this given initial $\mathcal{F}_0$-measurable random vector $u_h^0 \in \mathbb{R}^{d_h}$: For any $n \in \{0, \dots, N-1\}$, knowing $u_h^n = (u_K^n)_{K\in\mathcal{T}} \in \mathbb{R}^{d_h}$ we search for $u_h^{n+1} = (u_K^{n+1})_{K\in\mathcal{T}} \in \mathbb{R}^{d_h}$ such that, for almost every $\omega \in \Omega$, the vector u_h^{n+1} is solution to the following random equations

$$\begin{aligned}
&\frac{m_K}{\Delta t}(u_K^{n+1} - u_K^n) + \sum_{\sigma=K|L\in\mathcal{E}_{K,\text{int}}} m_\sigma v_{K,\sigma}^{n+1} f(u_\sigma^{n+1}) \\
&\quad + \sum_{\sigma=K|L\in\mathcal{E}_{K,\text{int}}} \frac{m_\sigma}{d_{K|L}}(u_K^{n+1} - u_L^{n+1}) \\
&= \frac{m_K}{\Delta t} g(u_K^n)(W^{n+1} - W^n) + m_K\beta(u_K^{n+1}), \quad \forall K \in \mathcal{T},
\end{aligned} \tag{3}$$

where, by denoting γ the $(d-1)$-dimensional Lebesgue measure,

$$v_{K,\sigma}^{n+1} = \frac{1}{\Delta t m_\sigma} \int_{t_n}^{t_{n+1}} \int_\sigma \mathbf{v}(t,x) \cdot \mathbf{n}_{K,\sigma}\, d\gamma(x) dt,$$

and u_σ^{n+1} denotes the upstream value at time t_{n+1} with respect to σ defined as follows: If $\sigma = K|L \in \mathcal{E}_{K,\text{int}}$ is the interface between the control volumes K and L, u_σ^{n+1} is equal to u_K^{n+1} if $v_{K,\sigma}^{n+1} \geqslant 0$ and to u_L^{n+1} if $v_{K,\sigma}^{n+1} < 0$. Note also that $W^{n+1} - W^n = W(t_{n+1}) - W(t_n)$ for $n \in \{0, \dots, N-1\}$.

Remark 2 Since $\operatorname{div}(\mathbf{v}) = 0$ in $[0,T] \times \Lambda$, for any $n \in \{0, \cdots, N-1\}$ and $K \in \mathcal{T}$ one has $\sum_{\sigma=K|L\in\mathcal{E}_{K,\text{int}}} m_\sigma v_{K,\sigma}^{n+1} = 0$. Thus, using that $v_{K,\sigma}^{n+1} = (v_{K,\sigma}^{n+1})^+ - (v_{K,\sigma}^{n+1})^-$ (where, for $r \in \mathbb{R}$, $r^+ := \max\{r, 0\}$ and $r^- := -\min\{0, r\}$) an equivalent formulation of the scheme (3) is given by

$$\begin{aligned} &\frac{m_K}{\Delta t}(u_K^{n+1} - u_K^n) + \sum_{\sigma=K|L\in\mathcal{E}_{K,\text{int}}} m_\sigma (v_{K,\sigma}^{n+1})^- \Big(f(u_K^{n+1}) - f(u_L^{n+1})\Big) \\ &\quad + \sum_{\sigma=K|L\in\mathcal{E}_{K,\text{int}}} \frac{m_\sigma}{d_{K|L}}(u_K^{n+1} - u_L^{n+1}) \\ &= \frac{m_K}{\Delta t} g(u_K^n)\left(W^{n+1} - W^n\right) + m_K \beta(u_K^{n+1}), \quad \forall K \in \mathcal{T}. \end{aligned} \tag{4}$$

4 Main Result

Proposition 1 (Existence of a discrete solution) *Assume that hypotheses H_1 to H_5 hold. Let $\mathcal{T}$ be an admissible finite-volume mesh of Λ in the sense of Definition 2 with a mesh size h and $N \in \mathbb{N}^*$. Then, there exists a unique solution $(u_h^n)_{1\leq n\leq N} \in (\mathbb{R}^{d_h})^N$ to Problem (3) associated with the initial vector u_h^0 defined by (2). Additionally, for any $n \in \{0, \dots, N\}$, u_h^n is a $\mathcal{F}_{t_n}$-measurable random vector.*

Proof We fix $n \in \{0, \dots, N-1\}$ and choose an arbitrary vector $u_h^n = (u_K^n)_{K\in\mathcal{T}} \in \mathbb{R}^{d_h}$. Firstly, we will show that there exists at least one random vector $u_h^{n+1} = (u_K^{n+1})_{K\in\mathcal{T}} \in \mathbb{R}^{d_h}$ such that (4) holds true $\mathbb{P}$-a.s in Ω. To this end, we define the mapping $\mathbf{P}^n : \mathbb{R}^{d_h} \to \mathbb{R}^{d_h}$, $\mathbf{P}^n = (P_1^n, \dots, P_{d_h}^n)$ such that for any $i \in \{1, \dots, d_h\}$

$$\begin{aligned} P_i^n(w_{K_1}, \dots, w_{K_{d_h}}) = &\frac{m_{K_i}}{\Delta t} w_{K_i} - m_{K_i}\beta(w_{K_i}) \\ &+ \sum_{\sigma=K_i|K_j\in\mathcal{E}_{K_i,\text{int}}} m_\sigma (v_{K_i,\sigma}^{n+1})^-(f(w_{K_i}) - f(w_{K_j})) \\ &+ \sum_{\sigma=K_i|K_j\in\mathcal{E}_{K_i,\text{int}}} \frac{m_\sigma}{d_{K_i|K_j}}(w_{K_i} - w_{K_j}) - \frac{m_{K_i}}{\Delta t}\xi_i^n \end{aligned}$$

where $\xi_i^n := u_{K_i}^n + g(u_{K_i}^n)(W^{n+1} - W^n)$. Obviously, $\mathbf{P}^n$ is a continuous mapping. Next, we show that there exists $\varrho > 0$ such that for all $w_h = (w_{K_i})_{1\leq i\leq d_h} \in \mathbb{R}^{d_h}$ such that $|w_h| = \varrho$,

$$(\mathbf{P}^n(w_h), w_h)_{\mathbb{R}^{d_h}} := \sum_{i=1}^{d_h} P_i^n(w_h)w_{K_i} \geq 0.$$

In this case, from [3, Lemma 4.3] it follows that there exists at least one $\overline{w}_h \in \mathbb{R}^{d_h}$ such that $|\overline{w}_h| \leq \varrho$ and $\mathbf{P}^n(\overline{w}_h) = 0$. We have

$$\begin{aligned}\sum_{i=1}^{d_h} P_i^n(w_h)w_{K_i} &= \sum_{i=1}^{d_h} \frac{m_{K_i}}{\Delta t} w_{K_i}^2 - \sum_{i=1}^{d_h} m_{K_i}\beta(w_{K_i})w_{K_i} - \sum_{i=1}^{d_h} \frac{m_{K_i}}{\Delta t}\xi_i^n w_{K_i}\\ &\quad+ \sum_{i=1}^{d_h} \sum_{\sigma=K_i|K_j\in\mathcal{E}_{K_i,\text{int}}} m_\sigma(v_{K_i,\sigma}^{n+1})^-(f(w_{K_i}) - f(w_{K_j}))w_{K_i}\\ &\quad+ \sum_{i=1}^{d_h} \sum_{\sigma=K_i|K_j\in\mathcal{E}_{K_i,\text{int}}} \frac{m_\sigma}{d_{K_i|K_j}}(w_{K_i} - w_{K_j})w_{K_i}\\ &=: I_1 + I_2 + I_3 + I_4 + I_5.\end{aligned}$$

Since β is Lipschitz continuous, the term I_2 satisfies

$$I_2 \geq -L_\beta\|w_h\|_{L^2(\Lambda)}^2. \tag{5}$$

Moreover, by discrete partial integration,

$$I_5 = \sum_{\sigma=K_i|K_j\in\mathcal{E}_{K_i,\text{int}}} \frac{m_\sigma}{d_{K_i|K_j}}|w_{K_i} - w_{K_j}|^2 \geq 0. \tag{6}$$

Now, we focus on the term I_4. Since f is Lipschitz continuous and nondecreasing, thanks to [1, Lemma 18.5], for any $r \in \mathbb{R}$, using the notation $\Phi(r) = \int_0^r f'(s)s\,ds$, for any $a, b \in \mathbb{R}$, one has

$$\begin{aligned}b(f(b) - f(a)) &= \int_a^b (sf(s))'ds - (b-a)f(a) = \int_a^b \Phi'(s)ds + \int_a^b (f(s) - f(a))ds\\ &\geq (\Phi(b) - \Phi(a)) + \frac{1}{2L_f}(f(b) - f(a))^2.\end{aligned}$$

Thus, since $\operatorname{div}\mathbf{v} = 0$ in $[0, T] \times \Lambda$ and $v_{K_i,\sigma}^{n+1} = -v_{K_j,\sigma}^{n+1}$, we obtain

$$\begin{aligned} I_4 &\geq \sum_{i=1}^{d_h} \sum_{\sigma=K_i|K_j\in\mathcal{E}_{K_i,\mathrm{int}}} m_\sigma (v_{K_i,\sigma}^{n+1})^- (\Phi(w_{K_i}) - \Phi(w_{K_j})) \\ &= \sum_{i=1}^{d_h} \Phi(w_{K_i}) \sum_{\sigma=K_i|K_j\in\mathcal{E}_{K_i,\mathrm{int}}} m_\sigma (v_{K_i,\sigma}^{n+1}) = 0. \end{aligned} \tag{7}$$

For the term I_3, since $-ab \geq -\frac{1}{2}(a^2+b^2)$ one has

$$I_3 \geq -\frac{1}{2\Delta t}\left(\|w_h\|^2_{L^2(\Lambda)} + \|\xi_h^n\|^2_{L^2(\Lambda)}\right). \tag{8}$$

From (5), (6), (7) and (8) for some $\alpha > 0$, choosing $\Delta t \leq \frac{1}{2(\alpha+L_\beta)}$ we now get

$$\begin{aligned} \sum_{i=1}^{d_h} P_i^n(w_h)w_{K_i} &\geq \frac{1}{2\Delta t}\|w_h\|^2_{L^2(\Lambda)} - L_\beta\|w_h\|^2_{L^2(\Lambda)} - \frac{1}{2\Delta t}\|\xi_h^n\|^2_{L^2(\Lambda)} \\ &\geq \alpha(\min_{K\in\mathcal{T}} m_K)|w_h|^2 - \frac{1}{2\Delta t}\|\xi_h^n\|^2_{L^2(\Lambda)}. \end{aligned} \tag{9}$$

Then, setting

$$\varrho := \sqrt{\frac{1}{2\alpha(\min_{K\in\mathcal{T}} m_K)\Delta t}}\|\xi_h^n\|_{L^2(\Lambda)} > 0$$

we get $(\mathbf{P}^n(w_h), w_h)_{\mathbb{R}^{d_h}} \geq 0$ from (9) for all $w_h \in \mathbb{R}^{d_h}$ such that $|w_h| = \varrho$. Hence, there exists at least one element $\overline{w}_h$ such that $\mathbf{P}^n(\overline{w}_h) = 0$. Thus, $u_h^{n+1} := \overline{w}_h \in \mathbb{R}^{d_h}$ is solution to the numerical scheme (4).

Next, we will prove the uniqueness of the solution. Therefore, we assume that there exist $w_h = (w_{K_i})_{1\leq i\leq d_h} \in \mathbb{R}^{d_h}$ and $z_h = (z_{K_i})_{1\leq i\leq d_h} \in \mathbb{R}^{d_h}$ satisfying $\mathbf{P}^n(w_h) = \mathbf{P}^n(z_h) = 0$. Taking $P_i^n(w_h) - P_i^n(z_h)$, and using the initial formulation of the scheme (3), for any $i = 1, \ldots, d_h$ we obtain

$$\begin{aligned} &\frac{m_{K_i}}{\Delta t}(w_{K_i} - z_{K_i}) - m_{K_i}(\beta(w_{K_i}) - \beta(z_{K_i})) \\ &\quad + \sum_{\sigma=K_i|K_j\in\mathcal{E}_{K_i,\mathrm{int}}} m_\sigma v_{K_i,\sigma}^{n+1}(f(w_\sigma) - f(z_\sigma)) \\ &\quad + \sum_{\sigma=K_i|K_j\in\mathcal{E}_{K_i,\mathrm{int}}} \frac{m_\sigma}{d_{K_i|K_j}}\left((w_{K_i} - w_{K_j}) - (z_{K_i} - z_{K_j})\right) = 0, \end{aligned}$$

where w_σ and z_σ are the upstream value with respect to σ.

Now, we adjust the method developed in the proof of [1, Proposition 26.1]: Using the monotonicity of f, the fact that $v_{K_i,\sigma}^{n+1} = (v_{K_i,\sigma}^{n+1})^+ - (v_{K_i,\sigma}^{n+1})^-$ and taking the absolute value, one has

$$\begin{aligned}
&\frac{m_{K_i}}{\Delta t}|w_{K_i}-z_{K_i}|+\sum_{\sigma=K_i|K_j\in\mathcal{E}_{K_i,\text{int}}}\frac{m_\sigma}{d_{K_i|K_j}}|w_{K_i}-z_{K_i}|\\
&\quad+\sum_{\sigma=K_i|K_j\in\mathcal{E}_{K_i,\text{int}}}m_\sigma(v^{n+1}_{K_i,\sigma})^+|f(w_{K_i})-f(z_{K_i})|\\
&\le m_{K_i}|\beta(w_{K_i})-\beta(z_{K_i})|+\sum_{\sigma=K_i|K_j\in\mathcal{E}_{K_i,\text{int}}}\frac{m_\sigma}{d_{K_i|K_j}}|w_{K_j}-z_{K_j}|\\
&\quad+\sum_{\sigma=K_i|K_j\in\mathcal{E}_{K_i,\text{int}}}m_\sigma(v^{n+1}_{K_i,\sigma})^-|f(w_{K_j})-f(z_{K_j})|.
\end{aligned}\tag{10}$$

For $\eta>0$, $x\in\mathbb{R}^d$ we define $\varphi(x)=\exp(-\eta|x|)$ and for $K_i\in\mathcal{T}$, $i=1,\dots,d_h$ let

$$\varphi_{K_i}:=\frac{1}{m_{K_i}}\int_{K_i}\varphi(x)\,dx.$$

Multiplying (10) by φ_{K_i}, taking the sum over $i=1,\dots,d_h$ and rearranging the sums on the right-hand side by fixing j and varying over i we obtain

$$\begin{aligned}
&\sum_{i=1}^{d_h}\frac{m_{K_i}}{\Delta t}\varphi_{K_i}|w_{K_i}-z_{K_i}|+\sum_{i=1}^{d_h}\varphi_{K_i}\sum_{\sigma=K_i|K_j\in\mathcal{E}_{K_i,\text{int}}}\frac{m_\sigma}{d_{K_i|K_j}}|w_{K_i}-z_{K_i}|\\
&+\sum_{i=1}^{d_h}\varphi_{K_i}\sum_{\sigma=K_i|K_j\in\mathcal{E}_{K_i,\text{int}}}m_\sigma(v^{n+1}_{K_i,\sigma})^+|f(w_{K_i})-f(z_{K_i})|\le I_1+I_2+I_3
\end{aligned}\tag{11}$$

where

$$\begin{aligned}
I_2\le&\sum_{i=1}^{d_h}|w_{K_i}-z_{K_i}|\sum_{\sigma=K_i|K_j\in\mathcal{E}_{K_i,\text{int}}}\frac{m_\sigma}{d_{K_i|K_j}}|\varphi_{K_i}-\varphi_{K_j}|\\
&+\sum_{i=1}^{d_h}|w_{K_i}-z_{K_i}|\sum_{\sigma=K_i|K_j\in\mathcal{E}_{K_i,\text{int}}}\frac{m_\sigma}{d_{K_i|K_j}}\varphi_{K_i}
\end{aligned}\tag{12}$$

and similarly, since $(v^{n+1}_{K_j,\sigma})^-=(-v^{n+1}_{K_i,\sigma})^-=(v^{n+1}_{K_i,\sigma})^+$ for $\sigma=K_i|K_j$, using the Lipschitz continuity of f

$$\begin{aligned}
I_3\le&\sum_{i=1}^{d_h}L_f|w_{K_i}-z_{K_i}|\sum_{\sigma=K_i|K_j\in\mathcal{E}_{K_i,\text{int}}}m_\sigma(v^{n+1}_{K_j,\sigma})^-|\varphi_{K_i}-\varphi_{K_j}|\\
&+\sum_{i=1}^{d_h}|f(w_{K_i})-f(z_{K_i})|\sum_{\sigma=K_i|K_j\in\mathcal{E}_{K_i,\text{int}}}m_\sigma(v^{n+1}_{K_i,\sigma})^+\varphi_{K_i}.
\end{aligned}\tag{13}$$

Now, plugging (12) and (13) into (11) and using the Lipschitz continuity of β we obtain for all $i = 1, \dots, d_h$

$$\sum_{i=1}^{d_h} a_i |w_{K_i} - z_{K_i}| \le \sum_{i=1}^{d_h} b_i |w_{K_i} - z_{K_i}| \tag{14}$$

with

$$a_i := m_{K_i} \left(\frac{1}{\Delta t} - L_\beta \right) \varphi_{K_i}$$
$$b_i := \sum_{\sigma = K_i|K_j \in \mathcal{E}_{K_i,\mathrm{int}}} \left(\frac{m_\sigma}{d_{K_i|K_j}} + m_\sigma (v_{K_j,\sigma}^{n+1})^- \right) |\varphi_{K_i} - \varphi_{K_j}|.$$

Now, taking $\Delta t \le \frac{1}{2L_\beta}$ using the same arguments as in the proof of [1, Proposition 26.1], we may choose $\eta > 0$ small enough such that $a_i > b_i$ for all $i = 1, \dots d_h$. Thus $w_{K_i} = z_{K_i}$ then follows from (14) for all $i = 1, \dots, d_h$. Since the initial vector u_h^0 is given, the existence of a unique solution $(u_h^n)_{1 \le n \le N} \in \mathbb{R}^{d_h}$ follows by iteration. It is left to prove that u_h^n is a $\mathcal{F}_{t_n}$-measurable random vector for all $n = 1, \dots, N$. We have already shown that for any given $\xi_h \in \mathbb{R}^{d_h}$ there exists a unique $w_h = (w_{K_i})_{1 \le i \le d_h} \in \mathbb{R}^{d_h}$ such that $\mathbf{Q}(w_h) = \xi_h$ where $\mathbf{Q} : \mathbb{R}^{d_h} \to \mathbb{R}^{d_h}$, $\mathbf{Q}(w_h) = (Q_1, \dots, Q_{d_h})(w_h)$ is defined by $Q_i(w_h) = P_i(w_h) - \frac{m_{K_i}}{\Delta t} \xi_i^n$ for all $i = 1, \dots, d_h$. Thus, $u_h^{n+1} = \mathbf{Q}^{-1}(\xi_h^n)$ $\mathbb{P}$-a.s. in Ω, where $\xi_h^n = (\xi_1^n, \dots, \xi_{d_h}^n)$. Since $\mathbf{Q}^{-1}$ is continuous, if ξ_h^n is $\mathcal{F}_{t_{n+1}}$-measurable the same holds true for u_h^{n+1}. Indeed, let $(\zeta^k)_k \subset \mathbb{R}^{d_h}$ be a sequence such that $\zeta^k \to \zeta$ for some $\zeta \in \mathbb{R}^{d_h}$ for $k \to \infty$. Then, for $w^k := \mathbf{Q}^{-1}(\zeta^k)$ from (9) and from the theorem of Bolzano-Weierstrass it follows that there exists $w \in \mathbb{R}^{d_h}$ such that, passing to an unlabelled subsequence if necessary, $w^k \to w$ for $k \to \infty$. This strong convergence is enough to pass to the limit in $\mathbf{Q}(w^k)$, and therefore $\mathbf{Q}(w) = \zeta$. Thanks to uniqueness, we get convergence of the whole sequence $(w^k)_k$, hence $\lim_{k \to \infty} \mathbf{Q}^{-1}(\zeta^k) = \mathbf{Q}^{-1}(\zeta)$ and $\mathbf{Q}^{-1}$ is continuous. □

Acknowledgements The authors would like to thank the German Research Foundation, the Institut de Mécanique et d'Ingenierie of Marseille, the Programmes for Project-Related Personal Exchange Procope and Procope Plus and the Procope Mobility Program DEU-22-0004 LG1 for financial support.

References

1. Eymard, R., Gallouët, T., Herbin, R.: Finite volume methods. In: Lions, J.L., Ciarlet, P. (eds.) Solution of Equation in $\mathbb{R}^n$ (Part 3), Techniques of Scientific Computing (Part 3), Handbook of Numerical Analysis, vol. 7, pp. 730–1020. Elsevier (2000)
2. Krylov, N.V., Rozovskii, B.L.: Stochastic evolution equations. J. Sov. Math. **16**(4), 1233–1277 (1981)
3. Lions, J.-L.: Quelques Méthodes de Résolution des Problèmes aux Limites Non Linéaires. Dunod, Paris (2002)
4. Boyaval, S., Martel, S., Reygner, J.: Finite-volume approximation of the invariant measure of a viscous stochastic scalar conservation law. IMA J. Numer. Anal. **42**(3), 2710–2770 (2022)

A New Analysis for a Super-Convergence Result in the Divergence Norm for Lowest Order Raviart–Thomas Mixed Finite Elements Combined with the Crank–Nicolson Method Applied to One Dimensional Parabolic Equations

Fayssal Benkhaldoun and Abdallah Bradji

Abstract We consider the $\mathbb{RT}_0$-MFEs (lowest order Raviart–Thomas mixed finite elements) combined with the Crank–Nicolson method applied to parabolic equations in one dimensional space. We first justify the super-convergence of "velocity" $p = -u'$ in H^1-norm when using $\mathbb{RT}_0$-MFEs applied to one dimensional elliptic equations. We subsequently extend the results to $\mathbb{RT}_0$-MFEs as discretization in space and Crank–Nicolson method applied to the one dimensional non-stationary heat equation. More precise, we state and prove the super-convergence of "velocity" $p(t) = -u_x(t)$ in the $L^2(H^1)$-norm. The super-convergence result of $\mathbb{RT}_0$-MFEs combined with the Crank–Nicolson method is obtained thanks to a novel discrete a priori estimate. This work is a continuation of the two papers [2, 3] which dealt with the convergence in the divergence norm of MFEMs applied to parabolic equations. It is also an initiation of a future work addressing the analysis of the super-convergence in the divergence norm of fully discrete MFE schemes applied to multi-dimensional parabolic equations.

Keywords Parabolic equations · Super-convergence in the divergence norm · $\mathbb{RT}_0$-MFEs · Crank–Nicolson method · One dimensional space

MSC2020 35K05 · 65N30 · 65M15 · 65M60

Supported by MCS team (LAGA Laboratory) of the "Université Sorbonne- Paris Nord".

F. Benkhaldoun · A. Bradji
LAGA, USPN, Villetaneuse, Paris, France
e-mail: fayssal@math.univ-paris13.fr
URL: https://www.math.univ-paris13.fr/~fayssal/

A. Bradji (✉)
LMA, Badji Mokhtar-Annaba University, Annaba, Algeria
e-mail: abdallah.bradji@univ-annaba.dz; abdallah.bradji@gmail.com; abdallah.bradji@etu.univ-amu.fr; bradji@math.univ-paris13.fr
URL: https://www.i2m.univ-amu.fr/perso/abdallah.bradji/

E. Franck et al. (eds.), *Finite Volumes for Complex Applications X—Volume 1, Elliptic and Parabolic Problems*, Springer Proceedings in Mathematics & Statistics 432,
https://doi.org/10.1007/978-3-031-40864-9_11

1 Motivation and Aim of This Note

Let us consider the one dimensional non-stationary heat equation, as a model for one dimensional parabolic equations

$$u_t(\boldsymbol{x},t) - u_{xx}(\boldsymbol{x},t) = f(\boldsymbol{x},t), \qquad (\boldsymbol{x},t) \in \mathbf{I} \times (0,T), \tag{1}$$

where $\mathbf{I} = (0,1)$, $T > 0$, and f is a given function defined on $\mathbf{I} \times (0,T)$. This equation is equipped with an initial condition given by:

$$u(\boldsymbol{x},0) = u^0(\boldsymbol{x}), \qquad \boldsymbol{x} \in \mathbf{I}, \tag{2}$$

where u^0 is a given function defined on $\mathbf{I}$, and the homogeneous Dirichlet boundary conditions

$$u(0,t) = u(1,t) = 0, \qquad t \in (0,T). \tag{3}$$

As a "formal" mixed formulation for (1)–(3) is (see for instance [7, p. 53]), for each $t \in (0,T)$, find $(p(t), u(t)) \in H_{\mathrm{div}}(\mathbf{I}) \times L^2(\mathbf{I})$ such that, for all $(\varphi, \psi) \in L^2(\mathbf{I}) \times H_{\mathrm{div}}(\mathbf{I})$

$$(u_t(t), \varphi)_{L^2(\mathbf{I})} + (\varphi, \operatorname{div} p(t))_{L^2(\mathbf{I})} = (\varphi, f(t))_{L^2(\mathbf{I})}\,, \tag{4}$$

$$(\psi, p(t))_{L^2(\mathbf{I})} = (\operatorname{div}\psi, u(t))_{L^2(\mathbf{I})}\,, \tag{5}$$

and

$$u(0) = u^0. \tag{6}$$

The space $H_{\mathrm{div}}(\mathbf{I})$ in the case of one dimension is given by the Sobolev space $H_{\mathrm{div}}(\mathbf{I}) = H^1(\mathbf{I})$. The mesh points of $\mathbf{I} = (0,1)$ are denoted by $0 = \boldsymbol{x}_0 < \boldsymbol{x}_1 \cdots < \boldsymbol{x}_{M+1} = 1$, with $M \in \mathbb{N} \setminus \{0\}$, and the constant step is given by $h = \boldsymbol{x}_{i+1} - \boldsymbol{x}_i = 1/(M+1)$. We consider the sub-intervals $\mathbf{I}_i = (\boldsymbol{x}_i, \boldsymbol{x}_{i+1})$, for $i \in [\![0, M]\!]$. The discretization of the spaces $H_{\mathrm{div}}(\mathbf{I})$ and $L^2(\mathbf{I})$ is performed using the $\mathbb{RT}_0$-MFEs (see [8, Sect. 7.2.2, pp. 235–236]):

$$V_h^{\mathrm{div}} = \{v \in H_{\mathrm{div}}(\mathbf{I}) : \quad v|_{\mathbf{I}_i} \in \mathbb{D}_0, \quad \forall i \in [\![0, M]\!]\} \tag{7}$$

and

$$W_h = \{u \in L^2(\mathbf{I}) : \quad u|_{\mathbf{I}_i} \in \mathbb{P}_0, \quad \forall i \in [\![0, M]\!]\}, \tag{8}$$

where $\mathbb{P}_0$ is the space of constant functions and

$$\mathbb{D}_0 = \mathbb{P}_0 \oplus \boldsymbol{x}\mathbb{P}_0.$$

The space V_h^{div} (resp. W_h) can be defined as the set of continuous functions (resp. functions of $L^2(\mathbf{I})$) which are linear (resp. constant) over each $\mathbf{I}_i$, see [8, Proposition 3.2.1, p. 75]. The basis of V_h^{div} is the set of usual piece-wise linear shape functions.

We shall use the following interpolation operators over the spaces V_h^{div} and W_h, see [9]:

- The usual linear interpolation operator $\varPi_h$ over V_h^{div}.
- The interpolation operator J_h over W_h given by $J_h u(\boldsymbol{x}) = J_i$, for $\boldsymbol{x} \in \mathbf{I}_i$ with J_i is given by the mean value over $\mathbf{I}_i$ (see [9, p. 2])

$$J_i = \frac{1}{h} \int_{\mathbf{I}_i} u(\boldsymbol{x}) d\boldsymbol{x}. \tag{9}$$

This gives the property $\int_0^1 (u - J_h u)(\boldsymbol{x}) d\boldsymbol{x} = 0$.

The discretization in time is performed using a constant time step $k = T/(N+1)$, where $N \in \mathbb{N}^\star$. We shall denote $t_n = nk$, for $n \in [\![0, N+1]\!]$. We need to use the discrete temporal derivative operator ∂^1 given by $\partial^1 v^n = (v^n - v^{n-1})/k$. We will use the notation $\partial^0 v^n = v^n$ and we will denote by $v^{n-\frac{1}{2}}$ the arithmetic mean value $v^{n-\frac{1}{2}} = (v^n + v^{n-1})/2$.

The letter C in this note is a positive constant independent of h and k.

In order that the scheme we shall consider be meaningful, we assume that $f \in \mathcal{C}([0, T]; \ L^2(\mathbf{I}))$ and $u^0 \in H^2(\mathbf{I})$. We are interested in this work with the following second order time accurate MFE scheme (based on the use of the Crank–Nicolson method) introduced in [2]: Find $(p_h^n, u_h^n) \in V_h^{\mathrm{div}} \times W_h$ such that:

- For any $n \in [\![0, N]\!]$ and for all $\varphi \in W_h$:

$$\left(\partial^1 u_h^{n+1}, \varphi\right)_{L^2(\mathbf{I})} + \left(\left(p_h^{n+\frac{1}{2}}\right)_{\boldsymbol{x}}, \varphi\right)_{L^2(\mathbf{I})} = \left(f(t_{n+\frac{1}{2}}), \varphi\right)_{L^2(\mathbf{I})}, \tag{10}$$

- For any $n \in [\![0, N+1]\!]$:

$$\left(p_h^n, \psi\right)_{L^2(\mathbf{I})} = \left(u_h^n, \psi_{\boldsymbol{x}}\right)_{L^2(\mathbf{I})}, \qquad \forall \psi \in V_h^{\mathrm{div}}, \tag{11}$$

where

$$p_h^0 = -\varPi_h (u^0)_{\boldsymbol{x}}. \tag{12}$$

Using some slight modifications on the proof of [2, (93), p. 96], we get:

$$\max_{n=0}^{N} \| u_x(t_{n+\frac{1}{2}}) + p_h^{n+\frac{1}{2}} \|_{H^1(\mathbf{I})} \le C \left(h + k^2\right). \tag{13}$$

The aim of this contribution is bi-fold:

- We improve the order in space (which is only h) in (13) to order h^2 by comparing the discrete solution $p_h^{n+\frac{1}{2}}$ with the linear interpolation Π_h of $p = -u_{\boldsymbol{x}}$, i.e. that is a super-convergence for the MFE scheme (10)–(12) (see [9]).
- Prove the order two in space stated in the previous item in the divergence norm (in space), i.e. H^1-norm. More precise, we shall prove this super-convergence result in $L^2(H^1)$–norm.

Indeed, to the best of our knowledge, we are not aware of existing results which state these super- convergence results for parabolic equations in the divergence norm in space. We find the result of [7] in which a super-convergence result is obtained for a first order time accurate $R\mathbb{T}_0$-rectangular scheme but the analysis is performed only in L^2–norm in space.

As stated in the abstract, this note is an extension to our previous works [2, 3] which dealt with the analysis of the convergence of respectively second order time accurate and first order time accurate MFE schemes for parabolic equation in the **divergence norm** for velocity. In fact, as mentioned in our previous works [2–4], in contrast to elliptic equations, there is a lack of literature dealing with the convergence of fully discrete MFEMs in the divergence norm for velocity $p = -\nabla u$ of parabolic equations, see [7].

2 Super-Convergence of MFE Applied to Elliptic Equations

Before we state our main contribution, i.e. super-convergence of MFEs applied to parabolic equation, let us first highlight the situation for elliptic equations. To the best our knowledge, the results of this section *are not stated explicitly in the existing literature but they can be deduced from some known results in MFEMs, e.g.* [8, Theorem 7.4.1, p. 249]. Let us consider the following second order elliptic equation in 1D:

$$-\,\omega_{\boldsymbol{xx}}(\boldsymbol{x}) = F(\boldsymbol{x}), \quad \boldsymbol{x} \in \mathbf{I} = (0, 1), \tag{14}$$

with $\omega(0) = \omega(1) = 0$ and $F \in L^2(\mathbf{I})$.

The mixed formulation of the problem (14) is given by: Find $(p, \omega) \in H^1(\mathbf{I}) \times L^2(\mathbf{I})$ such that, for all $(\varphi, \psi) \in L^2(\mathbf{I}) \times H^1(\mathbf{I})$

$$(p_{\boldsymbol{x}}, \varphi)_{L^2(\mathbf{I})} = (F, \varphi)_{L^2(\mathbf{I})} \quad \text{and} \quad (p, \psi)_{L^2(\mathbf{I})} = (\omega, \psi_{\boldsymbol{x}})_{L^2(\mathbf{I})}\,. \tag{15}$$

The MFE scheme for the problem (14) is: Find $(p_h, \omega_h) \in V_h^{\mathrm{div}} \times W_h$ such that, for all $(\varphi, \psi) \in W_h \times V_h^{\mathrm{div}}$

$$\left((p_h)_{\boldsymbol{x}}\,, \varphi\right)_{L^2(\mathbf{I})} = (F, \varphi)_{L^2(\mathbf{I})} \quad \text{and} \quad (p_h, \psi)_{L^2(\mathbf{I})} = (\omega_h, \psi_{\boldsymbol{x}})_{L^2(\mathbf{I})}\,. \tag{16}$$

Using the error estimate [8, (7.2.26), p. 237] yields the following first order estimate for the MFE scheme (16)

$$\|p_h - p\|_{1,\mathbf{I}} + \|\omega_h - \omega\|_{L^2(\mathbf{I})} \leq Ch. \tag{17}$$

The error estimate (17) is optimal and the comparison was performed with respect to p and ω. However, we shall justify, see (23) below, that the approximate solution (p_h, ω_h) is closer to some suitable interpolation of (p, ω) than the predicted by (17). We refer to [6] and references therein for more details on the super-convergence phenomenon and their uses.

From (15)–(16), we deduce that

$$\mathbf{a}(\mu, v) - \mathbf{b}(v, \eta) = l(v), \quad \forall v \in V_h^{\text{div}} \quad \text{and} \quad \mathbf{b}(\mu, \nu) = \sigma(\nu), \quad \forall \nu \in W_h. \tag{18}$$

where

$$(\mu, \eta) = (p_h - \Pi_h p, \omega_h - J_h\omega)$$

$$\mathbf{a}(\mu, v) = (\mu, v)_{L^2(\mathbf{I})}\,, \quad \mathbf{b}(v, \eta) = (\eta, v_{\boldsymbol{x}})_{L^2(\mathbf{I})}\,,$$

$$l(v) = (p - \Pi_h p, v)_{L^2(\mathbf{I})} - (\omega - J_h\omega, v_{\boldsymbol{x}})_{L^2(\mathbf{I})}\,, \quad \forall v \in V_h^{\text{div}},$$

and $\sigma(\nu) = \big((p - \Pi_h p)_{\boldsymbol{x}}\,, \nu\big)_{L^2(\mathbf{I})}$, for all $\nu \in W_h$. Applying Theorem [8, Theorem 7.4.1, p. 249] on (18) yields

$$\|p_h - \Pi_h p\|_{1,\mathbf{I}} + \|\omega_h - J_h\omega\|_{L^2(\mathbf{I})} \leq C\left(\|l\|_{X'} + \|\sigma\|_{M'}\right), \tag{19}$$

where $(X, M) = (V_h^{\text{div}}, W_h)$.

Let us now estimate $\|l\|_{X'}$ and $\|\sigma\|_{M'}$ which are involved in (19):

- **Estimate of** $\|l\|_{X'}$. Using the fact that $v_{\boldsymbol{x}}$ is constant on each $\mathbf{I}_i$ and the definition (9)

$$\begin{aligned} l(v) &= (p - \Pi_h p, v)_{L^2(\mathbf{I})} - \sum_{i=0}^{M} v_{\boldsymbol{x}} \int_{\boldsymbol{x}_i}^{\boldsymbol{x}_{i+1}} (\omega - J_h\omega)\,(\boldsymbol{x})d\boldsymbol{x} \\ &= (p - \Pi_h p, v)_{L^2(\mathbf{I})}\,. \end{aligned} \tag{20}$$

  Gathering this with the known L^2-estimate of the interpolation error gives $|l(v)| \leq Ch^2|p|_{2,\mathbf{I}}\|v\|_{1,\mathbf{I}}$. This implies that

$$\|l\|_{X'} \leq Ch^2|p|_{2,\mathbf{I}}. \tag{21}$$

- **Estimate of** $\|\sigma\|_{M'}$. Using the fact that $\nu \in W_h$ is constant on each $\mathbf{I}_i$

$$\sigma(\nu) = \sum_{i=0}^{M} \nu \int_{\boldsymbol{x}_i}^{\boldsymbol{x}_{i+1}} (p - \Pi_h p)_{\boldsymbol{x}} (\boldsymbol{x}) d\boldsymbol{x} = \sum_{i=0}^{M} \nu \left(p - \Pi_h p\right) |_{\boldsymbol{x}_i}^{\boldsymbol{x}_{i+1}} = 0. \quad (22)$$

Gathering (19), (21), and (22) yields the following Super-convergence result

$$\|p_h - \Pi_h p\|_{1,\mathbf{I}} + \|\omega_h - J_h \omega\|_{L^2(\mathbf{I})} \leq Ch^2. \quad (23)$$

3 Statement of the Main Results: Super-Convergence of MFE, in the Divergence Norm, for (1)–(3)

Theorem 1 (New error estimate for the MFE scheme (10)–(12)) *Let $\boldsymbol{I} = (0, 1)$. Assume that the solution of (1)–(3) satisfies $u \in \mathcal{C}^3([0, T]; H^3(\boldsymbol{I}))$. Let us consider the mesh points of $\boldsymbol{I}$: $0 = \boldsymbol{x}_0 < \boldsymbol{x}_1 \cdots < \boldsymbol{x}_{M+1} = 1$, where $M \in \mathbb{N}^\star$, with a constant step $h = 1/(M + 1)$. Let $\boldsymbol{I}_i$ be the the sub-intervals $\boldsymbol{I}_i = (\boldsymbol{x}_i, \boldsymbol{x}_{i+1})$. Let $k = T/(N + 1)$, where $N \in \mathbb{N}^\star$. We define $t_n = nk$, for $n \in [\![0, N + 1]\!]$. Let $V_h^{\text{div}} \subset H_{\text{div}}(\boldsymbol{I}) = H^1(\boldsymbol{I})$ and $W_h \subset L^2(\boldsymbol{I})$ be the two lowest order Raviart–Thomas finite element spaces given respectively by (7) and (8). Let Π_h be the usual linear interpolation over V_h^{div} and J_h be the interpolation over W_h given by the value of (9) on the sub-interval $\boldsymbol{I}_i$.*

Then, there exists a unique solution $\left((p_h^n, u_h^n)\right)_{n=0}^{N+1} \in \left(V_h^{\text{div}} \times W_h\right)^{N+2}$ for the MFE scheme (10)–(12) and the following $L^2(H_{\text{div}}(\boldsymbol{I}))$–error estimate holds:

$$\left(\sum_{n=0}^{N} k \left\| \Pi_h u_x(t_{n+\frac{1}{2}}) + p_h^{n+\frac{1}{2}} \right\|_{1,\boldsymbol{I}}^2 \right)^{\frac{1}{2}} \leq C(h + k)^2. \quad (24)$$

To prove Theorem 1, we need to use the following a priori estimate which we set as theorem for its own importance:

Theorem 2 (A new generic discrete well-posedness result) *Under the same hypotheses of Theorem 1, let $\left(\left(\xi_h^n\right)_{n=0}^{N+1}, \left(\overline{\xi}_h^n\right)_{n=0}^{N+1} \right) \in \left(V_h^{\text{div}}\right)^{N+2} \times W_h^{N+2}$ be satisfied:*

- *For any $n \in [\![0, N]\!]$, for all $\varphi \in W_h$:*

$$\left(\partial^1 \overline{\xi}_h^{n+1}, \varphi\right)_{L^2(\boldsymbol{I})} + \left(\left(\xi_h^{n+\frac{1}{2}}\right)_{\boldsymbol{x}}, \varphi\right)_{L^2(\boldsymbol{I})} = \sigma^{n+1}(\varphi), \quad (25)$$

- *For any $n \in [\![0, N + 1]\!]$, for all $\psi \in V_h^{\text{div}}$:*

$$\left(\xi_h^n, \psi\right)_{L^2(\boldsymbol{I})} - \left(\overline{\xi}^n, \psi_{\boldsymbol{x}}\right)_{L^2(\boldsymbol{I})} = l^n(\psi), \quad (26)$$

where $\sigma^{n+1} \in M'$ (resp. $l^n \in X'$), for all $n \in [\![0, N]\!]$ (resp. for all $n \in [\![0, N + 1]\!]$) with $(X, M) = (V_h^{\text{div}}, W_h)$.

Then, the following $L^2(H_{\mathrm{div}}(\boldsymbol{I}))$–estimate holds:

$$\left(\sum_{n=0}^{N} k\|\xi_h^{n+\frac{1}{2}}\|_{1,\boldsymbol{I}}^2\right)^{\frac{1}{2}} \leq C\left(\mathcal{L}^2 + \Sigma^2 + \|\xi_h^0\|_{L^2(\boldsymbol{I})}^2\right), \tag{27}$$

where we have denoted $\Sigma = \max_{n=0}^N \|\sigma^{n+1}\|_{M'}$ and $\mathcal{L} = \max_{n=0}^N \|\partial^1 l^{n+1}\|_{X'}$.

Proof The proof of Theorem 2 uses the techniques of the proof of [2, (23), p. 89]. Indeed, (27) generalizes a part in the estimate [2, (23), p. 89]. We will detail this in the future paper [1]. □

3.1 Sketch of Proof of Theorem 1

Taking $t = t_{n+\frac{1}{2}}$ in (1) and multiplying both sides by $\varphi \in L^2(\mathbf{I})$ yields (recall that $p = -u_{\boldsymbol{x}}$)

$$\left(\partial^1 \overline{\xi}_h^{n+1}, \varphi\right)_{L^2(\mathbf{I})} + \left(\left(\xi_h^{n+\frac{1}{2}}\right)_x, \varphi\right)_{L^2(\mathbf{I})} = \sigma^{n+1}(\varphi), \tag{28}$$

where $(\overline{\xi}^{n+1}, \xi_h^n) = (J_h u(t_{n+1}) - u_h^{n+1}, \Pi_h p(t_n) - p_h^n) \in W_h \times V_h^{\mathrm{div}}$, and, using reasoning similar to those used in (20) and (22)

$$\sigma^{n+1}(\varphi) = \left(\partial^1 u(t_{n+1}) - u_t(t_{n+\frac{1}{2}}), \varphi\right)_{L^2(\mathbf{I})}. \tag{29}$$

On the other hand, multiplying $p(t_n) = -u_{\boldsymbol{x}}(t_n)$ by $\psi \in V_h^{\mathrm{div}}$, using an integration by parts, and using again reasoning similar to that used in (20) yield

$$\left(\xi_h^n, \psi\right)_{L^2(\Omega)^d} - \left(\overline{\xi}^n, \psi_x\right)_{L^2(\mathbf{I})} = l^n(\psi), \tag{30}$$

where $l^n(\psi) = \left(\left(\Pi_h p - p\right)(t_n), \psi\right)_{L^2(\mathbf{I})}$. Using a Taylor expansion and the L^2-estimate of the interpolation error yields

$$|\sigma^{n+1}(\varphi)| \leq Ck^2\|\varphi\|_{L^2(\mathbf{I})} \quad \text{and} \quad |\partial^1 l^{n+1}(\psi)| \leq Ch^2\|\psi\|_{1,\mathbf{I}}. \tag{31}$$

Applying now the a priori estimate (27) on (28) and (30) and using estimates (31) together with the fact that $\xi_h^0 = 0$ (stems from the initial condition (2)) yields the desired estimate (24). This completes the proof of Theorem 1. □

Remark 1 (*An extension to 2D*) The present results can be extended, for instance, to two dimensional parabolic equations $u_t(\boldsymbol{x}, t) - \Delta u(\boldsymbol{x}, t) = f(\boldsymbol{x}, t)$, for $(\boldsymbol{x}, t) \in \Omega \times (0, T)$, where $\Omega \subset \mathbb{R}^2$ is a rectangular domain meshed with regular rectangular elements denoted by $\mathcal{T}_h$. Let us define the interpolation operator Π_h, see [9], for every $K \in \mathcal{T}_h$ and side σ of K

$$\Pi_h p|_K = \Pi_K p \quad \text{and} \quad \int_\sigma (p - \Pi_K p) \cdot \mathbf{n}_{K,\sigma} ds = 0,$$

where $\mathbf{n}_{K,\sigma}$ is the unit vector normal to σ outward to K. The MFE scheme is given by [2, (8)–(10), p. 87]. Following the same steps of the proof of Theorem 1 and using a multi-dimensional version of Theorem 2 together with [9, Lemma 2.1, p. 2], we are able to prove the following super-convergence result:

$$\left(\sum_{n=0}^{N} k \left\| \Pi_h \nabla u(t_{n+\frac{1}{2}}) + p_h^{n+\frac{1}{2}} \right\|_{H_{\rm div}(\Omega)}^2 \right)^{\frac{1}{2}} \leq C(h+k)^2.$$

This result will be detailed in [1].

4 Some Computational Results

We consider the particular case of (1) with $u(\boldsymbol{x}, t) = (1/\pi^2) \sin(\pi^2 t) \sin(\pi \boldsymbol{x})$ and $T = 1$. To show the numerical order (see [3]), we assume an expression for the error in a given norm denoted by $|\cdot|$ of the form $|e_{h,k}| = C \cdot h^r + D \cdot k^s$. This yields $s = \log_2 \left(\left(|e_{h,k}| - |e_{h,k/2}| \right) / \left(|e_{h,k/2}| - |e_{h,k/4}| \right) \right)$ and $r = \log_2 \left(\left(|e_{h,k}| - |e_{h/2,k}| \right) / \left(|e_{h/2,k}| - |e_{h/4,k}| \right) \right)$. We show numerically in the following two tables that the order of the errors in the approximation of the "velocity" $p = -u_{\boldsymbol{x}}$, is around 2 in space and time. This order holds not only in the norm of $L^2(H_{\rm div}(\mathbf{I}))$ but also in the $L^\infty(H_{\rm div}(\mathbf{I}))$–norm. This confirms the error estimate (24). The results of these two tables are computed respectively when $k = 1/1000$ and $h = 1/500$. The sentence "Convergence Order" is abbreviated to "**CO**". We will denote by **EP** the error in the approximation of $p = -u_{\boldsymbol{x}}$. The notation "–" corresponds to values of k or h of which the numerical order can not be computed using the above stated formulas.

h	$\lVert EP \rVert_{L^\infty(H_{\rm div})}$	
	Error	CO
1/20	7.559238e-04	–
1/40	1.869095e-04	–
1/80	4.467610e-05	2.0002046
1/160	9.484347e-06	2.014951

k	$\lVert EP \rVert_{L^\infty(H_{\rm div})}$	
	Error	CO
1/150	1.821064e-04	–
1/300	4.488298e-05	–
1/600	1.060886e-05	2.0013352
1/1200	2.169986e-06	2.0219972

5 Conclusion and Perspectives

A new analysis for a super-convergence result in the divergence norm towards the "velocity" is proved for lowest order Raviart–Thomas mixed finite elements combined with the Crank–Nicolson method applied to the non-stationary one dimensional heat equation. This work is an initiation to a full work addressing the analysis of the super-convergence in the divergence norm for fully discrete MFE schemes applied to multi-dimensional parabolic equations.

Acknowledgements The authors would like to thank the anonymous referees for their interesting suggestions which helped to improve the paper.

References

1. Benkhaldoun, F., Bradji, A.: A new generic super-convergence result in the divergence norm for primal dual mixed finite element schemes applied to parabolic equations and examples. In preparation
2. Benkhaldoun, F., Bradji, A.: Novel analysis approach for the convergence of a second order time accurate mixed finite element scheme for parabolic equations. Comput. Math. Appl. **133**, 85–103 (2023)
3. Benkhaldoun, F., Bradji, A.: Two new error estimates of a fully discrete primal-dual mixed finite element scheme for parabolic equations in any space dimension. Results Math. **76/4**, Paper No. 182 (2021)
4. Benkhaldoun, F., Bradji, A.: A new error estimate for a primal-dual Crank-Nicolson mixed finite element using lowest degree Raviart-Thomas spaces for parabolic equations. In: Large-Scale Scientific Computing, pp. 489–497. Lecture Notes in Computer Science, vol. 13127. Springer, Cham (2022)
5. Chen, H., Ewing, R., Lazarov, R.: Superconvergence of mixed finite element methods for parabolic problems with nonsmooth initial data. Numer. Math. **78**(4), 495–521 (1998)
6. Ewing, R.E., Lazarov, R.D.: Superconvergence of the mixed finite element approximations of parabolic problems using rectangular finite elements. East-West J. Numer. Math. **1**(3), 199–212 (1993)
7. Johnson, C., Thomee, V.: Error estimates for some mixed finite element methods for parabolic type problems. RAIRO Anal. Numér. **15**(1), 41–78 (1981)
8. Quarteroni, A., Valli, A.: Numerical Approximation of Partial Differential Equations. Springer Series in Computational Mathematics 23. Springer, Berlin (2008)
9. Yang, H., Shi, D.: Superconvergence analysis of the lowest order rectangular Raviart-Thomas element for semilinear parabolic equation. Appl. Math. Lett. **105**, 106280, 9 pp (2020)

An $L^\infty(H^1)$-Error Estimate for Gradient Schemes Applied to Time Fractional Diffusion Equations

Fayssal Benkhaldoun and Abdallah Bradji

Abstract In this work, we extend the results of [3] to some second order time accurate GSs (Gradient Schemes) applied to a general **TFDE** (Time Fractional Diffusion Equation) with a space-dependent conductivity. The time fractional derivative is taken in the Caputo sense. The space discretization is performed using the general framework of GDM (Gradient Discretization Method) which encompasses several numerical methods. The approximation of the Caputo derivative is given by the known $L2-1_\sigma$-formula. We prove a new discrete $L^\infty(H^1)$-a priori estimate which, in turn, helps establishing a new $L^\infty(H^1)$-error estimate for the stated second order time accurate GSs. The GDM considered in this work is restricted to the cases of the numerical methods in which $\|\varPi_\mathcal{D}\cdot\|_{L^2(\varOmega)}$ is a norm, where $\varPi_\mathcal{D}$ is the reconstruction operator of the approximate functions in the space $L^2(\varOmega)$.

Keywords Time Fractional Diffusion Equation · Space dependent conductivity · GDM · Second order in time · New $L^\infty(H^1)$-error estimate

MSC2020: 65M08 · 65M12 · 65M15

Supported by MCS team (LAGA Laboratory) of the “Université Sorbonne- Paris Nord”.

F. Benkhaldoun · A. Bradji
LAGA, USPN, Villetaneuse, Paris, France
e-mail: fayssal@math.univ-paris13.fr
URL: https://www.math.univ-paris13.fr/∼fayssal/

A. Bradji (✉)
LMA, Badji Mokhtar-Annaba University, Annaba, Algeria
e-mail: abdallah.bradji@univ-annaba.dz; abdallah.bradji@gmail.com; abdallah.bradji@etu.univ-amu.fr; bradji@math.univ-paris13.fr
URL: https://www.i2m.univ-amu.fr/perso/abdallah.bradji/

E. Franck et al. (eds.), *Finite Volumes for Complex Applications X—Volume 1, Elliptic and Parabolic Problems*, Springer Proceedings in Mathematics & Statistics 432,
https://doi.org/10.1007/978-3-031-40864-9_12

1 Problem to Be Solved and Aim of This Work

We consider the following **TFDE** with space-dependent conductivity:

$$\partial_t^\alpha u(\boldsymbol{x},t) - \nabla \cdot (\kappa \nabla u)(\boldsymbol{x},t) = f(\boldsymbol{x},t), \qquad (\boldsymbol{x},t) \in \Omega \times (0,T), \tag{1}$$

where Ω is an open polygonal bounded subset in $\mathbb{R}^d$, $T > 0$, $0 < \alpha < 1$, and κ and f are given functions defined respectively on Ω and $\Omega \times (0,T)$. The operator ∂_t^α is the Caputo derivative given by, for a function φ defined on $(0,T)$

$$\partial_t^\alpha \varphi(t) = \frac{1}{\Gamma(1-\alpha)} \int_0^t (t-s)^{-\alpha} \varphi'(s) ds. \tag{2}$$

Initial and homogeneous Dirichlet boundary conditions are given by, for a given function u^0 defined on Ω

$$u(\boldsymbol{x},0) = u^0(\boldsymbol{x}), \quad \boldsymbol{x} \in \Omega \quad \text{and} \quad u(\boldsymbol{x},t) = 0, \quad (\boldsymbol{x},t) \in \partial\Omega \times (0,T). \tag{3}$$

For the sake of simplicity of the present note, we assume that the conductivity κ satisfies $\kappa \in \mathcal{C}^1(\overline{\Omega})$ and for some given $\kappa_0 > 0$

$$\kappa(\boldsymbol{x}) > \kappa_0 > 0, \quad \forall \boldsymbol{x} \in \Omega. \tag{4}$$

Fractional Partial Differential equations are important models because of their ability to represent several phenomena in applications, see for instance [1] and references therein. This work is an extension and improvement of our paper [3] in which a new $L^\infty(H^1)$-error estimate for a first order time accurate cell-centered SUSHI scheme for a **TFDE** is proved. In this paper, we extend the results of [3] to some second order time accurate schemes applied to a more general **TFDE** using the general generic framework GDM developed in [8]. One of the features of GDM is that the results we obtain are valid for the numerous numerical methods which are encompassed by GDM, e.g. SUSHI, Conforming FEMs (Finite Element Methods), Mixed FEMs, and Multi-point Flux Approximation MPFA-O Scheme. This work also improves some results of [4] in which an $L^2(H^1)$-error estimate is proved for a first order time accurate GS applied to a **TFDE**. In this note, we apply GDM combined with the $L2-1_\sigma$-formula, which is (at least) a second order approximation of the Caputo derivative (2) (see (9)–(10)) and developed in [1] and references therein, to a **TFDE** with a space dependent conductivity. We shall show a new $L^\infty(H^1)$-error estimate. Such an error estimate is obtained through a new a priori estimate. The present results improve, in addition to the results of [3, 4], the $L^\infty(L^2)$-error estimate proved in [5] for a second order time accurate SUSHI scheme applied to a **TFDE**.

2 Space Discretization

Definition 1 (*Approximate gradient discretization, cf.* [8]) Let Ω be an open domain of $\mathbb{R}^d$, with $d \in \mathbb{N}^\star$. An approximate gradient discretization is given by $\mathcal{D} = (\mathcal{X}_{\mathcal{D},0}, \Pi_\mathcal{D}, \nabla_\mathcal{D})$, where

1. The set of discrete unknowns $\mathcal{X}_{\mathcal{D},0}$ is a finite dimensional vector space on $\mathbb{R}$.
2. The linear mapping $\Pi_\mathcal{D} : \mathcal{X}_{\mathcal{D},0} \to L^2(\Omega)$ is the reconstruction of the approximate function.
3. The gradient reconstruction $\nabla_\mathcal{D} : \mathcal{X}_{\mathcal{D},0} \to L^2(\Omega)^d$ is a linear mapping which reconstructs, from an element of $\mathcal{X}_{\mathcal{D},0}$, a "gradient" (vector-valued function) over Ω. The gradient reconstruction must be chosen such that $\|\nabla_\mathcal{D} \cdot \|_{L^2(\Omega)^d}$ is a norm on $\mathcal{X}_{\mathcal{D},0}$. Let us define the bi-linear form $\langle \cdot, \cdot \rangle_{\mathcal{D},\kappa}$ given by

$$\langle u, v\rangle_{\mathcal{D},\kappa} = \int_\Omega \kappa(\boldsymbol{x})\nabla_\mathcal{D} u(\boldsymbol{x}) \cdot \nabla_\mathcal{D} v(\boldsymbol{x}) d\boldsymbol{x}, \quad \forall (u, v) \in \mathcal{X}_{\mathcal{D},0} \times \mathcal{X}_{\mathcal{D},0}. \tag{5}$$

In order to analyse the convergence of the gradient schemes, we consider the following parameters related to the mesh $\mathcal{D}$ given in Definition 1. These parameters are given in [8, Definitions 2.2–2.5]:

1. The coercivity of the discretization is measured via the constant $C_\mathcal{D}$ given by

$$C_\mathcal{D} = \max_{v \in \mathcal{X}_{\mathcal{D},0} \setminus \{0\}} \frac{\|\Pi_\mathcal{D} v\|_{L^2(\Omega)}}{\|\nabla_\mathcal{D} v\|_{L^2(\Omega)^d}}.$$

This yields the Poincaré inequality: $\|\Pi_\mathcal{D} v\|_{L^2(\Omega)} \leq C_\mathcal{D} \|\nabla_\mathcal{D} v\|_{L^2(\Omega)^d}$.
2. The consistency of the discretization is measured through the interpolation error function $S_\mathcal{D} : H_0^1(\Omega) \to [0, +\infty)$ defined by, for all $\varphi \in H_0^1(\Omega)$

$$S_\mathcal{D}(\varphi) = \min_{v \in \mathcal{X}_{\mathcal{D},0}} \left(\|\Pi_\mathcal{D} v - \varphi\|^2_{L^2(\Omega)} + \|\nabla_\mathcal{D} v - \nabla\varphi\|^2_{L^2(\Omega)^d} \right)^{\frac{1}{2}}.$$

3. The limit-conformity of the discretization is measured through the conformity error function $W_\mathcal{D} : H_{\text{div}}(\Omega) \to [0, +\infty)$ defined by, for all $\varphi \in H_{\text{div}}(\Omega)$

$$W_\mathcal{D}(\varphi) = \max_{u \in \mathcal{X}_{\mathcal{D},0} \setminus \{0\}} \frac{1}{\|\nabla_\mathcal{D} u\|_{L^2(\Omega)^d}} \left| \int_\Omega \left(\nabla_\mathcal{D} u(\boldsymbol{x}) \cdot \varphi(\boldsymbol{x}) + \Pi_\mathcal{D} u(\boldsymbol{x}) \text{div}\varphi(\boldsymbol{x}) \right) d\boldsymbol{x} \right|.$$

Assumption 1 (*Additional assumption on the gradient discretisation*) We assume, in addition, that the generic mesh $\mathcal{D} = (\mathcal{X}_{\mathcal{D},0}, \Pi_\mathcal{D}, \nabla_\mathcal{D})$ is chosen such that $\|\Pi_\mathcal{D} \cdot \|_{L^2(\Omega)}$ is a norm on $\mathcal{X}_{\mathcal{D},0}$. This includes, for instance, Conforming FEMs, Cell-Centered SUSHI, and MFEMs.

As an example of schemes for which $\|\Pi_\mathcal{D} \cdot \|_{L^2(\Omega)}$ is not a norm on $\mathcal{X}_{\mathcal{D},0}$, we quote Hybrid Mimetic Mixed methods (see [8, Chap. 13]).

3 Approximation of the Caputo Derivative and Properties

The second order approximation of the Caputo derivative (2) that we use here was introduced in [1]. We refer also to [5] for some highlights and properties (which we shall use in this work) of such an approximation. The time discretization is performed with a constant time step $k = T/(N+1)$, 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 $\partial^1 v^{j+1} = (v^{j+1} - v^j)/k$. We will also use the notation $\partial^0 v^n = v^n$.

Throughout this paper, the letter C stands for a positive real number independent of the parameters of the space and time discretizations.

For the sake of completeness of this work, we recall the principles of the second order approximation of the fractional order derivative (2). Such approximation is performed thanks to the $L2 - 1_\sigma$-formula. Let us consider the "fractional mesh point" $t_{n+\sigma} = (n+\sigma)k$ where $\sigma = 1 - \alpha/2$. We define the two-point barycentric element $v^{n+\sigma}$ by $v^{n+\sigma} = \sigma v^{n+1} + (1-\sigma)v^n$.

Using (2), the value $\partial_t^\alpha u(t_{n+\sigma})$ is given by

$$\frac{1}{\Gamma(1-\alpha)}\left(\sum_{j=1}^{n}\int_{t_{j-1}}^{t_j}(t_{n+\sigma}-s)^{-\alpha}u_s(s)ds + \int_{t_n}^{t_{n+\sigma}}(t_{n+\sigma}-s)^{-\alpha}u_s(s)\right)ds. \tag{6}$$

The $L2 - 1_\sigma$-formula is based on the approximation of the terms of the sum (resp. the last term) using quadratic interpolations (resp. a linear interpolation) in (6) of $\partial_t^\alpha u(t_{n+\sigma})$. The stated quadratic interpolations interpolate the function u on each sub-interval (t_{j-1}, t_j) on the points t_{j-1}, t_j, t_{j+1}. This yields, see [1, (27)–(28), p. 429]

$$\partial_t^\alpha u(t_{n+\sigma}) \approx \frac{k^{1-\alpha}}{\Gamma(2-\alpha)}\sum_{j=0}^{n} c_{n-j}^{\sigma,n}\partial^1 u(t_{j+1}), \tag{7}$$

where $c_0^{\sigma,0} = \sigma^{1-\alpha}$ and for all $n \geq 1, c_0^{\sigma,n} = \sigma^{1-\alpha} + b_0^\sigma, c_n^{\sigma,n} = d_{n+\sigma-1,\alpha} - b_{n-1}^\sigma$, and

$$c_j^{\sigma,n} = d_{j+\sigma-1,\alpha} + b_j^\sigma - b_{j-1}^\sigma, \quad \forall j \in [\![1, n-1]\!] \tag{8}$$

with

$$b_l^\sigma = \frac{1}{2-\alpha}\left((l+\sigma+1)^{2-\alpha} - (l+\sigma)^{2-\alpha}\right) - \frac{1}{2}\left((l+\sigma+1)^{1-\alpha} + (l+\sigma)^{1-\alpha}\right)$$

and $d_{s,\alpha} = (s+1)^{1-\alpha} - s^{1-\alpha}$.

For any $n \in [\![0, N]\!]$ and $\varphi \in \mathcal{C}^3([0, T])$, we define the error $\mathbb{T}_1^n(\varphi)$ by

$$\mathbb{T}_1^n(\varphi) = \partial_t^\alpha \varphi(t_{n+\sigma}) - \sum_{j=0}^{n} k\lambda_j^{n+1}\partial^1\varphi(t_{j+1}), \tag{9}$$

where (see [1, Lemma 2])

$$\lambda_j^{n+1} = \frac{c_{n-j}^{\sigma,n}}{k^\alpha \Gamma(2-\alpha)} \quad \text{and} \quad \left|\mathbb{T}_1^n(\varphi)\right| \le Ck^{3-\alpha}\|\varphi^{(3)}\|_{\mathcal{C}([0,T])}. \tag{10}$$

The following properties hold for the approximation (9)–(10), see [1, 5]:

$$\lambda_{j+1}^{n+1} > \lambda_j^{n+1} > \lambda_0 = \frac{1}{2T^\alpha \Gamma(1-\alpha)} \quad \text{and} \quad \sum_{j=0}^{n} k\lambda_j^{n+1} \le \frac{T^{1-\alpha}}{\Gamma(2-\alpha)} \tag{11}$$

and for all $\left(\beta^j\right)_{j=0}^{N+1} \in \mathbb{R}^{N+2}$ and for all $n \in [\![0, N]\!]$

$$\left(\sigma\beta^{n+1} + (1-\sigma)\beta^n\right)\sum_{j=0}^{n}\lambda_j^{n+1}(\beta^{j+1} - \beta^j) \ge \frac{1}{2}\sum_{j=0}^{n}\lambda_j^{n+1}\left((\beta^{j+1})^2 - (\beta^j)^2\right). \tag{12}$$

4 Formulation of a GS and Statement of the Main Results

We have, thanks to a convenient Taylor expansion, for any function $\varphi \in \mathcal{C}^2([0, T])$

$$\sigma\varphi(t_{n+1}) + (1-\sigma)\varphi(t_n) = \varphi(t_{n+\sigma}) + \mathbb{T}_2^n(\varphi), \tag{13}$$

where

$$|\mathbb{T}_2^n(\varphi)| \le \frac{k^2}{2}\|\varphi\|_{\mathcal{C}^2([0,T])}. \tag{14}$$

Replacing t with $t_{n+\sigma}$ in Eq. (1) and taking into account (9), (10), (13), and (14), we suggest the following GS, for the problem (1)–(3), which is inspired by the finite volume scheme given in [5].

Definition 2 (*Formulation of a GS for* (1)–(3)) Let $\alpha \in (0, 1)$ be given and $\sigma = 1 - \alpha/2 \in (0, 1)$. Let $\mathcal{D} = (\mathcal{X}_{\mathcal{D},0}, \Pi_\mathcal{D}, \nabla_\mathcal{D})$ be a gradient discretization in the sense of Definition 1 and satisfies Assumption 1. Assume in addition that $f \in \mathcal{C}\left([0, T]; L^2(\Omega)\right)$ and $u^0 \in H^2(\Omega)$. As approximation for (1)–(3), we define the following GS:

1. Find $u_\mathcal{D}^0 \in \mathcal{X}_{\mathcal{D},0}$ such that for all $v \in \mathcal{X}_{\mathcal{D},0}$

$$\langle u_\mathcal{D}^0, v\rangle_{\mathcal{D},\kappa} = -\left(\nabla \cdot (\kappa\nabla u^0), \Pi_\mathcal{D} v\right)_{L^2(\Omega)}. \tag{15}$$

2. For all $n \in [\![0, N]\!]$, find $u_\mathcal{D}^{n+1} \in \mathcal{X}_{\mathcal{D},0}$ such that, for all $v \in \mathcal{X}_{\mathcal{D},0}$

$$\sum_{j=0}^{n} k\lambda_j^{n+1}\left(\partial^1 \Pi_{\mathcal{D}} u_{\mathcal{D}}^{j+1}, \Pi_{\mathcal{D}} v\right)_{L^2(\Omega)} + \langle u_{\mathcal{D}}^{n+\sigma}, v\rangle_{\mathcal{D},\kappa} = \left(f(t_{n+\sigma}), \Pi_{\mathcal{D}} v\right)_{L^2(\Omega)}. \tag{16}$$

Using the techniques of the proof of [5, Theorem 1] together with [8, Theorem 2.28], we obtain respectively the following $L^\infty(L^2)$ and $L^\infty(H_0^1)$-error estimates:

$$\max_{n=0}^{N+1} \|u(t_n) - \Pi_{\mathcal{D}} u_{\mathcal{D}}^n\|_{L^2(\Omega)} \le C(k^2 + \mathbb{E}_{\mathcal{D}}^k(u)) \tag{17}$$

and

$$\max_{n=0}^{N} \left(\lambda_n^{n+1}\right)^{-\frac{1}{2}} \|\nabla u^{n+\sigma} - \nabla_{\mathcal{D}} u_{\mathcal{D}}^{n+\sigma}\|_{L^2(\Omega)} \le C(k^2 + \mathbb{E}_{\mathcal{D}}^k(u)), \tag{18}$$

where for any function $u \in \mathcal{C}([0,T];\ H^2(\Omega))$, we denote by

$$\mathbb{E}_{\mathcal{D}}^k(u) = \max_{j\in\{0,1\}} \max_{n\in[\![j,N+1]\!]} \mathbb{E}_{\mathcal{D}}(\partial^j u(t_n)) \tag{19}$$

and, for any $\overline{u} \in H^2(\Omega)$,

$$\mathbb{E}_{\mathcal{D}}(\overline{u}) = \max\left(C_{\mathcal{D}} W_{\mathcal{D}}(\kappa\nabla\overline{u}) + (C_{\mathcal{D}}+1) S_{\mathcal{D}}(\overline{u}),\ W_{\mathcal{D}}(\kappa\nabla\overline{u}) + 2S_{\mathcal{D}}(\overline{u})\right). \tag{20}$$

As stated in [5, Remark 1], the $L^\infty(L^2)$-error estimate (17) is optimal in the sense that it is the same one known for a Crank-Nicolson finite volume scheme approximating the heat equation in [7]. Since λ_n^{n+1} is of order $k^{-\alpha}$, the $L^\infty(H_0^1)$-error estimate (18) implies that $\|\nabla_{\mathcal{D}} u_{\mathcal{D}}^{n+\sigma} - \nabla u^{n+\sigma}\|_{L^2(\Omega)}$ is of order $k^{-\frac{\alpha}{2}}(k^2 + \mathbb{E}_{\mathcal{D}}^k(u))$ which is a conditional convergence and consequently is not optimal. The aim of this paper is to prove an unconditional $L^\infty(H_0^1)$-error estimate not only for the SUSHI method but also for the large class of GDM applied to the general Eq. (1). We shall improve the estimate (18) in the sense that the coefficient $\left(\lambda_n^{n+1}\right)^{-\frac{1}{2}}$ of the left hand side is removed and that the convergence not only holds on the barycentric points $\{t_{n+\sigma}, n \in [\![0,N]\!]\}$, but also on the mesh points $\{t_n, n \in [\![0,N+1]\!]\}$ (see (21) below).

Theorem 1 (New $L^\infty(H^1)$-error estimate for the GS (15)–(16)) *In addition to the hypotheses of Definition 2, we assume that the solution u of (1)–(3) (resp. the conductivity $\kappa \in \mathcal{C}^1(\overline{\Omega})$) satisfies $u \in \mathcal{C}^3([0,T]; \mathcal{C}^2(\overline{\Omega}))$ (resp. (4)). Let $\mathbb{E}_{\mathcal{D}}^k(u)$ be the expression given by (19)–(20). Then, there exists a unique solution $\left(u_{\mathcal{D}}^n\right)_{n=0}^{N+1} \in \mathcal{X}_{\mathcal{D},0}^{N+2}$ for the GS (15)–(16) and the following $L^\infty(H^1)$-error estimate holds:*

$$\max_{n=0}^{n=N+1} \|\nabla u(t_n) - \nabla_{\mathcal{D}} u_{\mathcal{D}}^n\|_{L^2(\Omega)} \le C(k^2 + \mathbb{E}_{\mathcal{D}}^k(u)). \tag{21}$$

To prove Theorem 1, we first define an approximation of the operator $-\nabla(\kappa\nabla\cdot)$ in Definition 3 below. Such approximation, which exists and is unique thanks to Assumption 1, is inspired by the discrete Laplace operator introduced in [9] in the context of SUSHI. We subsequently need also to use the new discrete $L^\infty(H^1)$-a priori estimate given in Lemma 1 below.

Definition 3 (*Approximation of the operator* $-\nabla(\kappa\nabla\cdot)$) Under the hypotheses of Theorem 1 and for any $u \in \mathcal{X}_{\mathcal{D},0}$, we define the discrete operator $\Delta^\kappa_{\mathcal{D}}$ relative to $\nabla(\kappa\nabla)$, as the unique element of $\mathcal{X}_{\mathcal{D},0}$ satisfying

$$-\left(\Pi_{\mathcal{D}}\Delta^\kappa_{\mathcal{D}}u, \Pi_{\mathcal{D}}v\right)_{L^2(\Omega)} = \langle u, v\rangle_{\mathcal{D},\kappa}, \quad \forall v \in \mathcal{X}_{\mathcal{D},0}. \tag{22}$$

Indeed, when Assumption 1 is satisfied, $(\Pi_{\mathcal{D}}\cdot, \Pi_{\mathcal{D}}\cdot)_{L^2(\Omega)}$ becomes an inner product on $\mathcal{X}_{\mathcal{D},0}$. The existence and uniqueness of $\Delta^\kappa_{\mathcal{D}}u$ results then from the Riesz representation theorem in $\mathcal{X}_{\mathcal{D},0}$ for the inner product $(\Pi_{\mathcal{D}}\cdot, \Pi_{\mathcal{D}}\cdot)_{L^2(\Omega)}$.

Lemma 1 (New discrete $L^\infty(H^1)$-a priori estimate) *Under the hypotheses of Theorem 1, we assume that there exists* $\left(\eta^n_{\mathcal{D}}\right)_{n=0}^{N+1} \in \mathcal{X}^{N+2}_{\mathcal{D},0}$ *such that for all* $n \in [\![0, N]\!]$ *and* $v \in \mathcal{X}_{\mathcal{D},0}$

$$\sum_{j=0}^{n} k\lambda_j^{n+1}\left(\partial^1\Pi_{\mathcal{D}}\eta^{j+1}_{\mathcal{D}}, \Pi_{\mathcal{D}}v\right)_{L^2(\Omega)} + \langle\eta^{n+\sigma}_{\mathcal{D}}, v\rangle_{\mathcal{D},\kappa} = \left(\mathcal{S}^{n+1}, v\right)_{L^2(\Omega)}, \tag{23}$$

where $\mathcal{S}^{n+1} \in L^2(\Omega)$, *for all* $n \in [\![0, N]\!]$, *and* $\eta^0_{\mathcal{D}} = 0$. *Let us denote* $\mathcal{S} = \max_{n=0}^{N} \|\mathcal{S}^{n+1}\|_{L^2(\Omega)}$. *Then, the following* $L^\infty(H^1)$*-a priori estimate holds:*

$$\max_{n=0}^{N+1} \|\nabla_{\mathcal{D}}\eta^n_{\mathcal{D}}\|_{L^2(\Omega)} \le C\mathcal{S}. \tag{24}$$

Proof The proof follows that of [3, Lemma 2, p. 310] with some minor modifications. Taking $v = -\Delta^\kappa_{\mathcal{D}}\eta^{n+\sigma}_{\mathcal{D}}$ in (23), using twice the definition (22), and using the inequalities $xy \le x^2/2 + y^2/2$ and (12) together with hypothesis (4) yield

$$\|\nabla_{\mathcal{D}}\eta^{n+1}_{\mathcal{D}}\|^2_{L^2(\Omega)} \le \frac{1}{\lambda_n^{n+1}}\left(\sum_{j=1}^{n}(\lambda_j^{n+1} - \lambda_{j-1}^{n+1})\|\nabla_{\mathcal{D}}\eta^j_{\mathcal{D}}\|^2_{L^2(\Omega)} + \frac{(\mathcal{S})^2}{\kappa_0}\right). \tag{25}$$

Applying now a mathematical induction on (25), as in the proof of [3, Lemma 2], yields estimate (24). □

4.1 Sketch of Proof of Theorem 1

The existence and uniqueness for (15)–(16) can be proved as in [4, Proof of Theorem 4.1, p. 506]. To prove (21), we compare the GS (15)–(16) with the following auxiliary GS: For any $n \in [\![0, N+1]\!]$, find $\overline{u}^n_{\mathcal{D}} \in \mathcal{X}_{\mathcal{D},0}$ such that

$$\langle\overline{u}^n_{\mathcal{D}}, v\rangle_{\mathcal{D},\kappa} = (-\nabla\cdot(\kappa\nabla u)(t_n), \Pi_{\mathcal{D}}v)_{L^2(\Omega)}, \qquad \forall v \in \mathcal{X}_{\mathcal{D},0}. \tag{26}$$

1. Comparison between the solutions of (26) **and** (1)–(3). The following estimates hold, see [4]:

$$\max_{n=0}^{n=N+1} \|\nabla u(t_n) - \nabla_{\mathcal{D}} \overline{u}_{\mathcal{D}}^n\|_{L^2(\Omega)^d} + \max_{n=1}^{n=N+1} \|\partial^1 (u(t_n) - \overline{u}_{\mathcal{D}}^n)\|_{L^2(\Omega)} \le C\mathbb{E}_{\mathcal{D}}^k(u). \tag{27}$$

2. Comparison between (15)–(16) **and** (26). Let us define the *auxiliary* error $\eta_{\mathcal{D}}^n = u_{\mathcal{D}}^n - \bar{u}_{\mathcal{D}}^n \in \mathcal{X}_{\mathcal{D},0}$. Taking $n = 0$ in (26), using the fact that $u(0) = u^0$, and comparing with (15) imply that $\eta_{\mathcal{D}}^0 = 0$. From (26), we deduce that

$$\langle \overline{u}_{\mathcal{D}}^{n+\sigma}, v\rangle_{\mathcal{D},\kappa} = -\left(\sigma\nabla\cdot(\kappa\nabla u)(t_{n+1}) + (1-\sigma)\nabla\cdot(\kappa\nabla u)(t_n), \Pi_{\mathcal{D}} v\right)_{L^2(\Omega)}. \tag{28}$$

Subtracting (28) from (16), subtracting $\sum_{j=0}^n k\lambda_j^{n+1}\left(\partial^1 \Pi_{\mathcal{D}} \overline{u}_{\mathcal{D}}^{j+1}, \Pi_{\mathcal{D}} v\right)_{L^2(\Omega)}$ from both sides of the result, and using (13), (1), and (9), we find that $\left(\eta_{\mathcal{D}}^n\right)_{n=0}^{N+1}$ satisfies the hypothesis (23) of Lemma 1 with

$$\mathcal{S}^{n+1} = \sum_{j=0}^n k\lambda_j^{n+1}\partial^1\left(u(t_{j+1}) - \Pi_{\mathcal{D}}\overline{u}_{\mathcal{D}}^{j+1}\right) + \mathbb{T}_1^n(u) + \mathbb{T}_2^n(\nabla\cdot(\kappa\nabla u)). \tag{29}$$

Using (27), (11), (10), and (14) implies that $\|\mathcal{S}^{n+1}\|_{L^2(\Omega)} \le C(k^2 + \mathbb{E}_{\mathcal{D}}^k(u))$. This estimate and the a priori estimate (24) imply that $\max_{n=0}^{N+1} \|\nabla_{\mathcal{D}}\eta_{\mathcal{D}}^n\|_{L^2(\Omega)} \le C(k^2 + \mathbb{E}_{\mathcal{D}}^k(u))$. Gathering this estimate with the error estimate (27) implies the desired estimate (21). □

5 Conclusion and a Perspective

We applied the GDM combined with the $L2 - 1_\sigma$-formula to a **TFDE** with a space dependent conductivity. A new $L^\infty(H^1)$-error estimate is proved thanks to a new well developed a priori estimate. The GDM considered in this work is restricted to the particular case of Assumption 1. One of the main perspectives is to extend the results to GDM without the restriction of Assumption 1.

Acknowledgements The authors would like to thank the anonymous referees for their valuable comments and interesting suggestions which helped to improve the paper. The second author would also like to thank Professors M. Sini and N. Tatar for their help to correct the syntax of this paper.

References

1. Alikhanov, A.-A.: A new difference scheme for the fractional diffusion equation. J. Comput. Phys. **280**, 424–438 (2015)
2. Benkhaldoun, F., Bradji, A.: A new generic scheme and a novel convergence analysis approach for time fractional diffusion equation and applications. In progress
3. Bradji, A.: A new optimal $L^\infty(H^1)$-error estimate of a SUSHI scheme for the time fractional diffusion equation. In: FVCA IX–methods, theoretical aspects, examples, Bergen, Norway, June 2020, pp. 305–314. Springer Proceedings in Mathematics and Statistics, 323. Springer, Cham (2020)
4. 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)
5. Bradji, A.: A second order time accurate SUSHI method for the time-fractional diffusion equation. In: Numerical Methods and Applications, pp. 197–206. Lecture Notes in Computer Science, 11189. Springer, Cham (2019)
6. 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)
7. Bradji, A.: An analysis of a second-order time accurate scheme for a finite volume method for parabolic equations on general nonconforming multidimensional spatial meshes. Appl. Math. Comput. **219**(11), 6354–6371 (2013)
8. Droniou, J., Eymard, R., Gallouët, T., Guichard, C., Herbin, R.: The Gradient Discretisation Method. Mathématiques et Applications, 82. Springer Nature Switzerland AG, Switzerland (2018)
9. 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)
10. Eymard, R., Guichard, C., Herbin, R.: Small-stencil 3D schemes for diffusive flows in porous media. ESAIM Math. Model. Numer. Anal. **46**(2), 265–290 (2012)
11. 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)

Compatible Discrete Operator Schemes for Solidification and Segregation Phenomena

Jérôme Bonelle and Thomas Fonty

Abstract The appearance of macro-segregations in the ingot casting process has led to the development of a solidification model in the Computational Fluid Dynamics software code_saturne. It relies on a mixture model encompassing mass, momentum, energy and solute transport equations. After having implemented this model for the Finite Volume (FV) scheme of code_saturne, it is here implemented for the Compatible Discrete Operator (CDO) framework. The resulting solidification and segregation predictions are validated against an academic test case. Integral and local comparisons are performed and exhibit a good agreement of the CDO approach with results obtained with the FV scheme and with the commercial software SOLID®. Moreover, the CDO approach relying on a strong velocity-pressure coupling brings a significant improvement in terms of robustness with respect to the time step, allowing for faster computations.

Keywords CDO · Navier–Stokes · Solidification · Segregation

MSC (2010) 65N08 · 76D05 · 76D99 · 76M12

1 Introduction

The manufacturing process of ingot casting is widespread in the nuclear industry. High quality ingots are expected to forge a nuclear vessel reactor for instance. Understanding the solidification process and the solute redistribution is of paramount importance. This process can namely lead to a potential segregation of alloys (chemical heterogeneities) that are likely to alter the mechanical properties of the materials. A numerical segregation model has been introduced in a previous work [6] in

J. Bonelle (✉) · T. Fonty
EDF R&D, 6 Quai Watier, 78400 Chatou, France
e-mail: jerome.bonelle@edf.fr

T. Fonty
e-mail: thomas.fonty@edf.fr

E. Franck et al. (eds.), *Finite Volumes for Complex Applications X—Volume 1, Elliptic and Parabolic Problems*, Springer Proceedings in Mathematics & Statistics 432,
https://doi.org/10.1007/978-3-031-40864-9_13

code_saturne, the open-source and general-purpose Computational Fluid Dynamic (CFD) software [7] developed at EDF R&D. This work relies on a cell-centered colocated Finite Volume (FV) scheme and a fractional step algorithm for the velocity/pressure coupling. In this paper, we adapt the same segregation model to the Compatible Discrete Operator (CDO) framework [3, 4, 8] with a fully coupled velocity/pressure algorithm. We compare on an academic test case this new approach to the existing one in code_saturne and to the one inside SOLID® [1], a commercial software dedicated to this kind of simulation. SOLID® relies on a staggered FV scheme along with a fractional step algorithm for the velocity/pressure coupling.

2 Segregation Model

Segregation involves multi-scale and multi-physics phenomena in a solid and liquid phase (cf. [9] for a review). A trade-off between the accuracy of a model and its complexity has to be found. In this work, we focus on the simulation of the macro-segregation phenomenon of a binary alloy induced by the thermo-solutal convection. Second-order phenomena such as solid grain movement, solidification shrinkage or deformation of the solid network are ignored. The model implemented in code_saturne relies on the seminal works of Benon and Incropera [2] and that of Voller and Prakash [11]. For a given variable y, the associated mixture average variable is denoted by $y_m := g_l y_l + (1 - g_l) y_s$ where g_l is the average liquid fraction on a representative elementary volume and y_l (resp. y_s) is the average variable in the liquid (resp. solid) phase. The model at stake considers the conservation laws for the mixture. The mixture is always assumed to be at the thermodynamic equilibrium. No motion in the solid phase is assumed. The micro-segregation is modelled through a closure law named the *lever rule*. Starting from the Navier–Stokes equations for an incompressible flow and for a Newtonian and laminar fluid under the Boussinesq approximation, a drag force is added in the momentum Eq. (1) following the Kozeny–Carman model [5]. The incompressibility constraint (2) states the mass conservation. The solidification process is namely seen as an evolutive porous media where the porosity decreases when the solid portion increases. The conservation of the energy (3) is solved using the temperature T_m as main unknown and a source term is added to take into account the phase change. Since one assumes a thermodynamic equilibrium, $T_m = T_s = T_l$. The mass density ρ, the dynamic viscosity μ, the thermal (resp. solutal) conductivity λ_T (λ_C), the specific heat capacity c_p, the latent heat L, the thermal (resp. solutal) coefficient of expansion β_T (resp. β_C) are all assumed to be constant. Let Ω be the computational domain. One considers homogeneous Dirichlet boundary conditions for the velocity and a zero-mean constraint for the pressure. Dirichlet and homogeneous Neumann boundary conditions are set on the temperature while a no-flux boundary condition is enforced for the solute concentration C_m. Initially, the fluid is at rest and $\forall \underline{x} \in \Omega$, $T_m(\underline{x}, 0) = T_0$, $C_m(\underline{x}, 0) = C_0$. With all these assumptions, choices of modeling and settings, we end up with the following system to solve: Find $(\underline{U}_m, p, T_m, C_m) \in \underline{H}_0^1(\Omega) \times L_0^2(\Omega) \times H^1(\Omega) \times H^1(\Omega)$ s.t.

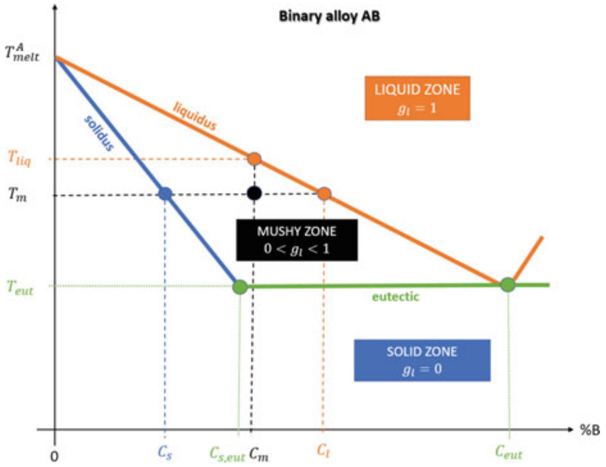


Fig. 1 Example of phase diagram in the case of a binary alloy

$$\begin{aligned}&\partial_t(\rho \underline{U}_m) + \mathrm{div}(\underline{U}_m \otimes \rho \underline{U}_m) - \underline{\mathrm{div}}\, \mu \,\underline{\underline{\mathrm{grad}}}(\underline{U}_m) + \frac{\mu}{K(g_l)}\underline{U}_m + \underline{\mathrm{grad}}(p)\\ &= \rho \underline{g}\left(1-\beta_T(T_m - T_{m,0}) + \beta_C(C_l - C_{l,0})\right),\end{aligned} \tag{1}$$

$$\mathrm{div}(\underline{U}_m) = 0, \tag{2}$$

$$c_p\left(\partial_t(\rho T_m) + \mathrm{div}(\rho \underline{U}_m T_m)\right) - \mathrm{div}\,\lambda_T\,\underline{\mathrm{grad}}(T_m) = -\rho L \partial_t(g_l(T_m, C_m)) + S_{th}, \tag{3}$$

$$\partial_t(\rho C_m) + \mathrm{div}\Big(\frac{1}{g_l + k_p(1-g_l)}C_m \cdot \rho \underline{U}_m\Big) - \mathrm{div}\,\lambda_C\,\underline{\mathrm{grad}}(C_m) = 0, \tag{4}$$

with k_p the partition coefficient and $\underline{g}$ the gravity vector. $K(g_l) := \alpha\frac{g_l^3}{(1-g_l)^2}$ where α is a scaling factor depending on the material. The function $g_l(T_m, C_m)$ is simplified by assuming a linearized liquidus slope (m_l) and the lever rule $C_s = k_p C_l$; see Fig. 1 for an example of phase diagram considered in the case of a binary alloy. From the definition of a mixture variable and the lever rule, one gets: $C_l = \frac{1}{g_l + k_p(1-g_l)}C_m$.

3 Numerical Scheme

The CDO framework [4] is used for the discretization of the system of equations described in Sect. 2. The discretization of the Navier–Stokes equations relies on the CDO face-based scheme as defined in [3, 8]. The degrees of freedom (DoFs) for the velocity are the component-wise mean-values over faces and cells, the mean-values over cells for the pressure. As proved in [8], an *inf-sup* property holds and we thus do not need any stabilization technique. The fully-coupled velocity-pressure coupling results in a saddle-point problem. DoFs for the temperature and solute concentration are also defined as the mean-value over faces and cells. All the algebraic systems to solve are reduced to only the face DoFs thanks to a static condensation technique. In all conservation equations (momentum, energy, solute), the time discretization relies on an implicit Euler time scheme and the advection scheme is centered for the momentum equation and is an upwind scheme for the other equations. Variable properties such as the liquid fraction g_l or the solute concentration in the liquid phase C_l are defined in each cell. The main algorithm to solve the new state of the system at time $t^{n+1} := t^n + \Delta t$ from the state at time t^n is described in Algorithm 1.

Algorithm 1 High-level algorithm

```
Require: $\underline{U}^n$, $p^n$, $C_m^n$, $T_m^n$, $g_l^n$
Initialize the thermo-solutal non-linear algo.: $k = 0$, $r_\infty = 2\epsilon$, $T_m^{n+1,0} \leftarrow T_m^n$ and $C_m^{n+1,0} \leftarrow C_m^n$
while $k < N_{max}$ AND $r_\infty > \epsilon$ do
    $C_m^{n+1,k+1} \leftarrow$ Solve the discrete counterpart of (4)
    Prepare the implicit/explicit source term contribution from $\partial_t g_l := \frac{\partial g_l}{\partial T_m}|_{C_m} \partial_t T_m + \frac{\partial g_l}{\partial C_m}|_{T_m} \partial_t C_m \approx \frac{\partial g_l}{\partial T_m}|_{C_m} \frac{T_m^{n+1,k+1} - T_m^n}{\Delta t} + \frac{\partial g_l}{\partial C_m}|_{T_m} \frac{C_m^{n+1,k+1} - C_m^n}{\Delta t}$
    $T_m^{n+1,k+1} \leftarrow$ Solve the discrete counterpart of (3)
    $g_l^{n+1,k+1} \leftarrow$ Solidification path from the knowledge of $T_m^{n+1,k+1}$ and $C_m^{n+1,k+1}$
    $r_\infty = \max_{c \in Cells} \left( |T_{m,c}^{n+1,k+1} - T_{m,c}^{n+1,k}|/T_0, |C_c^{n+1,k+1} - C_c^{n+1,k}|/C_0 \right)$ and $k \leftarrow k + 1$
end while
Update $C_l^{n+1}$, $K(g_l)$ (forcing term) and the buoyancy source term from the new thermo-solutal state.
$\underline{U}^{n+1}$, $p^{n+1} \leftarrow$ Compute the Navier–Stokes system (Picard algorithm on the convective term)
```

4 Numerical Results

We compare three approaches: (1) the code_saturne CDO approach presented in Sect. 3, (2) the colocated FV approach of code_saturne [7] referred as code_saturne FV in the following and (3) the staggered FV approach of the 2D commercial software SOLID® [10].

In code_saturne FV, the spatial discretization is centered without slope test for the advection, a two-point flux approximation is used for the diffusion operator and the Rhie and Chow filter is employed to prevent checker-board issues as all variables are co-located at the cell centers. A SIMPLEC algorithm is employed with a constant time step. An Euler implicit time scheme is used with up to 10 inner iterations to converge over the Navier–Stokes and thermo-solutal system of equations.

The saddle-point problems arising from the CDO discretization are solved using a generalized conjugate residual (GCR) with a symmetric Gauss-Seidel block preconditioning. SOLID® computations rely on an upwind advection scheme with a PISO-like velocity-pressure coupling. code_saturne CDO and FV approaches use the same solidification/segregation model which presents some simplifications with

Table 1 Test problem data for the adapted *Voller and Prakash* benchmark. The lower part corresponds to values specific to the simplified phase diagram

Specific heat	$c_p = 1$	Latent heat	$L = 5$
Density	$\rho = 1$	Viscosity	$\mu = 1$
Thermal conductivity	$\lambda_T = 0.001$	Thermal coefficient of expansion	$\beta_T = 0.01$
Solutal conductivity	$\lambda_C = 0$	Solutal coefficient of expansion	$\beta_C = 0.01$
Liquidus temperature	$T_{\text{liq}} = 0.1$	Solidus temperature	$T_{\text{sol}} = -0.1$
Initial temperature	$T_0 = 0.5$	Reference temperature	$T_{\text{ref}} = 0.5$
Hot wall temperature	$T_{\text{hot}} = 0.5$	Cold wall temperature	$T_{\text{cold}} = -0.5$
Gravity magnitude	$g = 1000$	Initial solute concentration	$C_0 = 1$
Melting temperature	$T_{\text{melt}} = 0.2$	Eutectic temperature	$T_{\text{eut}} = -0.1$
Liquidus slope	$m_l = -0.1$	Partition coefficient	$k_p = 0.1$

respect to the SOLID® model where the momentum equation is formulated over the liquid phase instead of the mixture and the energy equation is expressed in enthalpy. Computations were performed with code_saturne 8.0.

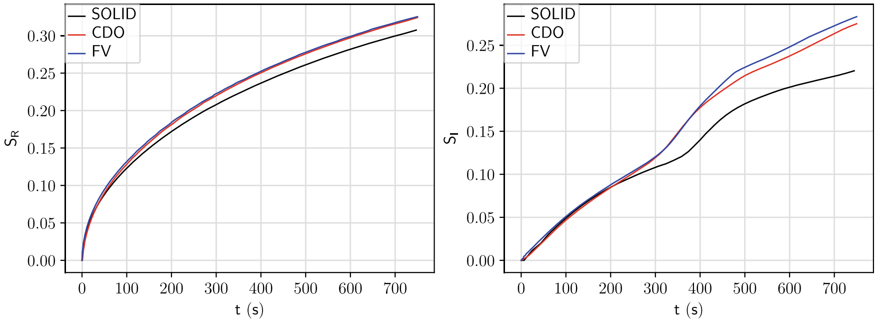
4.1 Case Description

We adapted the *Voller and Prakash*'s benchmark [11] to a segregation problem for a binary alloy (non-dimensionalized properties are listed in Table 1). The geometry is a 2D squared cavity of unit measure with wall boundary conditions. A cold wall condition ($T_m = T_{\text{cold}}$) is applied on the left, and a hot wall ($T_m = T_{\text{hot}}$) on the right. Adiabatic conditions are imposed on the upper and lower parts of the domain. The domain is initially at rest at $T_m = T_0$ and $C_m = C_0$. The simulation ends at 750 s. A uniform Cartesian mesh ($\Delta x = \Delta y = 1/100$) is used.

4.2 Results

Three criteria are used to compare the code_saturne CDO, code_saturne FV and SOLID® approaches:

1. integral quantities as the solidification rate S_R and the segregation index S_I (knowing that $C_0 = 1$)



Solidification rate (S_R) Segregation index (S_I)

Fig. 2 Time evolution of integral quantities

$$S_R := \frac{1}{V_{\text{tot}}} \sum_{i \in N_{\text{cells}}} \left(1 - g_{l,i}\right) V_i, \qquad S_I := \sqrt{\frac{1}{V_{\text{tot}}} \sum_{i \in N_{\text{cells}}} \left(C_{m,i} - 1\right)^2 V_i},$$

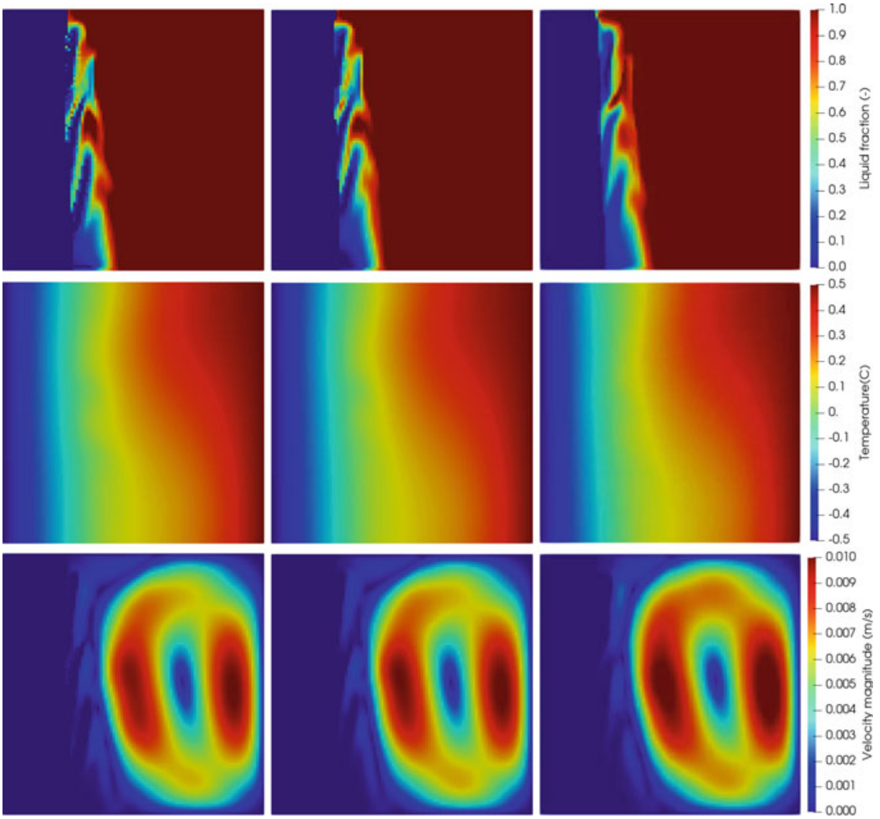
where $V_{\text{tot}} = \sum_{i \in N_{\text{cells}}} V_i$ with V_i the volume of cell i,

2. snapshots at the final time of the velocity, temperature and liquid fraction,
3. profiles at three different heights $y = (0.25, 0.5, 0.75)$ in the domain.

Results. The integral quantities are displayed on Fig. 2 and are similar between the code_saturne CDO and FV computations. The snapshots on Fig. 3 display the fields at the final time of the simulation for the liquid fraction, the temperature and the velocity magnitude. One observes a good visual agreement of the results.

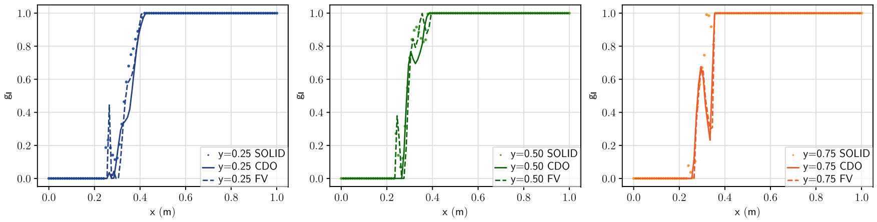
At the final time of the simulation, we plot the profiles of the values of liquid fractions (Fig. 4), of temperatures and velocity magnitudes (Fig. 5) with a comparison to SOLID® simulation results when available. One observes a good agreement of the computed profiles with respect to the SOLID® reference. Some discrepancies of code_saturne CDO and code_saturne FV approaches with respect to SOLID® are likely related to the distinct models used in these solvers.

Analysis. We obtained similar results between code_saturne CDO and code_saturne FV. SOLID® results present some differences as the system of equations is a bit different. Regarding the performances, the choice of the time step was driven by an upstream analysis that compared the error on the maximum velocity with respect to reference values associated to a SOLID® computation on a refined mesh. Values for the time step are 0.001 s for code_saturne FV approach, 1.0 s for code_saturne CDO and 0.01 s for SOLID®. code_saturne CDO appeared as more robust, keeping a good quality of results for large time steps and allowed for faster computations: the CDO computation took 40 min while the FV approach took around 40 h to run. Discrepancies between code_saturne and SOLID® approaches are linked to the difference of model and especially the way to handle the eutectic state. The non-linearity induced by this eutectic state will be the object of further investigations.



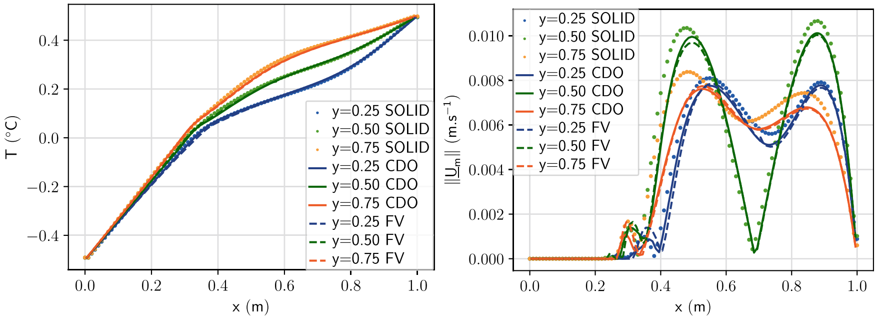
code_saturne FV code_saturne CDO SOLID

Fig. 3 Snapshots at the final simulation time



$y = 0.25$ $y = 0.50$ $y = 0.75$

Fig. 4 Liquid volume fraction profiles at different heights



(Temperature) (Velocity magnitude)

Fig. 5 Profiles at different heights

Acknowledgements The authors would like to thank Martin Ferrand, Charles Demay, Salma Belouah and Antoine Avrit for their valuable contributions to this topic.

References

1. Ahmad, N., Combeau, H., Desbiolles, J.-L., Jalanti, T., Lesoult, G., Rappaz, J., Rappaz, M., Stomp, C: Numerical simulation of macrosegregation: a comparison between finite volume method and finite element method predictions and a confrontation with experiments. Metall. Mater. Trans. A **29A**, 617–630 (1998)
2. Bennon, W.D., Incropera, F.P.: A continuum model for momentum, heat and species transport in binary solid-liquid phase change systems – I. Model formulation. Int. J. Heat Mass Trans. **30**(10), 2161–2170 (1987)
3. Bonelle, J., Ern, A., Milani, R.: Compatible discrete operator schemes for the steady incompressible Stokes and Navier–Stokes equations. In: International Conference on Finite Volumes for Complex Applications, pp. 93–101. Springer (2020)
4. Bonelle, J.: Compatible discrete operator schemes on polyhedral meshes for elliptic and Stokes equations. Ph.D. Thesis, Université Paris-Est (2014)
5. Carman, P.C.: Fluid flow through granular beds. Trans. Inst. Chem. Engrs **15**, 150–166 (1937)
6. Demay., C., Ferrand, M., Belouah, S., Robin, V.: Modelling and simulation of ingot solidification with the open-source software code_saturne. In: IOP Conference Series: Materials Science and Engineering, p. 012033 (2020)
7. EDF R&D: code_saturne. http://www.code-saturne.org/
8. Milani, R.: Compatible discrete operator schemes for the unsteady incompressible Navier–Stokes equations. Ph.D. Thesis, Université Paris-Est (2020)
9. Pickering, E.J.: Macrosegregation in steel ingots: the applicability of modelling and characterisation techniques. ISIJ Int. **53**(6), 935–949 (2013)
10. Sciences Computers Consultants: SOLID®. https://www.scconsultants.com/produits/gamme-solidification.html
11. Voller, V.R., Prakash, C.: A fixed grid numerical modelling methodology for convection-diffusion mushy region phase-change problems. Int. J. Heat Mass Trans. **30**(8), 1709–1719 (1987)

Trefftz Approximation Space for Poisson Equation in Perforated Domains

Miranda Boutilier, Konstantin Brenner, and Victorita Dolean

Abstract For the Poisson equation posed in a planar domain containing a large number of polygonal perforations, we propose a low-dimensional approximation space based on a coarse polygonal partitioning of the domain. Similar to other multi-scale numerical methods, this coarse space is spanned by basis functions that are locally discrete harmonic. We provide an error estimate in the energy norm that only depends on the regularity of the solution over the edges of the coarse skeleton. For a specific edge refinement procedure, this estimate allows us to establish superconvergence of the method, even if the true solution has low general regularity. Combined with the Restricted Additive Schwarz method, the proposed coarse space leads to an efficient two-level iterative linear solver which achieves the fine-scale finite element error in few iterations. The numerical experiment showcases the use of this coarse space over test cases involving singular solutions and realistic urban geometries.

Keywords Multi-scale finite element method · Trefftz method · Domain decomposition

1 Introduction

Let D be an open simply connected polygonal domain in $\mathbb{R}^2$. We denote by $\left(\Omega_{S,k}\right)_k$ a finite family of perforations in D such that each $\Omega_{S,k}$ is an open connected polygonal subdomain of D. The perforations are mutually disjoint such that $\overline{\Omega_{S,k}} \cap \overline{\Omega_{S,l}} = \emptyset$ for any $k \neq l$. We denote $\Omega_S = \bigcup_k \Omega_{S,k}$ and $\Omega = D \setminus \overline{\Omega_S}$, assuming that the family $\left(\Omega_{S,k}\right)_k$ is such that Ω is connected. Note that the latter assumption implies that $\Omega_{S,k}$ are simply connected.

M. Boutilier · K. Brenner (✉) · V. Dolean
LJAD, CNRS, INRIA, Université Côte d'Azur, Nice, France
e-mail: konstantin.brenner@univ-cotedazur.fr

V. Dolean
Department of Mathematics and Statistics, University of Strathclyde, Glasgow, Scotland

E. Franck et al. (eds.), *Finite Volumes for Complex Applications X—Volume 1, Elliptic and Parabolic Problems*, Springer Proceedings in Mathematics & Statistics 432,
https://doi.org/10.1007/978-3-031-40864-9_14

In this contribution we consider the following model problem

$$\begin{cases} -\Delta u = f & \text{in} \quad \Omega, \\ \dfrac{\partial u}{\partial \mathbf{n}} = 0 & \text{on} \quad \partial\Omega \cap \partial\Omega_S, \\ u = 0 & \text{on} \quad \partial\Omega \setminus \partial\Omega_S. \end{cases} \tag{1}$$

Our motivation behind this linear problem lies in the applications to urban flood modeling. In this context, u would represent the flow potential (pressure head) and $(\Omega_{S,k})_k$ can be thought of as a family of impervious structures such as buildings, walls, etc. Although the problem (1) is overly simplified to be directly used for urban hydraulic modeling, the more general nonlinear elliptic or parabolic models are common in free surface flow simulations. Such models arise from Shallow Water equations either by neglecting the inertia terms [1] or within the context of semi-implicit Froude-robust time discretizations [7].

One of the challenges of the numerical modeling of urban floods is that the small structural features may significantly affect the flow. Luckily, modern terrain survey techniques are able to generate high-resolution topographic data. For example, the data set used in this article provided by Métropole Nice Côte d'Azur allows for the infra-metric description of the urban geometries [2]. Depending on the geometrical complexity of the computational domain, the numerical resolution of (1) may become increasingly challenging. In this contribution, we present a numerical strategy involving two levels of space discretization. The first is based on a coarse polygonal partitioning of Ω, while the second is associated with the fine-scale triangulation and is designed to resolve the small scale details of the model domain. With these two levels of space discretization, we introduce a low-dimensional functional space that serves to approximate the locally harmonic component of the solution. This coarse approximation space, called here the Trefftz or discrete Trefftz space, is built upon basis functions that satisfy the local Laplace problems either exactly or numerically and have piecewise linear traces along the edges of the coarse mesh. We note that this coarse space can be readily extended toward higher-order polynomials approximation of the traces. The resulting coarse space can be employed either to approximate the solution over the coarse polygonal grid, or as a component of an iterative two-level domain decomposition (DD) solver for the algebraic problem resulting from the fine-scale discretization. Those are two approaches investigated in this contribution.

When using the Trefftz space to build the coarse approximation method, our methodology is similar to MsFEM [10] as we numerically compute localized basis functions, using fine-scale information in the computation. MsFEM-like methods for elliptic problems in domains containing a large number of perforations or inclusions can be found in [8, 11, 12], with application to both Dirichlet and Neumann conditions on the perforation boundaries. In comparison to these methods, our approach leads to a denser coarse space. In terms of size, our coarse problem is comparable to that obtained from polytopal methods such as the Virtual Element method (VEM) [4]. Due to degrees of freedom associated with the intersection between the perforations and the coarse skeleton, the coarse approximation achieves superconvergence for a specific edge refinement procedure.

As mentioned, our second approach combines the coarse approximation with local subdomain solves in the two-level Restricted Additive Schwarz (RAS) method [6], a DD method that can be used to efficiently solve sparse linear systems. This results in an efficient iterative solver for the algebraic system resulting from the fine-scale finite element (FE) method. The obtained algorithm improves the precision of the original coarse approximation in the spirit of iterative multi-scale methods (see e.g. [9]). We note that alternatively, the discrete Trefftz space can be used within a two-level DD preconditioner for a Krylov method [5].

2 Schur Complement Problem and Trefftz Approximation

We begin with a coarse discretization of Ω which involves a family of polygonal cells $\left(\Omega_j\right)_{j=1,\dots,N}$, the so-called coarse skeleton Γ, and the set of coarse grid nodes that will be referred to by $\mathcal{V}$. The construction is as follows. Consider a finite nonoverlapping polygonal partitioning of D denoted by $\left(D_j\right)_{j=1,\dots,N}$ and an induced nonoverlapping partitioning of Ω denoted by $\left(\Omega_j\right)_{j=1,\dots,N}$ such that $\Omega_j = D_j \cap \Omega$. We will refer to $\left(\Omega_j\right)_{j=1,\dots,N}$ as the coarse mesh over Ω. Additionally, we denote by Γ its skeleton, that is $\Gamma = \bigcup_{j=1,\dots,N} \partial\Omega_j \setminus \partial\Omega_S$.

Based on the coarse partitioning and with the problem (1) in mind, we can split $H^1(\Omega)$ into a direct sum of its subspaces $H^1_\Delta(\Omega)$ and $H^1_\Gamma(\Omega)$. Here, $H^1_\Gamma(\Omega)$ is a subspace of functions vanishing at Γ and $H^1_\Delta(\Omega)$ its H^1-orthogonal complement. The functions from $H^1_\Delta(\Omega)$ are referred to as locally harmonic; note that they also have zero normal traces on $\partial\Omega_S$. Using the orthogonal decomposition introduced above we can express the weak formulation of (1) as the following Schur complement problem: Find $u = u_\Delta + u_b$ with $u_\Delta \in H^1_\Delta(\Omega) \cap H^1_{\partial\Omega\setminus\partial\Omega_S}$ and $u_b \in H^1_\Gamma(\Omega)$ satisfying

$$\begin{cases} (u_\Delta, v)_{H^1(\Omega)} = (f, v)_{L^2(\Omega)} & \forall v \in H^1_\Delta(\Omega) \cap H^1_{\partial\Omega\setminus\partial\Omega_S}(\Omega), \\ (u_b, v)_{H^1(\Omega)} = (f, v)_{L^2(\Omega)} & \forall v \in H^1_\Gamma(\Omega). \end{cases} \tag{2}$$

We remark that the local "bubble" component of the solution u_b can be computed from (2) locally (and in parallel) on each Ω_j, while the problem for u_Δ is globally coupled over Ω.

We now proceed with the approximation of the locally harmonic component u_Δ. For this, we introduce the Trefftz coarse space, a finite-dimensional subspace V_H of $H^1_\Delta(\Omega)$ that is spanned by the functions that are piecewise linear on the skeleton Γ. Let $(e_k)_{k=1,\dots,N_e}$ denote a nonoverlapping partitioning of Γ such that each "coarse edge" e_k is an open planar segment, and we denote $H = max_{k=1,\dots,N_e}|e_k|$. We note that a straight segment of Γ may be subdivided into multiple edges (see Fig. 2). The set of coarse grid nodes is given by $\mathcal{V} = \bigcup_{k=1,\dots,N_e} \partial e_k$.

The coarse nodal basis is defined by the following set of boundary value problems. For all Ω_j and for all $s = 1, \ldots, N_{\mathcal{V}}$, find $\phi_s^j \in H^1(\Omega_j)$ such that ϕ_s^j is the weak solution to the following problem

$$\begin{cases} \Delta\phi_s^j = 0 \ \text{ in } \ \Omega_j, \\ \dfrac{\partial\phi_s^j}{\partial\mathbf{n}_j} = 0 \ \text{ on } \partial\Omega_j \cap \partial\Omega_S, \\ \phi_s^j = g_s \ \text{on } \partial\Omega_j \setminus \partial\Omega_S, \end{cases} \tag{3}$$

where g_s is a skeleton "hat" basis function such that $g_s : \Gamma \to \mathbb{R}$ is continuous on Γ, linear on each edge e_k and satisfies $g(\mathbf{x}_i) = \delta_{is}$ for all nodes $\mathbf{x}_i$.

Let $V_{H,0} = V_H \cap H^1_{\partial\Omega\setminus\partial\Omega_S}(\Omega)$. The Galerkin method based on the coarse space reads as follows: find $u_H \in V_{H,0}$ such that

$$(u_H, v_H)_{H^1(\Omega)} = (f, v_H)_{L^2(\Omega)} \qquad \forall v_H \in V_{H,0}. \tag{4}$$

Combining best approximation property of the coarse approximation u_H, classical FE interpolation theory in one dimension, and an interpolation inequality we obtain the following error estimate for the approximation of u_Δ.

Proposition 1 *Assume that there exists a finite nonoverlapping partitioning $(\gamma_l)_{l=1,\ldots,N_\gamma}$ of Γ such that the traces of u belong to $H^2(\gamma_l)$ for all l; assume in addition that the set of coarse edges $(e_k)_k$ is a subdivision of $(\gamma_l)_l$. There exists $C > 0$ depending only on $(D_j)_j$ such that*

$$\|\nabla(u_\Delta - u_H)\|_{L^2(\Omega)} \le CH^{\frac{3}{2}} \left(\sum_{l=1}^{N_\gamma} \|u\|^2_{H^2(\gamma_l)} \right)^{\frac{1}{2}}. \tag{5}$$

We remark that the broken H^2 norm in the right-hand side of (5) involves only the traces of the solution along the sections of the coarse skeleton. Therefore, the estimate is valid for u having low general regularity that is due, for example, to corner singularities; moreover the constant involved in the estimate is independent of the shape and distribution of the perforations. The estimate (5) provides an a priori criterion for the adaptation of the coarse mesh: one has to ensure that the edge norm in the right-hand side is small. For f regular enough, this can be achieved by moving the coarse edges away from the "bad" perforation corners. We further note that this estimate is especially valuable for a so-called space or edge refinement, which is a procedure that involves splitting the edges of an otherwise fixed coarse grid. In that case, one observes the superconvergence of the error with a rate of $3/2$.

3 Discrete Trefftz Space and Two-Level Schwarz Method

In this section, we introduce the two-level RAS method based on the discrete Trefftz method. Let us consider the triangulation of Ω which is assumed to be conforming with respect to the polygonal partitioning $\left(\Omega_j\right)_{j=1,\dots,N}$ (see Fig. 1). We denote by V_h the space of piecewise linear continuous function over this triangulation. The associated fine-scale FE discretization of (1) results in the linear system $\mathbf{Au} = \mathbf{f}$. Because the triangular mesh resolves the the perforations, the latter system may be quite large; moreover the size of the triangular elements may vary by several orders of magnitude. As a result, the matrix $\mathbf{A}$ is expected to be poorly conditioned.

Let us begin with the coarse space component. In most practical situations, the coarse basis functions defined by (3) cannot be computed analytically. Therefore, we consider the FE approximation of V_H denoted by $V_{H,h}$. The basis of the discrete Trefftz space is obtained through the FE approximation of (3). Let $\mathbf{R}_H$ be a transition matrix from the discrete Trefftz basis of $V_{H,h}$ toward the standard FE nodal basis of V_h. The FE counterpart of (4) can be expressed algebraically as $\mathbf{u}_H = M_H^{-1}\mathbf{f}$, where $M_H^{-1} = \mathbf{R}_H^T(\mathbf{R}_H\mathbf{A}\mathbf{R}_H^T)^{-1}\mathbf{R}_H$.

Below, we will show how the coarse space introduced in the previous section can be combined with RAS to construct a simple yet efficient iterative linear solver for the fine-scale finite element method. Let $\left(\Omega'_j\right)_{j=1,\dots,N}$ denote the overlapping partitioning of Ω such that $\Omega_j \subset \Omega'_j$. In practice, each Ω'_j is constructed by propagating Ω_j by a few layers of triangles. Consider classical boolean restriction matrices $\mathbf{R}'_j$ associated to the family $\left(\Omega'_j\right)_{j=1,\dots,N}$. The iterative procedure is given by

$$\begin{aligned} \mathbf{u}^{n+\frac{1}{2}} &= \mathbf{u}^n + M_{RAS}^{-1}(\mathbf{f} - \mathbf{A}\mathbf{u}^n) \\ \mathbf{u}^{n+1} &= \mathbf{u}^{n+\frac{1}{2}} + M_H^{-1}(\mathbf{f} - \mathbf{A}\mathbf{u}^{n+\frac{1}{2}}) \end{aligned} \tag{6}$$

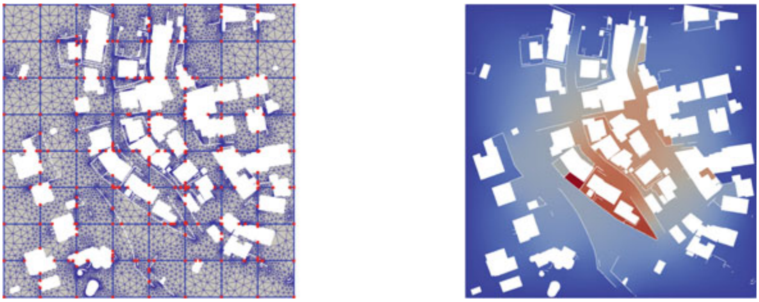


Fig. 1 Left: coarse (thick lines) and fine (thin lines) discretizations, with the coarse nodes shown by red dots. Right: FE solution with $f = 1$

where $M_{RAS}^{-1} = \sum_{j=1}^{N} \left(\mathbf{R}'_j\right)^T \mathbf{D}_j (\mathbf{A}'_j)^{-1} \mathbf{R}'_j$ and $\mathbf{A}'_j = \mathbf{R}'_j \mathbf{A} \left(\mathbf{R}'_j\right)^T$, while $\mathbf{D}_j$ denote the partition-of-unity matrices. We note that this iterative two-level RAS iteration is classical in domain decomposition literature for a general coarse matrix M_H^{-1}. As well, the local matrix solves and assembly of restriction and extension matrices can be done in parallel, forming an efficient algorithm for sparse matrices $\mathbf{A}$.

4 Numerical Results

In this section, we illustrate the performance of the discrete Trefftz space within two different scenarios involving either a standalone Galerkin approximation (4) or an iterative approach (6). In particular, we will provide numerical evidence of the error estimate (5) over the case involving a solution with a corner singularity. The numerical investigation of the iterative algorithm (6) shows that the fine-scale FE error can be achieved in few iterations.

Convergence of the discrete Trefftz approximation method: We begin with a test case involving a classical L-shaped domain with a reentering corner (Fig. 2). The domain is defined by $D = (-1, 1)^2$, $\Omega_S = (0, 1)^2$, and $\Omega = D \setminus \overline{\Omega_S}$. We consider the problem (1) with a zero right-hand side and a non-homogeneous Dirichlet boundary condition on $\partial\Omega \setminus \partial\Omega_S$ provided by the singular exact solution $u(r, \theta) = r^{\frac{2}{3}} \cos(\frac{2}{3}(\theta - \pi/2))$.

In order to assess the convergence of the discrete Trefftz method, we consider two strategies regarding the refinement of the coarse partitioning. The procedure involving the reduction of the diameter of the coarse cells will be referred to as *mesh refinement*. The sequence of such meshes will be constructed as follows: first the background domain D is partitioned into $N = (2p + 1)^2$, $p \in \mathbb{N}$, squares, then the coarse cells Ω_j are generated by excluding Ω_S. The choice of N being a square of an odd number ensures the consistency of the mesh sequence in terms of the shape of the

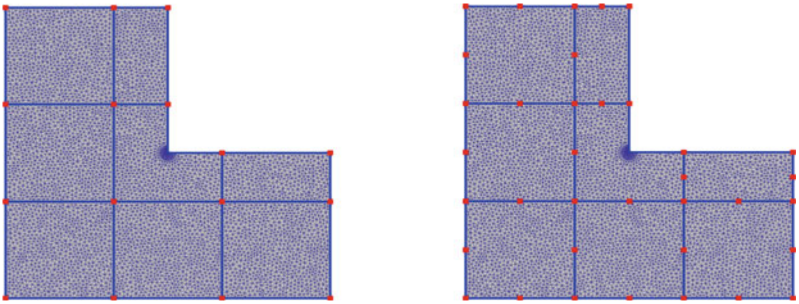


Fig. 2 Coarse and fine discretizations of the L-shaped domain with no (left) and one (right) additional degree(s) of edge refinement

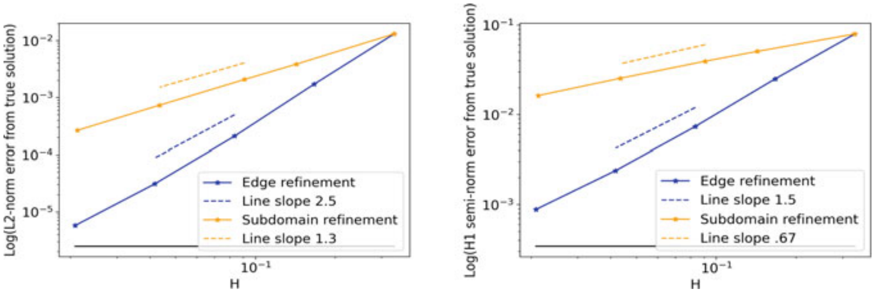


Fig. 3 Coarse approximation error for L-shaped domain with edge (blue) and mesh (orange) refinement in L^2 norm (left) and the energy norm (right)

elements. Alternatively, we consider the *edge refinement* procedure, which accounts for subdividing the edges of an original "3 × 3 grid". This edge refinement approach is illustrated by Fig. 2 and is inspired by the Multiscale Hybrid-Mixed method [3], where the superconvergence in the energy norm has equally been observed. Let us stress that under these refinement procedures, none of the coarse grids will have degree of freedom located at the corner (0, 0). As a result, the corner singularity will be captured by the basis functions associated with the L-shaped domain. We also note that in order to improve the precision of the fine-scale FE method, the size of the triangles is graded in the vicinity of the corner (0, 0).

Figure 3 reports the error in L^2 and the energy norms as functions of maximal coarse edge length H. The black horizontal line represents the typical fine-scale FE error. As expected, we observe that the convergence of the discrete Trefftz method deteriorates as the coarse error approaches this value. In accordance with the error estimate (5), for the edge refinement, we observe superconvergence in the energy norm with the rate slightly superior to $3/2$. We note that that the convergence rate in L^2 appears to be superior to $5/2$. In contrast, the convergence resulting from the mesh refinement seems to be controlled by the low global regularity of the solution. For the mesh refinement, we observe the convergence rates typical for FE methods on quasi-uniform meshes.

Convergence of the two-level iterative method: Next, we examine the performance of the iterative scheme (6) over the L-shaped domain considered previously and a domain based on realistic urban geometries for which the datasets were kindly provided by Métropole Nice Côte d'Azur. The domain involving a small portion of the structural topography of the city of Nice is shown in Fig. 1. The dataset contains two kinds of structural elements, namely buildings (and assimilated small elevated structures) and walls.

We report on Fig. 4 the convergence history of the iterative method for both L-shaped and urban domains; more precisely, we report the convergence of the full L^2 error, that is the norm of the difference between the intermediate approximation and some accurate solution of (1). Each figure reports the convergence history for linear systems based on increasingly refined background triangulations. The typical

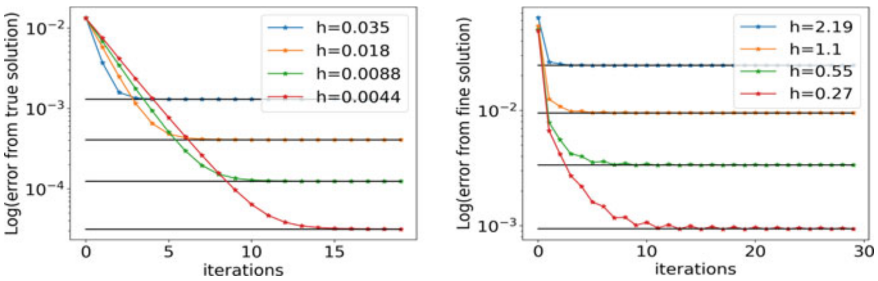


Fig. 4 Convergence of the two-level iterative method for the L-shaped domain on 3 × 3 subdomains (left) and the urban dataset on 8 × 8 subdomains (right). The black horizontal lines show the error of the fine-scale FE method

width of the overlap in RAS method is of $\mathrm{diam}(\Omega_j)/20$. For the realistic data set, we solve (1) with $f = 1$; the fine-scale FE solution is reported on Fig. 1. As the exact solution is not available for the case based on the urban data, we use a reference solution obtained on a very fine grid.

We observe that the error of the fine-scale FE method (black lines) can be reached relatively fast. Further iterations do not improve the overall precision of the approximate solution even though the algebraic error may decrease. For the L-shaped domain, the convergence of the full error is essentially exponential; moreover the decay rate of the full error remains consistent with respect to the finite size h of the fine-scale triangulation.

Acknowledgements This work has been supported by ANR Project Top-up (ANR-20-CE46-0005). The high-resolution structural data has been provided by Métropole Nice Côte d'Azur. We warmly thank Florient Largeron, chief of MNCA's SIG 3D project, for his help in preparation of the data and for the multiple fruitful discussions.

References

1. Alonso, R., Santillana, M., Dawson, C.: On the diffusive wave approximation of the shallow water equations. Eur. J. Appl. Math. **19**(5), 575–606 (2008)
2. Andres, L.: L'apport de la donnée topographique pour la modélisation 3D fine et classifiée d'un territoire. Rev. XYZ **133**(4), 24–30 (2012)
3. Araya, R., Harder, C., Paredes, D., Valentin, F.: Multiscale hybrid-mixed method. SIAM J. Numer. Anal. **51**(6), 3505–3531 (2013)
4. Beirão da Veiga, L., Brezzi, F., Cangiani, A., Manzini, G., Marini, L.D., Russo, A.: Basic principles of virtual element methods. Math. Model. Method. Appl. Sci. **23**(01), 199–214 (2013)
5. Boutilier, M., Brenner, K., Dolean, V.: A trefftz-like coarse space for the two-level Schwarz method on perforated domains (2022). arXiv:2211.05880
6. Cai, X.C., Sarkis, M.: A restricted additive schwarz preconditioner for general sparse linear systems. Siam J. Sci. Comput. **21**(2), 792–797 (1999)
7. Casulli, V.: Semi-implicit finite difference methods for the two-dimensional shallow water equations. J. Comput. Phys. **86**(1), 56–74 (1990)

8. Chu, C.C., Graham, I., Hou, T.Y.: A new multiscale finite element method for high-contrast elliptic interface problems. Math. Comput. **79**(272), 1915–1955 (2010)
9. Hajibeygi, H., Bonfigli, G., Hesse, M.A., Jenny, P.: Iterative multiscale finite-volume method. J. Comput. Phys. **19**(277), 8604–8621 (2008)
10. Hou, T.Y., Wu, X.H.: A multiscale finite element method for elliptic problems in composite materials and porous media. J. Comput. Phys. **134**(1), 169–189 (1997)
11. Le Bris, C., Legoll, F., Lozinski, A.: An MsFEM type approach for perforated domains. Multiscale Model. Simul. **12**(3), 1046–1077 (2014)
12. Le Bris, C., Legoll, F., Madiot, F.: Multiscale finite element methods for advection-dominated problems in perforated domains. Multiscale Model. & Simul. **17**(2), 773–825 (2019)

Structure Preserving Finite Volume Approximation of Cross-Diffusion Systems Coupled by a Free Interface

Clément Cancès, Jean Cauvin-Vila, Claire Chainais-Hillairet, and Virginie Ehrlacher

Abstract We propose a two-point flux approximation finite-volume scheme for the approximation of two cross-diffusion systems coupled by a free interface to account for one-dimensional vapor deposition. The moving interface is addressed with a cut-cell approach, where the mesh is locally deformed around the interface. The scheme preserves the structure of the continuous system, namely: mass conservation, nonnegativity, volume-filling constraints and decay of the free energy. Numerical results illustrate the properties of the scheme.

Keywords Cross-diffusion system · Cut-cell method · Finite volume scheme · Free energy dissipation · Moving interface · Vapor deposition

1 A Free Interface Cross-Diffusion Model

We address a toy model to describe a physical vapor deposition process used for the fabrication of thin film layers [3]. We consider the evolving domain

$$\Omega(t) = (0, X(t)) \cup (X(t), 1), \; t > 0,$$

where $\mathbb{R}_+ \ni t \to X(t) \in [0, 1]$ is the free interface between the solid (left) and the gas (right). Traces and jumps at the interface are respectively denoted by f^s, f^g, $[[f]] = f^g - f^s$. We consider n different chemical species represented by their densities of molar concentration $\boldsymbol{c} = (c_1, \dots, c_n)^T$. The local conservation of matter reads:

$$\partial_t \boldsymbol{c} + \partial_x \boldsymbol{J} = 0, \; t > 0, \; x \in \Omega(t), \tag{1a}$$

C. Cancès · C. Chainais-Hillairet
UMR 8524 - Laboratoire Paul Painlevé, CNRS, Inria, University of Lille, 59000 Lille, France

J. Cauvin-Vila (✉) · V. Ehrlacher
CERMICS (ENPC) & INRIA, Paris, France
e-mail: jean.cauvin-vila@enpc.fr

E. Franck et al. (eds.), *Finite Volumes for Complex Applications X—Volume 1, Elliptic and Parabolic Problems*, Springer Proceedings in Mathematics & Statistics 432,
https://doi.org/10.1007/978-3-031-40864-9_15

for some molar fluxes $\boldsymbol{J} := (J_1, \dots, J_n)^T$. Cross-diffusion phenomena are modelled differently in each phase. In the solid phase, the fluxes are given by

$$J_i = -\sum_{j=1}^{n} \kappa_{ij}^s \left(c_j \partial_x c_i - c_i \partial_x c_j \right), \text{ in } (0, X), \; i \in \{1, \dots, n\},$$

with cross-diffusion coefficients $\kappa_{ij}^s = \kappa_{ji}^s > 0$, which rewrite more compactly

$$\boldsymbol{J} = -\boldsymbol{A}_s(\boldsymbol{c})\partial_x \boldsymbol{c}, \text{ in } (0, X), \tag{1b}$$

with a linear diffusion matrix $\boldsymbol{A}_s(\boldsymbol{c})$ (see [1]). In the gaseous phase, the fluxes are defined implicitly via the Maxwell–Stefan linear system (see [2])

$$\boldsymbol{A}_g(\boldsymbol{c})\boldsymbol{J} = -\partial_x \boldsymbol{c}, \text{ and } \sum_{i=1}^{n} J_i = 0, \text{ in } (X, 1), \tag{1c}$$

where $\boldsymbol{A}_g(\boldsymbol{c})$ is identical to $\boldsymbol{A}_s(\boldsymbol{c})$, except for possibly different cross-diffusion coefficients $\kappa_{ij}^g = \kappa_{ji}^g > 0$. The system is completed with an initial condition $(\boldsymbol{c}^0, X^0)$, no-flux conditions on the fixed boundary and the following conditions across the moving interface:

$$\boldsymbol{J}^s(t) - X'(t)\boldsymbol{c}^s(t) = \mathbb{1}_{\{X(t)\in(0,1)\}}\boldsymbol{F}(t) = \boldsymbol{J}^g(t) - X'(t)\boldsymbol{c}^g(t), \; t > 0, \tag{1d}$$

where $\boldsymbol{F}$ accounts for reaction mechanisms [6, 7] and is defined, for some constant reference chemical potentials $\mu_i^{*,s}, \mu_i^{*,g} \in \mathbb{R}$, by the Butler–Volmer formulas: for $i \in \{1, \dots, n\}$,

$$\begin{aligned} F_i &= c_i^s \exp\left(\frac{\mu_i^{*,g} - \mu_i^{*,s}}{2}\right) - c_i^g \exp\left(\frac{\mu_i^{*,s} - \mu_i^{*,g}}{2}\right), \\ &= 2\sqrt{c_i^s c_i^g} \sinh\left(-\frac{1}{2}[[\log(c_i) - \mu_i^*]]\right). \end{aligned} \tag{1e}$$

Finally, the interface evolves according to

$$X'(t) = -\mathbb{1}_{\{X(t)\in(0,1)\}} \sum_{i=1}^{n} F_i. \tag{1f}$$

Note that, in the limit cases $X(t) = 0$ or $X(t) = 1$, (1d)–(1f) imply that we recover a single phase problem with zero-flux boundary conditions. The system enjoys several important properties we aim at preserving at the discrete level: First, mass conservation follows from the local conservation (1a), no-flux conditions on the fixed boundary and the conservative condition (1d). Second, the system preserves the nonnegativity of the concentrations and the volume-filling constraints $\sum_{i=1}^{n} c_i = 1$

(satisfied by the initial condition), and we refer to such a solution as *admissible*. Finally, the functional

$$\mathcal{H}(\boldsymbol{c}, X) = \int_0^X h_s(\boldsymbol{c}) + \int_X^1 h_g(\boldsymbol{c}), \tag{2}$$

with density $h_\alpha(\boldsymbol{c}) = \sum_{i=1}^n c_i(\log(c_i) - \mu_i^{*,\alpha}) - c_i + 1$, for $\alpha \in \{s, g\}$, can be shown to formally satisfy, for some positive semi-definite mobility matrices $\boldsymbol{M}_s, \boldsymbol{M}_g$, the free energy dissipation relation [4, 5]

$$\begin{aligned} \frac{d}{dt}\mathcal{H}(\boldsymbol{c}(t), X(t)) &= -\int_0^{X(t)} \partial_x \log(\boldsymbol{c})^T \boldsymbol{M}_s(\boldsymbol{c}) \partial_x \log(\boldsymbol{c}) \\ &- \int_{X(t)}^1 \partial_x \log(\boldsymbol{c})^T \boldsymbol{M}_g(\boldsymbol{c}) \partial_x \log(\boldsymbol{c}) + \boldsymbol{F}(t)^T [[\log(\boldsymbol{c}) - \boldsymbol{\mu}^*]] \leq 0. \end{aligned} \tag{3}$$

One deduces from the dissipation inequality that stationary solutions $(\bar{\boldsymbol{c}}, \bar{X})$ must be constant in (each connected part of) $\bar{\Omega} = (0, \bar{X}) \cup (\bar{X}, 1)$ and moreover, if $\bar{X} \in (0, 1)$, $F_i(\bar{c}_i^s, \bar{c}_i^g) = 0$ should hold for any i. We characterize in [3] the stationary states of (1), as partially stated in Proposition 1.

Proposition 1 (Stationary states) *Let* $\boldsymbol{m}^0 = \boldsymbol{m}^{0,s} + \boldsymbol{m}^{0,g} > 0$ *be the initial amount of matter in the system. The one-phase solutions* $(\boldsymbol{m}^0, 0, 1)$ *and* $(0, \boldsymbol{m}^0, 0)$ *are stationary. Define the coefficients* $\beta_i = \exp\left([[\mu_i^*]]\right)$. *There exists a stationary solution where the two phases coexist (*i.e. *such that* $\bar{X} \in (0, 1)$*) if and only if*

$$\min\left(\sum_{i=1}^n m_i^0 \beta_i, \sum_{i=1}^n m_i^0 \frac{1}{\beta_i}\right) > 1. \tag{4}$$

Moreover, under (4)*, this stationary state is unique and explicitly computable from* $\bar{X}$*, which is itself solution to a convex scalar equation.*

Let us remark that, under condition (4), one-phase stationary states are not expected to be stable.

2 Finite Volume Scheme

We consider $N \in \mathbb{N}^*$ reference cells of uniform size $\Delta x = \frac{1}{N}$. The $N + 1$ edge vertices are denoted by $0 = x_{\frac{1}{2}}, x_{\frac{3}{2}}, \ldots, x_{N+\frac{1}{2}} = 1$. We consider a time horizon $T > 0$ and a time discretization with mesh parameter Δt defined such that $N_T \Delta t = T$ with $N_T \in \mathbb{N}^*$. The concentrations are discretized as $\boldsymbol{c}_{\Delta x}^p = (c_{i,K}^p)_{i\in\{1,\ldots,n\},\ K\in\{1,\ldots,N\}}$ for $p \in \{0, \ldots, N_T\}$. The interface is discretized in time as X^p for $p \in \{0, \ldots, N_T\}$, and we denote by $x_{K^p+\frac{1}{2}} \in [0, 1]$ the closest vertex to X^p (the left vertex in case of equality). At time $t^{p-1} = (p-1)\Delta t$, the mesh is locally modified around X^{p-1}: the cells K^{p-1} and $K^{p-1}+1$ are deformed, as presented initially in Fig. 1, where we denote by K the interface cell to alleviate the notations. To account for this deformation, we introduce Δ_K^{p-1} the size of cell K at discrete time t^{p-1}:

$$\Delta_K^{p-1} = \begin{cases} (X^{p-1} - x_{K^{p-1}-\frac{1}{2}}) & \text{if } K = K^{p-1}, \\ (x_{K^{p-1}+\frac{3}{2}} - X^{p-1}) & \text{if } K = K^{p-1}+1, \\ \Delta x & \text{otherwise.} \end{cases}$$

With this notation, the initial concentrations $\boldsymbol{c}^0 \in L^\infty(\Omega_0; \mathbb{A})$ are naturally discretized as $c_{i,K}^0 = \frac{1}{\Delta_K^0}\int_K c_i^0 \, dx$. Starting from the knowledge of $\boldsymbol{c}_{\Delta x}^{p-1}$, X^{p-1}, our scheme consists in

(i) solving the conservation laws and updating the interface position, leading to $(\boldsymbol{c}_{\Delta x}^{p,\star}, X^p)$.
(ii) updating the mesh to Δ_K^p and post-processing the interface concentrations into the final values $\boldsymbol{c}_{\Delta x}^p$.

2.1 Conservation Laws

The conservation laws (1a) are discretized implicitly as, for $K \in \{1, \dots, N\}, i \in \{1, \dots, n\}$,

$$\frac{1}{\Delta t}(\Delta_K^{p,\star} c_{i,K}^{p,\star} - \Delta_K^{p-1} c_{i,K}^{p-1}) + J_{i,K+\frac{1}{2}}^{p,\star} - J_{i,K-\frac{1}{2}}^{p,\star} = 0. \tag{5a}$$

where we have introduced the intermediate quantity (see the intermediate mesh in Fig. 1)

$$\Delta_K^{p,\star} = \begin{cases} (X^p - x_{K^{p-1}-\frac{1}{2}}) & \text{if } K = K^{p-1}, \\ (x_{K^{p-1}+\frac{3}{2}} - X^p) & \text{if } K = K^{p-1}+1, \\ \Delta x & \text{otherwise.} \end{cases}$$

The bulk fluxes (1b)–(1c) are discretized in a way that preserves the bulk part of the dissipation structure (3). We refer to [4] (resp. [5]) for the discretization of (1b) (resp. (1c)) in a single-phase and fixed domain context, since we prefer to highlight our contribution to the treatment of the interface coupling. Because of the moving interface, a correction term $-X'(t)\boldsymbol{c}$ appears in (5a) in the interface cells, see (1d), and the numerical interface fluxes are given by a discretization of (1e) as

$$J_{i,K^{p-1}+\frac{1}{2}}^{p,\star} = F_i^{p,\star} = c_{i,K^{p-1}}^{p,\star} \exp\left(\frac{\mu_i^{*,g} - \mu_i^{*,s}}{2}\right) - c_{i,(K^{p-1}+1)}^{p,\star} \exp\left(\frac{\mu_i^{*,s} - \mu_i^{*,g}}{2}\right),$$

Finally, (1f) is discretized as

$$X^p = X^{p-1} - \Delta t \sum_{i=1}^{n} F_i^{p,\star}. \tag{5b}$$

We denote a solution to (5) by $(\boldsymbol{c}_{\Delta x}^{p,\star}, X^p)$.

2.2 Post-processing

When X^p crosses the center of a cell, one needs to update the interface cell from K^{p-1} to K^p and to adjust the concentrations accordingly. First, we can derive from (5b) a linear CFL condition to enforce $|X^p - X^{p-1}| \leq \frac{\Delta x}{2}$, which in particular ensures that $|K^p - K^{p-1}| \leq 1$ and simplifies the post-processing process (X^p cannot cross $x_{K+\frac{3}{2}}$ in Fig. 1). If $K^p = K^{p-1}$, then we can directly iterate the scheme with $\boldsymbol{c}_{\Delta x}^{p} = \boldsymbol{c}_{\Delta x}^{p,\star}$. Otherwise, let us illustrate the case of a right displacement $K^p = K^{p-1} + 1$ and let us use again the notation $K := K^{p-1}$ for simplicity. We perform the following steps (see the final mesh in Fig. 1)

(i) *Projection:* The value $c_{i,K}^{p,\star}$ is assigned to the virtual cell $(x_{K-\frac{1}{2}}, X^p)$. We assign this value to both the fixed cell $K = (x_{K-\frac{1}{2}}, x_{K+\frac{3}{2}})$ and the new interface cell $(K+1) = (x_{K+\frac{1}{2}}, X^p)$:

$$c_{i,K}^{p} = c_{i,K+1}^{p} = c_{i,K}^{p,\star}. \tag{6}$$

(ii) *Average:* X^p replaces x_{K+1} as the interface node. We average the value in the cell $(K+2) = (X^p, x_{K+2})$:

$$c_{i,K+2}^{p} = \frac{1}{\Delta x + \Delta_{K+1}^{p,\star}} \left[\Delta_{K+1}^{p,\star} c_{i,K+1}^{p,\star} + \Delta x \; c_{i,K+2}^{p,\star}\right]. \tag{7}$$

(iii) In all other cells, $c_{i,K}^{p} = c_{i,K}^{p,\star}$.

2.3 Numerical Analysis

Let us introduce the discrete version of the free energy functional (2):

$$\mathcal{H}^p(\boldsymbol{c}_{\Delta x}^{p}, X^p) = \sum_{i=1}^{n} \sum_{K \leq K^p} \Delta_K^p h^s(c_{i,K}^p) + \sum_{i=1}^{n} \sum_{K \geq K^p+1} \Delta_K^p h^g(c_{i,K}^p). \tag{8}$$

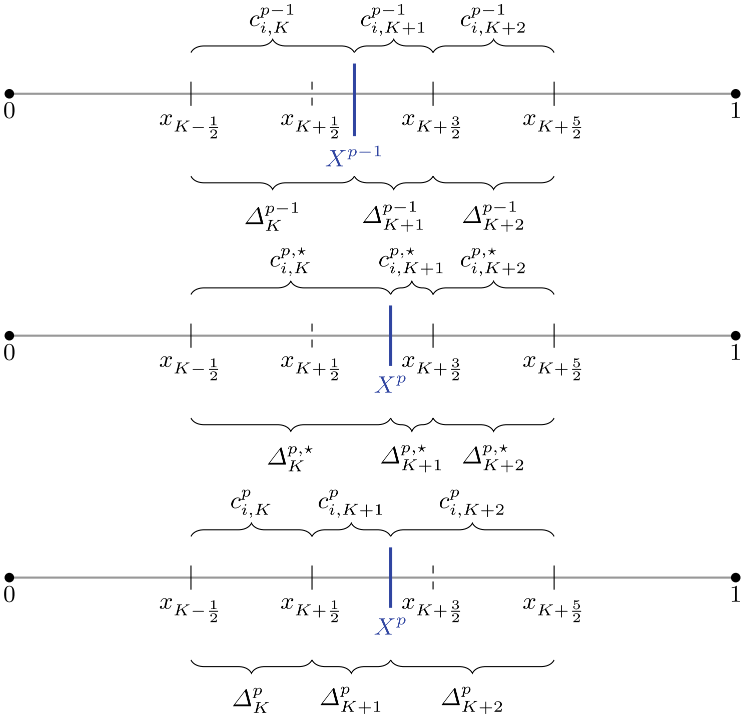


Fig. 1 A virtual mesh displacement between $t^{p-1} = (p-1)\Delta t$ and $t^p = p\Delta t$

Proposition 2 gives some a priori estimates fulfilled by a solution to the scheme, leading to existence of a solution.

Proposition 2 (Structure preservation) *Given an admissible solution* $(\boldsymbol{c}_{\Delta x}^{p-1}, X^{p-1})$, *there exists an admissible solution* $(\boldsymbol{c}_{\Delta x}^{p}, X^p)$ *to the scheme (5). Moreover, the amount of matter of each species is conserved and a discrete version of the dissipation relation* (3) *is satisfied:*

$$\boldsymbol{c}_{\Delta x}^{p} \geq 0, \; \textit{and} \; \sum_{i=1}^{n} c_{i,K}^{p} = 1, \; K \in \{1, \ldots, N\},$$
$$\sum_{K=1}^{N} \Delta_K^p c_{i,K}^p = m_i^0, \; i \in \{1, \ldots, n\},$$
$$\mathcal{H}^p(\boldsymbol{c}_{\Delta x}^{p}, X^p) \leq \mathcal{H}^p(\boldsymbol{c}_{\Delta x}^{p-1}, X^{p-1}).$$

We sketch some ingredients of the proof below, see [3] for details.

Proof Concerning conservation of matter, it follows from summing the conservation laws (5a) over the cells K and the fact that the fluxes are conservative that, for any $i \in \{1, \ldots, n\}$,

$$\sum_{K=1}^{N} \Delta_K^{p,\star} c_{i,K}^{p,\star} = \sum_{K=1}^{N} \Delta_K^{p-1} c_{i,K}^{p-1}.$$

If $K^p = K^{p-1}$, the result follows immediately. Otherwise, it follows by construction of the post-processing formulas (6)–(7).

The proof of the nonnegativity of the concentrations follows from a contradiction argument with an appropriate truncation of the fluxes. One even obtains strict positivity if $\boldsymbol{c}_{\Delta x}^{p-1} > 0$.

The volume-filling constraints are proved by summing the conservation laws (5a) over i and using a normalized version of (5b).

Thanks to strict positivity, a chain rule holds [4, 5] and the continuous dissipative structure (3) can be translated at the discrete level. Besides, convexity implies that the post-processing (6)–(7) cannot make the free energy increase.

Finally, the existence proof follows from a topological degree argument, arguing by deformation to two independent one-phase systems in fixed domains.

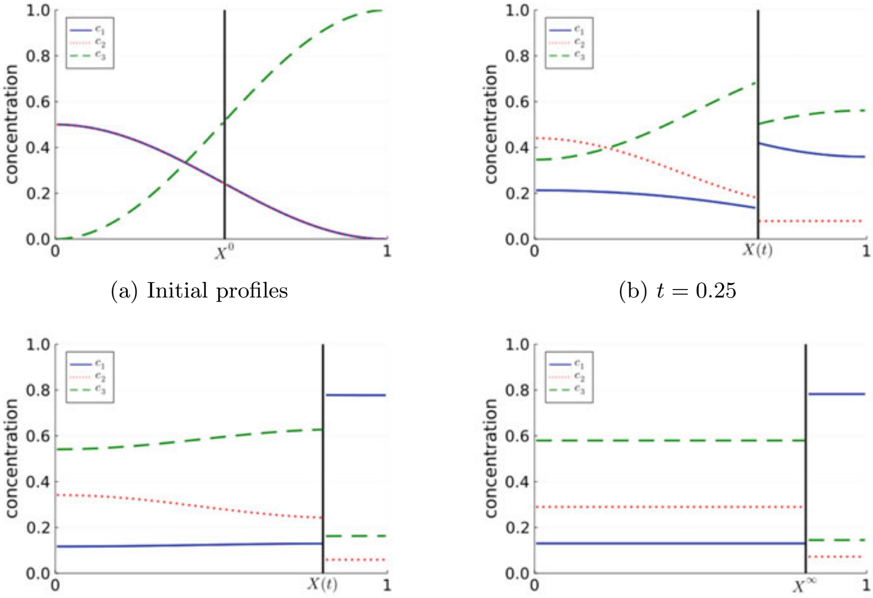
3 Numerical Results

The numerical scheme has been implemented in the Julia language. The nonlinear system is solved with Newton method and stopping criterion $\|Res\|_{l^2(\Delta x)} \leq 10^{-12}$, where Res is the residual of the scheme. Jacobians are efficiently automatically computed thanks to the ForwardDiff and SparseDiffTools packages.

Let us introduce a test case: we fix an initial interface $X^0 = 0.51$ and smooth initial concentrations $c_1^0(x) = c_2^0(x) = \frac{1}{4}\,(1 + \cos(\pi x))\,,\ c_3^0(x) = \frac{1}{2}\,(1 - \cos(\pi x))$ that will be suitably discretized. The cross-diffusion coefficients are taken equal in each phase, with values $\kappa_{12} = \kappa_{21} = 0.2,\ \kappa_{23} = \kappa_{32} = 0.1,\ \kappa_{13} = \kappa_{31} = 1$ (diagonal coefficients do not play any role). The reference chemical potential $\boldsymbol{\mu}^{*,s}$, $\boldsymbol{\mu}^{*,g}$ are given by $e^{\boldsymbol{\mu}^{*,s}} = [0.2\ 0.4\ 0.4],\ e^{\boldsymbol{\mu}^{*,g}} = [1.2\ 0.1\ 0.1]$, so as to fulfill the equilibrium condition (4).

We illustrate the properties of the scheme on a uniform mesh of $N = 100$ cells with time step $\Delta t_1 = 6 \times 10^{-4}$ and a final time $T_1 = 5$. Snapshots of the simulation are presented in Fig. 2, where one notices the formation of a discontinuity at the free interface and convergence to the two-phase stationary solution. We also study the long-time asymptotics: we first compute accurately the stationary solution $(\boldsymbol{c}^\infty, X^\infty)$ obtained in Proposition 1. Then we study the relative free energy $\mathcal{H}^p(\boldsymbol{c}_{\Delta x}^p, X^p) - \mathcal{H}^\infty(\boldsymbol{c}^\infty, X^\infty)$ and relative interface $X^\infty - X^p$ over time. The results are given in Fig. 3a, indicating exponential speed of convergence and decrease of both functionals. In particular, our scheme is well-balanced and preserves the asymptotics of the continuous system.

Our second test is devoted to a convergence analysis with respect to the size of the mesh. We consider a fixed time step $\Delta t_2 = 10^{-4}$, a final time $T_2 = 0.25$, uniform meshes from 2^3 to 2^{10} cells and we compare the different solutions with respect to a



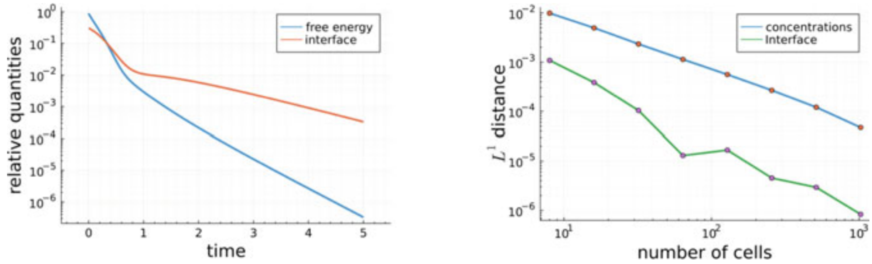
(a) Initial profiles

(b) $t = 0.25$

(c) $t = 1.0$

(d) Stationary profiles

Fig. 2 Concentration profiles at different times



(a) Long-time asymptotics

(b) Convergence analysis

Fig. 3 $(\mathcal{H}(\boldsymbol{c}(t), X(t)) - \mathcal{H}(\boldsymbol{c}^\infty, X^\infty))$ and $(X^\infty - X(t))$ as functions of time (left). Convergence analysis of the solution under space grid refinement (right)

reference solution computed on a finer grid of 2^{11} cells. The space-time (resp. time) L^1 error on the concentrations (resp. on the interface) are displayed in Fig. 3b. One clearly observes convergence, at first order in space for the concentrations. These results should be compared with the second order accurate one-phase schemes [4, 5]. On the one hand, it is plausible that the interface treatment induces the loss of order. On the other hand, the discrete $L^1((0, 1))$ space distance we use to compare solutions is not perfectly adapted since the solutions are defined in slightly different domains. Rescaling all quantities might offer more insights into the convergence properties.

Acknowledgements The authors acknowledge support from the ANR project COMODO (ANR-19-CE46-0002) which funds the Ph.D. of Jean Cauvin-Vila.

References

1. Bakhta, A., Ehrlacher, V.: Cross-diffusion systems with non-zero flux and moving boundary conditions. ESAIM:M2AN **52**(4), 1385–1415 (2018)
2. Bothe, D.: On the Maxwell-Stefan approach to multicomponent diffusion. Parabol. Probl. **80**, 81–93 (2011)
3. Cancès, C., Cauvin-Vila, J., Chainais-Hillairet, C., Ehrlacher, V.: A Convergent Finite Volume Scheme for a Free Interface Cross-Diffusion Model. In preparation
4. Cancès, C., Gaudeul, B.: A convergent entropy diminishing finite volume scheme for a cross-diffusion system. SINUM **58**(5), 2684–2710 (2020)
5. Cancès, C., Ehrlacher, V., Monasse L.: Finite Volumes for the Stefan-Maxwell Cross-Diffusion System (2020). arXiv:2007.09951
6. Glitzky, A., Mielke, A.: A gradient structure for systems coupling reaction-diffusion effects in bulk and interfaces. ZAMP **64**, 29–52 (2013)
7. Mielke, A., Peletier, M.A., Renger, D.M.: On the relation between gradient flows and the large-deviation principle, with applications to Markov chains and diffusion. Potential Anal. **41**, 1293–1327 (2014)

Finite Volume Scheme for the Diffusive Field-Road Model: Study of the Long Time Behaviour

Matthieu Alfaro and Claire Chainais-Hillairet

Abstract We consider the field-road system, a model for *fast diffusion channels* in population dynamics, consisting of two parabolic equations posed on sets of different dimensions and coupled through exchange terms on "the road". We propose a finite volume scheme for this model, the analysis of which requires an unconventional discrete Poincaré-Wirtinger inequality, and establish some numerical analysis results.

Keywords Field-road model · Finite volume method · Long time behaviour · Entropy dissipation

1 Presentation of the Field-Road Model

The *field-road* model was introduced in [2] to describe the spread of invasive species in presence of networks with fast propagation. It consists in a reaction-diffusion equation on the half plane $\mathbb{R}^{N-1} \times \{x_N > 0\}$ (the field, $N \geq 2$) coupled with a diffusion equation on the hyperplane $\mathbb{R}^{N-1} \times \{x_N = 0\}$ (the road). The coupling is ensured by the boundary terms on the field and the zeroth-order term on the road. This problem has been widely studied in the literature from a theoretical viewpoint. In the present work, we consider the *purely diffusive* field-road model, see [1], in a bounded domain and we study its numerical approximation by a TPFA finite volume scheme. We will briefly give some properties satisfied by the scheme and state its long time behaviour obtained by an entropy method. The originality of this work comes from the difference of dimension between the field and the road and the exchange terms between both. In particular a refinement of Poincaré-Wirtinger inequality will be required for the analysis.

M. Alfaro
University of Rouen Normandie, LMRS, CNRS, Rouen, France
e-mail: matthieu.alfaro@univ-rouen.fr

C. Chainais-Hillairet (✉)
Univ. Lille, CNRS, Inria, UMR 8524 - Laboratoire Paul Painlevé, 59000 Lille, France
e-mail: claire.chainais@univ-lille.fr

E. Franck et al. (eds.), *Finite Volumes for Complex Applications X—Volume 1, Elliptic and Parabolic Problems*, Springer Proceedings in Mathematics & Statistics 432,
https://doi.org/10.1007/978-3-031-40864-9_16

We restrict the presentation to the case $N = 2$. The road is defined by ω, an open bounded interval, while the field is defined by $\Omega = \omega \times (0, L)$ with $L > 0$. We denote by $\mathbf{n}$ the outward normal to Ω, and also to ω. The field-road model writes as follows:

$$\begin{cases} \partial_t v = d\Delta v, & \text{in } \Omega \times \{t > 0\}, \\ d\nabla v \cdot \mathbf{n} = \mu u - \nu v|_\omega, & \text{on } \omega \times \{t > 0\}, \\ d\nabla v \cdot \mathbf{n} = 0, & \text{on } \partial\Omega \setminus \omega \times \{t > 0\}, \\ \partial_t u = D\Delta u + \nu v|_\omega - \mu u, & \text{in } \omega \times \{t > 0\}, \\ D\nabla u \cdot \mathbf{n} = 0, & \text{on } \partial\omega \times \{t > 0\}, \end{cases} \tag{1}$$

where $v = v(x, y, t)$ (*resp.* $u = u(x, t)$) denotes the density of individuals in the field (*resp.* in the road), with $x \in \omega$, $0 < y < L$. Furthermore, d and D are positive diffusion coefficients, and μ and ν positive transfer coefficients between the field and the road. The system (1) is supplemented with (nonnegative) initial conditions $v_0 = v_0(x, y)$ on Ω and $u_0 = u_0(x)$ on ω.

In Sect. 2, we present the finite volume scheme. Its first properties, including *dissipation*, are presented in Sect. 3. Exponential decay in time to the associated steady-state is proved in Sect. 4. It requires to relate entropy and dissipation, via an unconventional discrete Poincaré-Wirtinger inequality adapted to the field-road model. Last, in Sect. 5, we present some simulations, sustaining our convergence result and exploring further effects.

2 The TPFA Finite Volume Scheme

2.1 Meshes and Notation

Let us first consider a mesh of Ω made of a family of control volumes, a family of edges and a family of points: $\mathcal{M}_\Omega = (\mathcal{T}_\Omega, \mathcal{E}_\Omega, \mathcal{P}_\Omega)$. We use classical notations: K for a control volume, σ for an edge, x_K for an interior point of K (named as the center of K). The mesh is admissible in the sense that it satisfies the usual orthogonality property, see [4]. We also consider an admissible mesh of ω, $\mathcal{M}_\omega = (\mathcal{T}_\omega, \mathcal{E}_\omega, \mathcal{P}_\omega)$. We denote by K^* a control volume of $\mathcal{T}_\omega$, σ^* an edge (a point in practice) of $\mathcal{E}_\omega$ and x_{K^*} an interior point of K^*.

In $\mathcal{T}_\Omega$, we can distinguish the control volumes that have an edge on the road from the other ones that are strictly included in the field, which writes $\mathcal{T}_\Omega = \mathcal{T}_\Omega^r \cup \mathcal{T}_\Omega^f$. For the edges of $\mathcal{E}_\Omega$ we can also distinguish the interior edges from the boundary edges, included in ω or included in $\partial\Omega \setminus \omega$ (considered as exterior edges). We have $\mathcal{E}_\Omega = \mathcal{E}_\Omega^{\text{int}} \cup \mathcal{E}_\Omega^r \cup \mathcal{E}_\Omega^{\text{ext}}$. For an interior edge $\sigma \in \mathcal{E}_\Omega^{\text{int}}$, we may write $\sigma = K|L$.

We assume the compatibility of the two meshes $\mathcal{M}_\Omega$ and $\mathcal{M}_\omega$: every control volume of $\mathcal{T}_\omega$ must coincide with an edge of $\mathcal{E}_\Omega^r$. More precisely, for all $\sigma \in \mathcal{E}_\Omega^r$, there exists a unique $K \in \mathcal{T}_\Omega^r$ such that σ is an edge of K and a unique $K^* \in \mathcal{T}_\omega$ such that σ coincide with K^*. Therefore, we will use the notation $\sigma = K|K^*$.

The measures of control volumes or edges are denoted by m_K, m_{K^*}, m_σ, m_{σ^*} (which is set equal to 1 in our case). We also define by d_σ or d_{σ^*} the distance associated to an edge $\sigma \in \mathcal{E}_\Omega$ or $\sigma^* \in \mathcal{E}_\omega$, usually defined as the distance between the centers of two neighbouring cells (or the distance from the center to the boundary), so that the transmissivities are defined by

$$\tau_\sigma = \frac{m_\sigma}{d_\sigma}\ \forall \sigma \in \mathcal{E}_\Omega, \quad \tau_{\sigma^*} = \frac{m_{\sigma^*}}{d_{\sigma^*}}\ \forall \sigma^* \in \mathcal{E}_\omega.$$

In view of time discretization, we consider a time step δt.

2.2 The Scheme

Let us denote by $((v_K^n)_{K\in\mathcal{T}_\Omega, n\geq 0}, (v_{K^*}^n)_{K^*\in\mathcal{T}_\omega, n\geq 1}, (u_{K^*}^n)_{K^*\in\mathcal{T}_\omega, n\geq 0})$ the discrete unknowns. We start with the discretization of the initial conditions: v_K^0 and $u_{K^*}^0$ are defined as the mean values of v_0 and u_0 respectively over $K \in \mathcal{T}_\Omega$ and $K^* \in \mathcal{T}_\omega$.

The scheme we propose is a backward Euler scheme in time and a two-point flux approximation finite volume scheme in space. It writes as follows:

$$\begin{cases} m_K \dfrac{v_K^n - v_K^{n-1}}{\delta t} + d \displaystyle\sum_{\sigma=K|L} \tau_\sigma (v_K^n - v_L^n) + d \sum_{\sigma=K|K^*} \tau_\sigma (v_K^n - v_{K^*}^n) = 0, \forall K \in \mathcal{T}_\Omega, \\ -d\tau_\sigma (v_K^n - v_{K^*}^n) = m_{K^*}(\mu u_{K^*}^n - \nu v_{K^*}^n),\ \forall \sigma \in \mathcal{E}_\Omega^r, \sigma = K|K^*, \\ m_{K^*} \dfrac{u_{K^*}^n - u_{K^*}^{n-1}}{\delta t} + D \displaystyle\sum_{\sigma^*=K^*|L^*} \tau_{\sigma^*}(u_{K^*}^n - u_{L^*}^n) \\ \qquad\qquad\qquad + m_{K^*}(\mu u_{K^*}^n - \nu v_{K^*}^n) = 0, \forall K^* \in \mathcal{T}_\omega. \end{cases} \tag{2}$$

At each time step, the scheme consists in a square linear system of equations of size $\#\mathcal{T}_\Omega + 2\#\mathcal{T}_\omega$. We can obtain a weak formulation of the scheme by multiplying the equations in (2) by some test values and summing over $\mathcal{T}_\Omega$, $\mathcal{E}_\Omega^r$, $\mathcal{T}_\omega$. For a given vector $((\varphi_K)_{K\in\mathcal{T}_\Omega}, (\varphi_{K^*})_{K^*\in\mathcal{T}_\omega}, (\psi_{K^*})_{K^*\in\mathcal{T}_\omega})$, we obtain

$$\begin{aligned} &\sum_{K\in\mathcal{T}_\Omega} m_K \varphi_K \frac{v_K^n - v_K^{n-1}}{\delta t} + \sum_{K^*\in\mathcal{T}_\omega} m_{K^*}\psi_{K^*} \frac{u_{K^*}^n - u_{K^*}^{n-1}}{\delta t} \\ &+ d \sum_{\sigma=K|K^*} \tau_\sigma (v_K^n - v_{K^*}^n)(\varphi_K - \varphi_{K^*}) = -D \sum_{\sigma^*=K^*|L^*} \tau_{\sigma^*}(u_{K^*}^n - u_{L^*}^n)(\psi_{K^*} - \psi_{L^*}) \\ &- d \sum_{\sigma=K|L} \tau_\sigma (v_K^n - v_L^n)(\varphi_K - \varphi_L) - \sum_{K^*\in\mathcal{T}_\omega} m_{K^*}(\mu u_{K^*}^n - \nu v_{K^*}^n)(\psi_{K^*} - \varphi_{K^*}). \end{aligned} \tag{3}$$

This weak formulation (3) is equivalent to the scheme (2).

3 Existence of a Solution and First Properties

Existence, Uniqueness and Positivity of the Solutions to the Scheme

Assuming that $v_K^{n-1} = 0$ for all $K \in \mathcal{T}_\Omega$ and $u_{K^*}^{n-1} = 0$ for all $K^* \in \mathcal{T}_\omega$ and choosing $\varphi_K = \nu v_K^n$, $\varphi_{K^*} = \nu v_{K^*}^n$, $\psi_{K^*} = \mu u_{K^*}^n$ in (3) yields existence and uniqueness of a solution to the scheme (2) at each time step.

Assuming now that $v_K^{n-1} \geq 0$ for all $K \in \mathcal{T}_\Omega$ and $u_{K^*}^{n-1} \geq 0$ for all $K^* \in \mathcal{T}_\omega$ choosing $\varphi_K = \nu (v_K^n)^-$, $\varphi_{K^*} = \nu (v_{K^*}^n)^-$, $\psi_{K^*} = \mu (u_{K^*}^n)^-$ (where x^- denotes the negative part of $x \in \mathbb{R}$) in (3) yields by induction the nonnegativity of the solution to the scheme (2) at each time step:

$$v_K^n \geq 0 \; \forall K \in \mathcal{T}_\Omega, \quad v_{K^*}^n \geq 0, u_{K^*}^n \geq 0 \; \forall K^* \in \mathcal{T}_\omega.$$

Mass conservation and steady-state

Choosing the test vector constant equal to 1 in (3) leads to the conservation of the total mass:

$$\sum_{K \in \mathcal{T}_\Omega} m_K v_K^n + \sum_{K^* \in \mathcal{T}_\omega} m_{K^*} u_{K^*}^n = \int_\Omega v^0 dx dy + \int_\omega u^0 dx, \quad \forall n \geq 0. \tag{4}$$

We will denote by M^0 the total mass.

A steady-state verifies $v_K^n = v_K^{n-1} = v_K^\infty$ for all $K \in \mathcal{T}_\Omega$ and $u_{K^*}^n = u_{K^*}^{n-1} = u_{K^*}^\infty$ for all $K^* \in \mathcal{T}_\omega$, with $v_{K^*}^n = v_{K^*}^\infty$ for all $K^* \in \mathcal{T}_\omega$. With $\varphi_K = \nu v_K^\infty$, $\varphi_{K^*} = \nu v_{K^*}^\infty$, $\psi_{K^*} = \mu u_{K^*}^\infty$ in (3), we obtain that the steady-state is constant in space, i.e. $v_K^\infty = v^\infty = v_{K^*}^\infty$ and $u_{K^*}^\infty = u^\infty$ and, from the mass conservation,

$$\begin{cases} \nu v^\infty - \mu u^\infty = 0, \\ m_\Omega v^\infty + m_\omega u^\infty = M^0, \end{cases} \quad \text{so that } \frac{v^\infty}{\mu} = \frac{M^0}{\mu m_\Omega + \nu m_\omega} = \frac{u^\infty}{\nu}. \tag{5}$$

From now on, we assume that M^0 is positive, so that v^∞ and u^∞ are also positive.

Relative entropies and dissipations

For any twice differentiable function Φ satisfying

$$\Phi'' > 0, \;\; \Phi(1) = 0, \;\; \Phi'(1) = 0,$$

we define, as in [3], a discrete entropy, relative to the steady state (v^∞, u^∞), by

$$\mathcal{H}_\Phi^n = \sum_{K\in\mathcal{T}_\Omega} m_K v^\infty \Phi(\frac{v_K^n}{v^\infty}) + \sum_{K^*\in\mathcal{T}_\omega} m_{K^*} u^\infty \Phi(\frac{u_{K^*}^n}{u^\infty}), \quad \forall n \geq 0. \tag{6}$$

Lemma 1 *Let* $((v_K^n)_{K\in\mathcal{T}_\Omega, n\geq 0}, (v_{K^*}^n)_{K^*\in\mathcal{T}_\omega, n\geq 1}, (u_{K^*}^n)_{K^*\in\mathcal{T}_\omega, n\geq 0})$ *be a solution to the scheme* (2) *and* (v^∞, u^∞) *the associated steady-state defined by* (5)*, then the discrete entropy defined by* (6) *is dissipated along time, as follows:*

$$\frac{\mathcal{H}_\Phi^n - \mathcal{H}_\Phi^{n-1}}{\delta t} \leq -\mathcal{D}_\Phi^n \ \leq 0 \ \ \forall n \geq 0, \tag{7}$$

$$\begin{aligned} \textit{with } \mathcal{D}_\Phi^n =& d \sum_{\sigma=K|K^*} \tau_\sigma (v_K^n - v_{K^*}^n) \left(\Phi'(\frac{v_K^n}{v^\infty}) - \Phi'(\frac{v_{K^*}^n}{v^\infty}) \right) \\ &+ d \sum_{\sigma=K|L} \tau_\sigma (v_K^n - v_L^n) \left(\Phi'(\frac{v_K^n}{v^\infty}) - \Phi'(\frac{v_L^n}{v^\infty}) \right) \\ &+ D \sum_{\sigma^*=K^*|L^*} \tau_{\sigma^*} (u_{K^*}^n - u_{L^*}^n) \left(\Phi'(\frac{u_{K^*}^n}{u^\infty}) - \Phi'(\frac{u_{L^*}^n}{u^\infty}) \right) \\ &+ \mu u^\infty \sum_{K^*\in\mathcal{T}_\omega} m_{K^*} \left(\frac{u_{K^*}^n}{u^\infty} - \frac{v_{K^*}^n}{v^\infty} \right) \left(\Phi'(\frac{u_{K^*}^n}{u^\infty}) - \Phi'(\frac{v_{K^*}^n}{v^\infty}) \right) \geq 0. \end{aligned} \tag{8}$$

Proof Due to the convexity of Φ, we have

$$\frac{\mathcal{H}_\Phi^n - \mathcal{H}_\Phi^{n-1}}{\delta t} \leq \sum_{K\in\mathcal{T}_\Omega} m_K \frac{v_K^n - v_K^{n-1}}{\delta t} \Phi'(\frac{v_K^n}{v^\infty}) + \sum_{K^*\in\mathcal{T}_\omega} m_{K^*} \frac{u_{K^*}^n - u_{K^*}^{n-1}}{\delta t} \Phi'(\frac{u_{K^*}^n}{u^\infty}).$$

Then, we apply (3) with $\varphi_K = \Phi'(\frac{v_K^n}{v^\infty})$, $\varphi_{K^*} = \Phi'(\frac{v_{K^*}^n}{v^\infty})$, $\psi_{K^*} = \Phi'(\frac{u_{K^*}^n}{u^\infty})$, which leads to the entropy-dissipation relation (7). The dissipation term $\mathcal{D}_\Phi^n$ rewrites as (8) thanks to (5). Moreover it is nonnegative due to the monotonicity of Φ'. □

4 Long-Time Behaviour: Convergence to the Steady-State

From now on, we will focus on the special case where $\Phi(x) = (x-1)^2/2$ (frequently denoted as Φ_2). Forgetting the superscript n, the corresponding entropy and dissipation are denoted by $\mathcal{H}_2$ and $\mathcal{D}_2$ and they are equal to:

$$\mathcal{H}_2 = \frac{1}{2}\sum_{K\in\mathcal{T}_\Omega} m_K \frac{(v_K - v^\infty)^2}{v^\infty} + \frac{1}{2}\sum_{K^*\in\mathcal{T}_\omega} m_{K^*}\frac{(u_{K^*}-u^\infty)^2}{u^\infty},$$

$$\mathcal{D}_2 = d\sum_{\sigma=K|K^*}\tau_\sigma\frac{(v_K - v_{K^*})^2}{v^\infty} + d\sum_{\sigma=K|L}\tau_\sigma\frac{(v_K - v_L)^2}{v^\infty}$$
$$+D\sum_{\sigma^*=K^*|L^*}\tau_{\sigma^*}\frac{(u_{K^*}-u_{L^*})^2}{u^\infty} + \mu u^\infty\sum_{K^*\in\mathcal{T}_\omega} m_{K^*}\left(\frac{u_{K^*}}{u^\infty} - \frac{v_{K^*}}{v^\infty}\right)^2.$$

We notice that the relative entropy $\mathcal{H}_2$ corresponds to a weighted L^2 distance between the solution to the scheme and its steady-state, that have the same total mass, while the dissipation $\mathcal{D}_2$ corresponds to a weighted L^2 norm of a discrete gradient of the solution on the field and the road, with additional exchange terms on the road. Proposition 1 states a crucial relation between entropy and dissipation which can be seen as a kind of discrete Poincaré-Wirtinger inequality adapted to the field-road model.

Proposition 1 *Let us consider a set* $((v_K)_{K\in\mathcal{T}_\Omega}, (v_{K^*})_{K^*\in\mathcal{T}_\omega}, (u_{K^*})_{K^*\in\mathcal{T}_\omega})$ *of discrete values with a total mass*

$$M^0 = \sum_{K\in\mathcal{T}_\Omega} m_K v_K + \sum_{K^*\in\mathcal{T}_\omega} m_{K^*}u_{K^*},$$

and (v^∞, u^∞) *defined by* (5). *There exists a constant* Λ *depending on the domain* Ω *and on the data* M^0, μ, ν, d, D *such that*

$$\mathcal{H}_2 \le \frac{1}{\Lambda}\mathcal{D}_2. \tag{9}$$

Proof We only give a sketch of the proof. Let $\ell > 0$, we denote by $\Omega_\ell = \omega\times(-\ell, 0)$ the "thickened" road and $\Omega^+ = \omega\times(-\ell, L)$ the enlarged domain. Thanks to (5), we can define a probability measure γ on Ω^+ by

$$d\gamma(x,y) = \left(\frac{v^\infty}{M^0}\mathbf{1}_\Omega(x,y) + \frac{1}{\ell}\frac{u^\infty}{M^0}\mathbf{1}_{\Omega_\ell}(x,y)\right)dxdy.$$

We also introduce two piecewise constant functions f and $\bar{v}$ on Ω^+ defined by

$$f(x,y) = \sum_{K\in\mathcal{T}_\Omega}\frac{v_K}{v^\infty}\mathbf{1}_K(x,y) + \sum_{K^*\in\mathcal{T}_\omega}\frac{u_{K^*}}{u^\infty}\mathbf{1}_{K^*\times(-\ell,0)}(x,y),$$
$$\bar{v}(x,y) = \sum_{K^*\in\mathcal{T}_\omega}\frac{v_{K^*}}{v^\infty}\mathbf{1}_{K^*\times(-\ell,0)}(x,y).$$

We notice that $\gamma(f) = 1$, so that $\mathcal{H}_2$ can be rewritten as

$$\begin{aligned}\mathcal{H}_2 &= \frac{M^0}{2}\int_{\Omega^+} (f(x,y) - \gamma(f))^2 d\gamma(x,y)\\ &= \frac{M^0}{4}\int_{\Omega^+}\int_{\Omega^+} (f(x,y) - f(x',y'))^2 d\gamma(x,y)d\gamma(x',y').\end{aligned}$$

The integral over $\Omega^+ \times \Omega^+$ can be split into three terms corresponding to the integrals over $\Omega \times \Omega$, $\Omega_\ell \times \Omega_\ell$ and $\Omega \times \Omega_\ell$ (counted twice). We may apply the proof of the discrete mean Poincaré inequality in [4] to the first two terms, so that they are bounded (up to a multiplicative constant) respectively by

$$\sum_{\sigma=K|L} \tau_\sigma \frac{(v_K - v_L)^2}{v^\infty} \text{ and } \sum_{\sigma^*=K^*|L^*} \tau_{\sigma^*} \frac{(u_{K^*} - u_{L^*})^2}{u^\infty}.$$

In the cross term, we may introduce $\bar{v}$ in this way:

$$(f(x,y) - f(x',y'))^2 \le 2(f(x,y) - \bar{v}(x',y'))^2 + 2(f(x',y') - \bar{v}(x',y'))^2.$$

The integral over $\Omega \times \Omega_\ell$ of the first term in the above right hand side is bounded, up to a multiplicative constant, by

$$\sum_{\sigma=K|K^*} \tau_\sigma \frac{(v_K - v_{K^*})^2}{v^\infty},$$

thanks to a refinement of [4] to deal with the non symmetry of the domain. Last $\int_\Omega\int_{\Omega_\ell} (f(x',y') - \bar{v}(x',y'))^2 d\gamma(x,y)d\gamma(x',y')$ is nothing else than

$$\frac{u^\infty}{M^0}\sum_{K^*\in\mathcal{T}_\omega} m_{K^*}\left(\frac{u_{K^*}}{u^\infty} - \frac{v_{K^*}}{v^\infty}\right)^2$$

which is a term appearing in the expression of $\mathcal{D}_2$. □

Lemma 1 and Proposition 1, combined with a discrete Gronwall lemma, lead to the exponential decay of the relative entropy in time, and therefore of the distance in L^2-norm between the solution at time step n and the steady-state.

Theorem 1 *Let $((v_K^n)_{K\in\mathcal{T}_\Omega,n\ge0}, (v_{K^*}^n)_{K^*\in\mathcal{T}_\omega,n\ge1}, (u_{K^*}^n)_{K^*\in\mathcal{T}_\omega,n\ge0})$ be a solution to the scheme (2) and (v^∞, u^∞) the associated steady-state defined by (5), then we have, for all $n \ge 0$,*

$$\mathcal{H}_2^n \le (1 + \Lambda\delta t)^{-n}\left(\frac{1}{2v^\infty}\int_\Omega (v^0 - v^\infty)^2 dxdy + \frac{1}{2u^\infty}\int_\omega (u^0 - u^\infty)^2 dx\right).$$

5 Numerical Experiments

We consider two test cases inspired from [1]. Let $L = 20$, the road and the field are defined by $\omega = (-2L, 2L)$ and $\Omega = \omega \times (0, L)$. For the first test case, there is no individual on the road at $t = 0$; the initial condition is given by

$$v_0 = \mathbf{1}_{[-2.5,2.5]\times[0,5]} \text{ and } u_0 = 0 \Longrightarrow M^0 = 25.$$

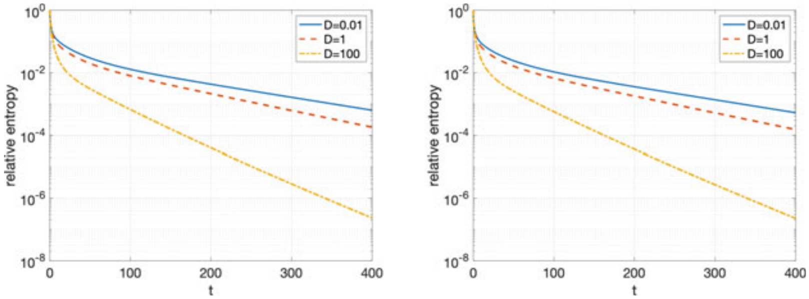
For the second test case, we assume that there are individuals on the road at $t = 0$, but v_0 and u_0 ensure that the total mass M_0 remains the same:

$$v_0 = 1.5 \cdot \mathbf{1}_{[-2.5,2.5]\times[2.5,5]} \text{ and } u_0 = 1.25 \cdot \mathbf{1}_{[-2.5,2.5]} \Longrightarrow M^0 = 25.$$

Moreover, for both test cases, we choose $\mu = 1$ and $\nu = 5$, so that they have the same steady state: $(v^\infty, u^\infty) = (1/80, 5/80)$.We also fix the diffusion coefficient for the field $d = 1$ and we choose different values for the diffusion coefficient on the road D. The mesh we use for the simulations is made of 14336 triangles.

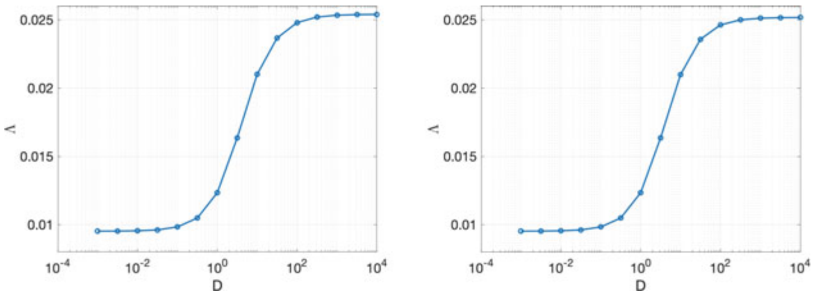
We plot on Fig. 1 the evolution of the relative entropy with respect to time for $D = 0.01, 1, 100$. As expected, we observe an exponential decay towards 0 in time, with a decay rate depending on D. We also notice that the results are almost the same for both test cases.

Figure 2 shows the evolution of the decay rate Λ (computed experimentally) as a function of D. We observe that Λ behaves as a monotone function of D for both test cases. We notice some speed-up of the decay rate when D si sufficiently large. Moreover, we observe that the two curves are almost the same. This suggests that the decay rate essentially depends on D and the steady state (v^∞, u^∞). The latter keeps a trace of the initial total mass but not of the initial "fragmentation" (Test Case 1 vs. Test Case 2). The dependence on the transfer coefficients, μ and ν, would deserve further investigations.



Test case 1 Test case 2

Fig. 1 Discrete relative entropy $\mathcal{H}_2^n/\mathcal{H}_2^0$ as a function of time for different values of D.



Test case 1

Test case 2

Fig. 2 Decay rate Λ as a function of the diffusion coefficient D.

Acknowledgements M. A. is supported by the région Normandie project BIOMANORMAN 21E04343 and the ANR project DEEV ANR-20-CE40-0011-01. C. C.-H. acknowledges support from the Labex CEMPI (ANR-11-LABX-0007-01).

References

1. Alfaro, M., Ducasse, R., Tréton, S.: The field-road diffusion model: fundamental solution and asymptotic behavior. J. Differ. Equat. **367**, 332–365 (2023). https://doi.org/10.1016/j.jde.2023.05.002
2. Berestycki, H., Roquejoffre, J.-M., Rossi, L.: The influence of a line with fast diffusion on Fisher-KPP propagation. J. Math. Biol. **66**, 743–766 (2013)
3. 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. **40**, 2473–2504 (2020)
4. Eymard, R., Gallouët, T., Herbin, R.: Finite volume methods. Handb. Numer. Anal. **7**, 713–1020 (2000)

An Approximate Two-Point Dirichlet Flux for Quasilinear Convection Diffusion Equations

C. Chainais-Hillairet, R. Eymard, and J. Fuhrmann

Abstract We define, in the case of quasilinear convection-diffusion equations, an approximation of the numerical fluxes obtained by extending the Scharfetter and Gummel fluxes (defined in the case of linear convection—diffusion). We show that this approximation is compatible with the asymptotic thermal equilibrium on an application example.

Keywords Quasilinear convection–diffusion equation · Scharfetter–Gummel flux · Long time behavior · Log-sobolev inequalities

1 Numerical Fluxes for Quasilinear Convection-Diffusion

Quasilinear convection-diffusion equations occur in a number of interesting applications in semiconductor physics, biology and other fields. Thermodynamical consistency of numerical fluxes is a requirement for two point flux finite volume discretizations of these problems. For the linear case, Scharfetter and Gummel [1] defined such fluxes based on the analytical solution of a two point Dirichlet boundary value problem defined at the interfaces between neighboring control volumes. In [2], this approach was generalized to the quasilinear case, resulting in a nonlinear integral equation for the numerical flux. Recently, in [3] a new approximation scheme for this problem was found. In this contribution, we review the results from [3] and apply

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

R. Eymard
LAMA - UMR 8050, Université Gustave Eiffel, Champs-sur-Marne, France
e-mail: robert.eymard@univ-eiffel.fr

J. Fuhrmann (✉)
Weierstrass Institute for Applied Analysis and Stochastics, Berlin, Germany
e-mail: juergen.fuhrmann@wias-berlin.de

E. Franck et al. (eds.), *Finite Volumes for Complex Applications X—Volume 1, Elliptic and Parabolic Problems*, Springer Proceedings in Mathematics & Statistics 432,
https://doi.org/10.1007/978-3-031-40864-9_17

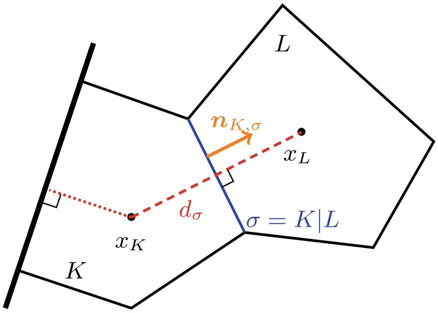


Fig. 1 Two control volumes in a finite volume scheme

the method to a two-dimensional demonstration example. We consider the nonlinear conservation law

$$u_t + \operatorname{div} \boldsymbol{J} = 0 \tag{1}$$

with nonlinear fluxes depending on the unknown function u and its gradient in a bounded domain $\Omega \subset \mathbb{R}^d$ ($d \geq 1$), supplemented by initial and boundary conditions (letting $\partial\Omega = \Gamma^D \cup \Gamma^N$, we consider Dirichlet boundary conditions on Γ^D and no-flux boundary conditions on Γ^N). Let ζ and η be nonlinear functions depending on the unknown u, such that

(i) $\eta \in \mathcal{C}(\mathbb{R}, \mathbb{R})$ is Lipschitz continuous with Lipschitz constant L_η.
(ii) $\zeta \in \mathcal{C}^1(\mathbb{R}, \mathbb{R})$ is Lipschitz continuous with Lipschitz constant L_ζ and $\exists r > 0$ such that $\zeta'(s) \geq r \quad \forall s \in \mathbb{R}$.
(iii) $\boldsymbol{q} \in \mathcal{C}^1(\bar{\Omega}, \mathbb{R}^d)$.

Define the nonlinear flux

$$\boldsymbol{J} = -\nabla\zeta(u) + \eta(u)\boldsymbol{q}. \tag{2}$$

Consider an implicit Euler, two-point flux finite volume scheme (Fig. 1)

$$m_K \frac{u_K^{n+1} - u_K^n}{\Delta t} + \sum_{\sigma \in \mathcal{E}_{K,\mathrm{int}} \cup \mathcal{E}_K^D} m_\sigma \mathcal{F}_{K\sigma}^{n+1} = 0. \tag{3}$$

The numerical flux $\mathcal{F}_{K\sigma}^{n+1}$ through the edge σ of a control volume K is sought as a consistent and conservative approximation of $\frac{1}{m_\sigma}\int_\sigma \boldsymbol{J} \cdot n_{K,\sigma}$. It is defined as a function of u_K^{n+1} and u_L^{n+1} if σ is the common edge of two neighboring control volumes K and L (we will denote $\sigma = K|L$, and $\sigma \in \mathcal{E}_{K,\mathrm{int}}$). It is a function of u_K^{n+1} and u_σ^D if σ is an edge of K included in the Dirichlet boundary Γ^D (we will denote $\sigma \in \mathcal{E}_K^D$ in this case and define by u_σ^D an approximation of the Dirichlet data u^D on σ). Let

$$q_{K,\sigma} = \frac{1}{m_\sigma}\int_\sigma \boldsymbol{q} \cdot \boldsymbol{n}_{K\sigma} \mathrm{d}s \quad \forall \sigma \in \mathcal{E}_K \tag{4}$$

Then we set $\mathcal{F}^{n+1}_{K\sigma} = \mathcal{F}(u^{n+1}_K, u^{n+1}_{K\sigma}, q_{K,\sigma}, d_\sigma)$, where $\mathcal{F}$ verifies the following properties, which are shown to be sufficient for the convergence of the numerical scheme in the case div$\boldsymbol{q} = 0$ [2].

Definition 1 The function $\mathcal{F} : (a, b, q, h) \in \mathbb{R}^3 \times \mathbb{R}_+ \mapsto \mathcal{F}(a, b, q, h) \in \mathbb{R}$ defines an admissible numerical flux if:

(i) $\mathcal{F}$ is Lipschitz-continuous with respect to a and b.
(ii) $\mathcal{F}$ is increasing with respect to a, decreasing with respect to b.
(iii) $\mathcal{F}(a, b, q, h) + \mathcal{F}(b, a, -q, h) = 0$ for all $(a, b, q, h) \in \mathbb{R}^3 \times \mathbb{R}_+$.
(iv) There exists $c \in [a\bot b, a\top b]$ such that $\mathcal{F}(a, b, q, h) = q\eta(c) - \dfrac{\zeta(b) - \zeta(a)}{h}$.
(v) $(a - b)\mathcal{F}(a, b, q, h) \geq -\displaystyle\int_a^b q\eta(s)\mathrm{d}s + \frac{(\xi(b) - \xi(a))^2}{h}$ with $\xi(s) = \displaystyle\int_0^s \sqrt{\zeta'(t)}\mathrm{d}t$.

For linear functions η, ζ, Scharfetter and Gummel [1] in the context of semiconductor device simulation define these numerical fluxes by $qy - y'$ if $y : [0, h] \to \mathbb{R}$ is such that $(qy - y')' = 0$, $y(0) = a$, $y(y) = b$. Eymard, Fuhrmann and Gärtner in [2] generalized this idea to the nonlinear case: Search for the solution $y : [0, h] \to \mathbb{R}$ to

$$\Big(q\ \eta(y) - (\zeta(y))'\Big)' = 0,\ \ y(0) = a, y(h) = b, \tag{5}$$

and define $\mathcal{F}(a, b, q, h)$ as the constant value of $q\eta(y(s)) - (\zeta(y))'(s)$. For $a < b$, the numerical flux $\mathcal{F}(a, b, q, h)$ can be obtained as the unique solution of

$$H(\mathcal{F}(a, b, q, h)) = h \text{ where } H(x) = \int_a^b \frac{\zeta'(s)}{q\eta(s) - x}\mathrm{d}s. \tag{6}$$

For $a > b$, it is given by $\mathcal{F}(a, b, q, h) = \mathcal{F}(b, a, -q, h)$, and for $a = b$, it is given by $\mathcal{F}(a, b, q, h) = q\eta(a)$. It is then proved to be admissible in the sense of Definition 1.

In general one cannot analytically solve the integral equation (6). A few special cases beyond the linear one resulting in the classical Scharfetter-Gummel scheme allow for a simplified treatment [4, 5]. In [6], the integral in (6) has been replaced by various fixed quadrature rules. We propose in [3] an adaptive quadrature approach, based on the definition of $\mathcal{F}_\delta(a, b, q, h)$ for an approximation parameter $\delta > 0$.

First, set $\mathcal{F}_\delta(a, a, q, h) = q\eta(a)$ and $\mathcal{F}_\delta(a, b, q, h) = \mathcal{F}_\delta(b, a, -q, h)$ for $a > b$ as for the SG-nl fluxes so that it is sufficient to consider now that $a < b$.

Let $(\overline{y}_i)_{i\in\mathbb{Z}} \in \mathbb{R}^{\mathbb{Z}}$ be a given sequence, independent of a, b, q, h, such that:

1. the sequence $(\overline{y}_i)_{i\in\mathbb{Z}}$ is increasing,
2. $\overline{y}_i$ tends to $\pm\infty$ as $i \to \pm\infty$,
3. $\sup_{i\in\mathbb{Z}}(\overline{y}_{i+1} - \overline{y}_i) = \delta \in (0, +\infty)$.

Given $a < b$, we denote $i_a \in \mathbb{Z}$ the index such that $a \in [\overline{y}_{i_a}, \overline{y}_{i_a+1})$ and $i_b \in \mathbb{Z}$ the index such that $b \in (\overline{y}_{i_b}, \overline{y}_{i_b+1}]$. If $i_a < i_b$, we define $N = 1 + i_b - i_a$ and we

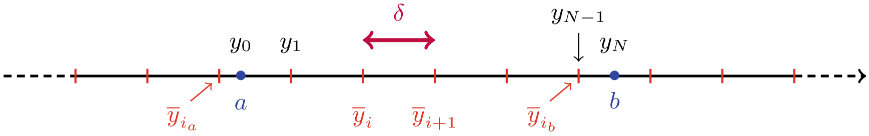


Fig. 2 Presentation of the approximation points $(\overline{y}_i)_{i\in\mathbb{Z}}$ and $(y_i)_{0\le i\le N}$

set $y_0 = a$, $y_1 = \overline{y}_{i_a+1}, \dots, y_{N-1} = \overline{y}_{i_b}$, $y_N = b$. This case is illustrated in Fig. 2. If $i_a = i_b$, we let $N = 1$ and we define $y_0 = a$, $y_1 = b$.

Denoting by $\mathcal{F}^{(q)}_{\text{god}}$ the Godunov flux function, define $H_\delta(x)$, for any $x \in (-\infty, \mathcal{F}^{(q)}_{\text{god}}(a, b))$, by

$$H_\delta(x) = \sum_{i=0}^{N-1} \frac{\zeta(y_{i+1}) - \zeta(y_i)}{\mathcal{F}^{(q)}_{\text{god}}(y_i, y_{i+1}) - x} \tag{7}$$

where

$$\mathcal{F}^{(q)}_{\text{god}}(u, v) = \begin{cases} \min\limits_{s\in[u,v]} \; q\eta(s) \text{ if } u \le v, \\ \max\limits_{s\in[v,u]} \; q\eta(s) \text{ if } v \le u. \end{cases} \tag{8}$$

Remark that $\mathcal{F}^{(q)}_{\text{god}}(a, b) \le \mathcal{F}^{(q)}_{\text{god}}(y_i, y_{i+1})$ for $i = 0, \dots, N-1$. Then $\mathcal{F}_\delta(a, b, q, h)$ is defined as the solution to the nonlinear equation:

$$h = H_\delta(\mathcal{F}_\delta(a, b, q, h)). \tag{9}$$

One should be aware that $\mathcal{F}_\delta$ depends on the full sequence $(\overline{y}_i)_{i\in\mathbb{Z}}$, not only on δ.

Lemma 1 (δ-fluxes are well defined and admissible) *[3] Let $a < b$. Then the function H_δ defined by* (7) *is increasing and strictly convex, and there holds*

$$\lim_{x\to-\infty} H_\delta(x) = 0 \textit{ and } \lim_{x \underset{<}{\to} \mathcal{F}^{(q)}_{\text{god}}(a,b)} H_\delta(x) = +\infty. \tag{10}$$

As a consequence, $\mathcal{F}_\delta(a, b, q, h)$ is well defined as the unique solution to the equation $H_\delta(x) = h$. Moreover, the δ-fluxes defined by (7)–(9) *are admissible in the sense of Definition 1.*

Lemma 2 (For $\delta \to 0$, the δ-fluxes approximate the SG-nl fluxes) *[3] There exists $C_F > 0$, only depending on Q, M, L_η, L_ζ and r where $M \ge h$ and $Q > |q|$, such that, for $a \le b$,*

$$0 \le \mathcal{F}(a, b, q, h) - \mathcal{F}_\delta(a, b, q, h) \le C_F\delta, \tag{11}$$

Due to (11), we can consider that $\mathcal{F}_0$ coincides with $\mathcal{F}$, the SG-nl flux. Therefore, in the sequel, we can use $\mathcal{F}_\delta$ with $\delta = 0$ to denote the SG-nl flux.

2 The Drift-Diffusion Case and Thermal Equilibrium

For $\Gamma^D = \emptyset$, also consider the special case $\boldsymbol{q} = -\nabla V$ with $V \in L^\infty(\Omega) \cap H^1(\Omega)$ such that $\int_\Omega V = 0$ and $\boldsymbol{q} \cdot \boldsymbol{n} = 0$ on $\Gamma = \partial\Omega$ in a weak sense. Under the additional condition $\eta(0) = 0$ and $\eta(s) > 0$ for all $s > 0$, with $\mu(t) = \int_1^t \frac{\zeta'(s)}{\eta(s)}\, ds$ one can express

$$\boldsymbol{J} = -\eta(u)\nabla\left(\mu(u) + V(\boldsymbol{x})\right). \tag{12}$$

Let $g := -\Delta V \in L^2(\Omega)$ and set $g_K = \dfrac{1}{m_K}\displaystyle\int_K g(x)dx$. Define $(V_K)_{K\in\mathcal{T}}$ by

$$-\sum_{\sigma=K|L} \tau_\sigma(V_L - V_K) = m_K g_K, \quad \forall K \in \mathcal{T}, \quad \text{and} \sum_{K\in\mathcal{T}} m_K V_K = 0. \tag{13}$$

Now, instead of Eq. (4), express $q_{K,L}$ in the numerical scheme (3) by

$$q_{K,L} = -\frac{V_L - V_K}{\mathrm{d}_\sigma} \quad \forall \sigma = K|L \in \mathcal{E}_{\mathrm{int}}. \tag{14}$$

In [3] it is shown that the resulting δ-scheme is mass conserving, has a unique solution, and yields nonnegative solutions for nonnegative initial values. Moreover, the convergence proof from [2] can be adapted to this case.

Lemma 3 (Consistency of the SG-nl flux with the special form of the continuous flux) *For all $a, b \in (0, +\infty)$, there exists $c \in [a \bot b, a \top b]$, denoted in the sequel $c = \chi(a, b)$, such that*

$$\mathcal{F}(a, b, q, h) = -\eta(c)\left(\frac{\mu(b) - \mu(a)}{h} - q\right). \tag{15}$$

This definition of $\chi(a, b)$ allows to characterize the entropy behavior with respect to the thermal equilibrium defined by the SG-nl scheme.

Lemma 4 (Thermal equilibrium) *For any $M^0 > 0$, there exists one and only one $(u_K)_{K\in\mathcal{T}}$ with $u_K \geq 0$ such that*

$$\sum_{K\in\mathcal{T}} m_K u_K = M^0 > 0, \tag{16}$$

and

$$\forall K \in \mathcal{T},\ \forall \sigma \in \mathcal{E}_{K,\mathrm{int}}, \sigma = K|L,\ \mathcal{F}(u_K, u_L, q_{K,L}, d_\sigma) = 0. \tag{17}$$

Moreover, we have $u_K > 0$ for all $K \in \mathcal{T}$, and there exists one and only one $\lambda \in \mathbb{R}$ such that

$$\forall K \in \mathcal{T},\ \mu(u_K) + V_K = \lambda.$$

Consequently, the SG-nl scheme preserves the thermal equilibrium of the continuous problem characterized by the existence of $\lambda \in \mathbb{R}$ such that $\mu(u) + V = \lambda$.

Lemma 5 (Discrete solutions uniformly bounded for δ small enough) *Let $(u_K^0)_{K\in\mathcal{T}}$ be given, with $u_K^0 > 0$ for all $K \in \mathcal{T}$. Then there exist $\underline{B} > 0$ and $\overline{B} > 0$, only depending on $\max_K u_K^0$, $\min_K u_K^0$, $\max_K V_K$, $\min_K V_K$ and μ, and $\delta_0 > 0$, only depending on $\max_K u_K^0$, $\min_K u_K^0$, $\max_K V_K$, $\min_K V_K$, μ and on $\mathcal{M}$, such that, for all $\delta < \delta_0$, any solution $(u_K^n)_{K\in\mathcal{T},n\geq 0}$ of the δ- scheme is such that*

$$\forall n \in \mathbb{N},\ \forall K \in \mathcal{T},\ 0 < \underline{B} \leq u_K^n \leq \overline{B}.$$

Lemma 6 (Decay of the relative entropy) *Let $(u_K^0)_{K\in\mathcal{T}}$ with $u_K^0 > 0$ be given. Let $(u_K^\infty)_{K\in\mathcal{T}}$ be the thermal equilibrium given by Lemma 4 for $M^0 = \sum_{K\in\mathcal{T}} m_K u_K^0$, and let $(u_K^n)_{K\in\mathcal{T},n\geq 0}$ be the solution to the δ-scheme ($\delta \geq 0$). For any $n \in \mathbb{N}$, we define the discrete entropy E^n and the associated discrete dissipation D^{n+1} by:*

$$E^n = \sum_{K\in\mathcal{T}} m_K \Big(\Phi(u_K^n) - \Phi(u_K^\infty) - \mu(u_K^\infty)(u_K^n - u_K^\infty) \Big) \tag{18}$$

$$D^{n+1} = \sum_{\sigma\in\mathcal{E}_{\text{int}},\sigma=K|L} \tau_\sigma \eta(\overline{u}_\sigma^{n+1}) \Big(\mu(u_K^{n+1}) - \mu(u_L^{n+1}) - \mu(u_K^\infty) + \mu(u_L^\infty) \Big)^2, \tag{19}$$

with $\overline{u}_\sigma^{n+1} = \chi(u_K^{n+1}, u_L^{n+1})$ defined by (15). Then there exists $\beta \geq 0$, only depending on Ω, $\|g\|_{L^2(\Omega)}$, L_η, L_ζ, such that

$$\frac{E^{n+1} - E^n}{\Delta t} + D^{n+1} \leq \beta\delta. \tag{20}$$

Theorem 1 (Convergence to the discrete thermal equilibrium) *Let (u_K^0) be given, with $u_K^0 > 0$ for all $K \in \mathcal{T}$ and M^0 the associate mass. Let $(u_K^n)_{K\in\mathcal{T},n\geq 0}$ the transient discrete solution and $(u_K^\infty)_{K\in\mathcal{T}}$ the thermal equilibrium defined by Lemma 4. Let $\delta_0 > 0$, $\underline{B} > 0$ and $\overline{B} > 0$ be given by Lemma 5, and $\beta > 0$ be given by Lemma 6. Then there exists $\alpha > 0$ only depending on μ, η, $\underline{B}$ and $\overline{B}$, such that, for any $\delta \in [0, \delta_0)$ and for any $n \in \mathbb{N}$, it holds*

$$\frac{1}{2} \min_{[\underline{B},\overline{B}]} \mu' \sum_{K\in\mathcal{T}} m_K (u_K^n - u_K^\infty)^2 \leq E^n \leq \beta\delta \frac{(1 - (1 + \alpha\Delta t)^{-n})}{\alpha} + E^0 (1 + \alpha\Delta t)^{-n}.$$

Note that, for $\Delta t \leq 1/\alpha$, it holds $(1 + \alpha\Delta t)^{-n} \leq \exp(-\frac{1}{2}\alpha n \Delta t)$, which shows in this case the exponential decay of E^n, up to δ.

3 A Numerical Experiment

Amending the numerical results provided in [3], in the present contribution we provide results of a 2D simulation inspired by [7]. Let $\Omega = (-10, 10) \times (-10, 10)$. For $V(\mathbf{x}) = -\frac{|\mathbf{x}|^2}{2}$, regard the convective porous medium equation

$$-\nabla(u^m + \alpha u) + u\nabla V = 0 \tag{21}$$

with homogeneous Neumann boundary conditions. For $\mathbf{x}_a = (2, -2)$ and $\mathbf{x}_b = (-2, 2)$, define the initial value

$$u_0(\mathbf{x}) = \begin{cases} \exp\left(-\frac{1}{6-|\mathbf{x}-\mathbf{x}_a|^2}\right) & \text{if } |\mathbf{x}-\mathbf{x}_a| < \sqrt{6} \\ \exp\left(-\frac{1}{6-|\mathbf{x}-\mathbf{x}_b|^2}\right) & \text{if } |\mathbf{x}-\mathbf{x}_b| < \sqrt{6} \\ 0 & \text{else} \end{cases} \tag{22}$$

For $\alpha = 0$, this problem has an equilibrium solution

$$u_{eq}(\mathbf{x}) = \left(C - \frac{m-1}{m} V(\mathbf{x})\right)_+^{\frac{1}{m-1}} \tag{23}$$

where the constant C is given from mass conservation $\int_\Omega u_{eq} = \int_\Omega u_0$. We discretize Ω by a grid of 60×60 discretization points and a timestep of 0.1. For $\alpha > 0$, the equilibrium solution can be obtained numerically from

$$\alpha \log u_{eq}(\mathbf{x}) + \frac{m}{m-1} u_{eq}(\mathbf{x})^{m-1} - (C - V(\mathbf{x})) = 0 \tag{24}$$

Figure 3 shows the solution at three moments of the evolution. Figure 4 demonstrates the evolution of the discrete relative entropy, its finite difference time derivative and the discrete dissipation rate for the degenerate case (violating the assumption

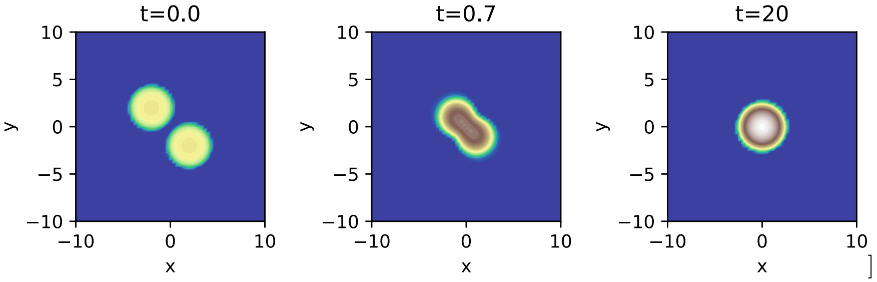


Fig. 3 Evolution from u_0 (22) to u_{eq} (23) under (21) for $\alpha = 0$

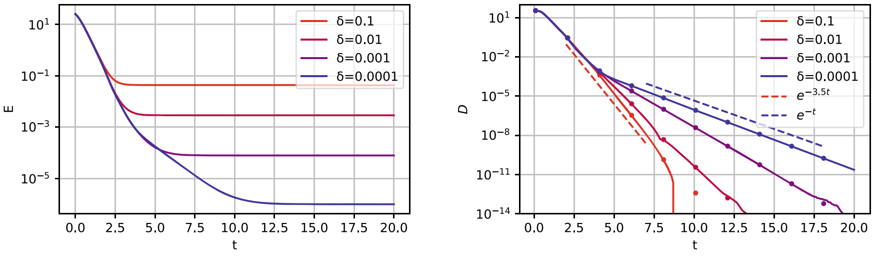


Fig. 4 Relative entropy E (18) (left), dissipation rate D (19) (right, lines) and $\partial_t E$ (right, dots) for different values of δ during evolution from u_0 (22) to u_{eq} (23) under (21) for $\alpha = 0$

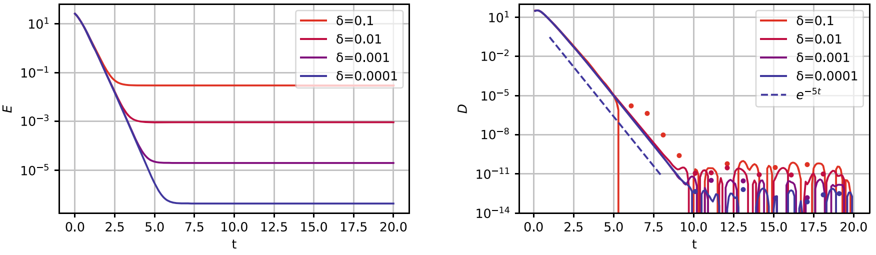


Fig. 5 Relative entropy E (18) (left), dissipation rate D (19) (right, lines) and $\partial_t E$ (right, dots) for different values of δ during evolution from u_0 (22) to u_{eq} (23) under (21) for $\alpha = 0.1$

Table 1 Estimated values of β from Lemma 6

δ	$\alpha = 0$	$\alpha = 0.1$
10^{-1}	$1.06 \cdot 10^{-12}$	$6.60 \cdot 10^{-13}$
10^{-2}	$7.59 \cdot 10^{-12}$	$4.51 \cdot 10^{-12}$
10^{-3}	$7.53 \cdot 10^{-11}$	$9.90 \cdot 10^{-11}$
10^{-4}	0	$5.60 \cdot 10^{-10}$

$\zeta' \geq r > 0$). We note that D becomes slightly negative for $\delta = 0.1$. The same data for $\alpha = 0.1$ is shown in Fig. 5. The solution for $\alpha = 0.1$ is visually not distinguishable from the one shown in Fig. 3. In both cases, a decrease of δ leads to a closer approach to the equilibrium for large times. As in the 1D case, the interesting observation is that for the nondegenerate case, the slope of the numerical dissipation rate does not depend on α.

From the computation, we also can estimate the value of β from Lemma 6 the resulting values shown in Table 1 essentially are not distinguishable from roundoff error, supporting the hypothesis that $\beta = 0$ could be proven.

References

1. Scharfetter, D.L., Gummel, H.K.: Large-signal analysis of a silicon Read diode oscillator. IEEE Trans. Electron Devices **16**(1), 64–77 (1969)
2. Eymard, R., Fuhrmann, J., Gärtner, K.: A finite volume scheme for nonlinear parabolic equations derived from one-dimensional local Dirichlet problems. Numer. Math. **102**(3), 463–495 (2006)
3. Chainais-Hillairet, C., Eymard, R., Fuhrmann, J.: A monotone numerical flux for quasilinear convection diffusion equation (2022). https://hal.science/hal-03791166
4. Fuhrmann, J.: Evaluation of numerical fluxes for a locally exact finite volume scheme using hypergeometric functions. In: Benkhaldoun, F., Ouazar, D., Raghay, S. (eds.) Finite Volumes in Complex Applications IV: Proceedings Marrakech, pp. 337–344. HERMES, Paris (2005). ISBN 905209-48-7
5. Koprucki, Th., Gärtner, K.: Discretization scheme for drift-diffusion equations with strong diffusion enhancement. Opt. Quant. Electron. **45**(7), 791–796 (2013)
6. Patriarca, M., Farrell, P., Fuhrmann, J., Koprucki, Th.: Highly accurate quadrature-based Scharfetter-Gummel schemes for charge transport in degenerate semiconductors. Comput. Phys. Commun. **235**, 40–49 (2019)
7. Bessemoulin-Chatard, M., Filbet, F.: A finite volume scheme for nonlinear degenerate parabolic equations. SIAM J. Sci. Comp. **34**(5), B559–B583 (2012)

The Two-Point Finite Volume Scheme for the Microscopic Bidomain Model of Electrocardiology

Zeina Chehade and Yves Coudière

Abstract We are interested in the approximation of a cardiac microscopic model. It is a set of Laplace equations with non-standard, time-dependent, transmission conditions, for which finite volume methods are of real interest. The transmission conditions state as ordinary differential equations on the jump of the potential, namely the transmembrane voltage, so that we keep this voltage as an unknown in our scheme. Here we extend the two-point flux approximation to the discretization of this model, show that it converges, and compute error estimates.

Keywords Finite volumes · Error estimate · Cardiac EMI model

1 Introduction

The standard mathematical model for describing the spread of excitation of a cardiac tissue is a system of reaction-diffusion equations, obtained by homogenization of a microscopic model. It fails to capture effects of small scale tissue organization, relevant to cardiac fibrillation [3]. For these reasons, we aim at finding numerical approximations of the microscopic cardiac model, called the extracellular-membrane-intracellular (EMI) model, as defined in [1, 5].

The general EMI model is a set of $N_C + 1$ Laplace equations. For sake of simplicity, we will work only with $N_C = 1$ cell, denoted Ω_1 within a tissue sample Ω ($\overline{\Omega_1} \subset \Omega$). The extracellular matrix (ECM) is denoted by $\Omega_0 := \Omega\backslash\overline{\Omega_1}$ (see Fig. 1). Scalar electrical conductivity coefficients $\sigma_0 > 0$ and $\sigma_1 > 0$ are given in the two subdomains. The main physical unknowns are the electrical potentials u_0 and u_1. The normal flux of current at the interface between subdomains is continuous, whereas

Z. Chehade (✉) · Y. Coudière
University of Bordeaux, CNRS, Inria, Bordeaux INP, IMB, UMR 5251, IHU Liryc, 33400 Talence, France
e-mail: zeina.chehade@u-bordeaux.fr

Y. Coudière
e-mail: yves.coudiere@u-bordeaux.fr

E. Franck et al. (eds.), *Finite Volumes for Complex Applications X—Volume 1, Elliptic and Parabolic Problems*, Springer Proceedings in Mathematics & Statistics 432,
https://doi.org/10.1007/978-3-031-40864-9_18

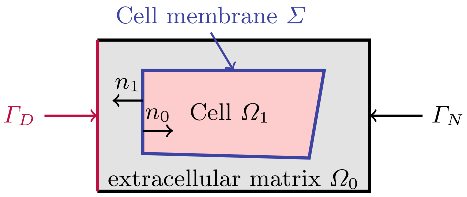


Fig. 1 Geometrical setup of the EMI model with one cell

the electrical potential jumps between subdomains. This jump defines the voltage $v = u_1 - u_0$ on the cell membrane $\Sigma := \overline{\Omega}_0 \cap \overline{\Omega}_1$. The system is closed by ordinary differential equations (ODEs) that link the flux of current to the voltage on the cell membrane. The equations read

$$-\sigma_k \Delta u_k = 0, \qquad \text{in } \Omega_k,\ k = 0, 1, \tag{1}$$

$$-\sigma_0 \nabla u_0 \cdot n_0 = \sigma_1 \nabla u_1 \cdot n_1 = -\left(c_m \partial_t v + f(v)\right), \qquad \text{on } \Sigma, \tag{2}$$

where the unit vectors n_0 and n_1 are normal to Σ outward to Ω_0 and Ω_1, as depicted in Fig. 1. The function f defines the ionic current, and models membrane electrophysiology, while $c_m > 0$ is the membrane surface capacitance. In general, the function f also depends on a set of state variables w defined on Σ (i.e. $f = f(v, w)$), and the model is closed by a system of ODE, $\partial_t w = g(v, w)$ on Σ. Here we consider a simplified model without state variables w. The equations are completed by mixed boundary conditions $-\sigma_0 \nabla u_0 \cdot n_0 = g^N$ on Γ^N, and $u_0 = g^D$ on Γ^D, where $\Gamma^D \cup \Gamma^N = \partial\Omega$ is the boundary of the tissue sample. The equations are supplemented with initial data on Σ, $v(0, \cdot) = v^0$, and we look for a solution u_0, u_1 for $t \in [0, T]$ for any $T > 0$. From previous work [1], we know that there exists a weak solution in $L^2(0, T; H^1(\Omega_0) \times H^1(\Omega_1))$.

There has been some attempts at solving this system of equations by finite elements or boundary element methods [1, 4, 5], but not with finite volumes methods (FVM). Since the dynamics of the phenomena is written on the interface between subdomains (through jump conditions), and the model relies on the continuity of the current flux we believe that FVM are a relevant choice for the EMI model. We start here by studying the simple two-point flux approximation (TPFA) for Eqs (1) and (2). It is essential to carefully write the discrete flux on the cell membrane Σ. Following the usual track, and assuming enough regularity (see Sect. 4), we are able to compute some error estimates for the jump v in $L^\infty(0, T; L^2(\Sigma))$, as well as in a discrete $L^2(0, T, H^1)$-like norm for u_0 and u_1. For this purpose, we assume that Ω_0 and Ω_1 are polygonal domains.

Sections 2, 3, and 4 describe the scheme, its well-posedness, and the error estimates; Sect. 5 discusses the current result, and on-going work.

2 The Numerical Scheme

We consider a standard FV-admissible mesh $\mathcal{T}$, as defined in e.g. [2], with control volumes $K \in \mathcal{T}$, and cell centers $(x_K)_{K\in\mathcal{T}}$. We assume that the mesh $\mathcal{T}$ is consistent with the subdomains Ω_0 and Ω_1: any $K \in \mathcal{T}$ is such that, either $K \subset \Omega_0$, or $K \subset \Omega_1$, so that we define $\mathcal{T}_i = \{K \in \mathcal{T},\ K \subset \Omega_i\}$ $(i = 0, 1)$. We also split the set of the mesh interfaces into: the set $\mathcal{E}^\star$ of interfaces $e = K|L$ (using notations from [2]) such that $(K, L) \in \mathcal{T}_i \times \mathcal{T}_i$ for some $i \in \{0, 1\}$; the set $\mathcal{E}^\Sigma$ of interfaces $e = K|L$ such that $(K, L) \in \mathcal{T}_0 \times \mathcal{T}_1$; and the set $\mathcal{E}^D$ (resp. $\mathcal{E}^N$) of the boundary faces $e = K|$ for which $K \subset \Omega_0$ and $e \subset \Gamma^D$ (resp. Γ^N). Let $\mathcal{E} = \mathcal{E}^\star \cup \mathcal{E}^\Sigma \cup \mathcal{E}^D \cup \mathcal{E}^N$, and denote by $(x_e)_{e\in\mathcal{E}}$ the points x_e at the intersection of the interface e and the line (x_K, x_L) if $e \in \mathcal{E}^\star \cup \mathcal{E}^\Sigma$, and the perpendicular projection of x_K on e if $e \in \mathcal{E}^D \cup \mathcal{E}^N$. Regarding the time discretization, we set $\Delta t = \frac{T}{N}$, for any $N > 0$ and $t^n = n\Delta t$, for $n = 0, \dots N$.

An edge $e = K|L \in \mathcal{E}^\Sigma$ may be such that $(K, L) \in \mathcal{T}_0 \times \mathcal{T}_1$ or $(K, L) \in \mathcal{T}_1 \times \mathcal{T}_0$. *Below, we always use the convention that $K \in \mathcal{T}_0$ and $L \in \mathcal{T}_1$.*

The integral form of Eq. (1) on any cell $K \in \mathcal{T}$ at time t^n reads

$$-\sum_{e\in\mathcal{E}_K} \int_e \sigma_k \nabla u_k(t^n, \cdot) \cdot n_{Ke} = 0, \tag{3}$$

where $\mathcal{E}_K$ is the set of interfaces $e \in \mathcal{E}$ that form the boundary of $K \in \mathcal{T}$, and the vector n_{Ke} is the unit normal to e outward of K. The integral of Eq. (2) on any interface $e = K|L \in \mathcal{E}^\Sigma$ (with the above convention) at time t^n reads

$$\begin{aligned} -\int_e \sigma_0 \nabla u_0(t^n, \cdot) \cdot n_{Ke} = \int_e \sigma_1 \nabla u_1(t^n, \cdot) \cdot n_{Le} \\ = -\left(c_m \int_e \partial_t v(t^n, \cdot) + \int_e f(v(t^n, \cdot))\right). \end{aligned} \tag{4}$$

There are $N_\mathcal{T} = \mathrm{card}(\mathcal{T})$ equations (3), and $N_\Sigma = \mathrm{card}(\mathcal{E}^\Sigma)$ equations (4), so that we take $N_\mathcal{T} + N_\Sigma$ unknowns at each time t^n $(n = 1 \dots N)$, which are the vectors $u^n_\mathcal{T} := (u^n_K)_{K\in\mathcal{T}} \in \mathbb{R}^{N_\mathcal{T}}$, approximating the $u(t^n, x_K)$, and $v^n_{\mathcal{E}^\Sigma} := (v^n_e)_{e\in\mathcal{E}^\Sigma} \in \mathbb{R}^{N_\Sigma}$, approximating the $v(t^n, x_e)$. Using a semi-implicit Euler time-stepping method, our numerical scheme states, for $n = 1, 2 \dots N$, as the $N_\mathcal{T} + N_\Sigma$ equations

$$-\sum_{e\in\mathcal{E}_K} F^n_{Ke} = 0, \qquad \text{for all } K \in \mathcal{T}, \tag{5}$$

$$-F^n_{Ke} = F^n_{Le} = -\left(\frac{c_m}{\Delta t}(v^n_e - v^{n-1}_e) + f(v^{n-1}_e)\right)|e|\,, \quad \text{for all } e = K|L \in \mathcal{E}^\Sigma, \tag{6}$$

with some given initial values $v^0_{\mathcal{E}^\Sigma} = (v^0_e)_{e\in\mathcal{E}^\Sigma} \in \mathbb{R}^{N_\Sigma}$. The numerical flux formula F^n_{Ke} for $e \in \mathcal{E}_K$ and $K \in \mathcal{T}$ approximate the exact flux:

$$F^n_{Ke} = \tau_e(u^n_L - u^n_K) \qquad \text{if } e = K|L \in \mathcal{E}_K \cap \mathcal{E}^\star,\ K \in \mathcal{T}, \tag{7}$$

$$F^n_{Ke} = \tau_e(u^n_L - u^n_K - v^n_e) \qquad \text{if } e = K|L \in \mathcal{E}_K \cap \mathcal{E}^\Sigma,\ K \in \mathcal{T}_0,\ L \in \mathcal{T}_1, \tag{8}$$

$$F^n_{Ke} = \tau_e(g^{D,n}_e - u^n_K) \qquad \text{if } e = K| \in \mathcal{E}_K \cap \mathcal{E}^D,\ K \in \mathcal{T}_0, \tag{9}$$

$$F^n_{Ke} = -\, g^{N,n}_e\, |e| \qquad \text{if } e = K| \in \mathcal{E}_K \cap \mathcal{E}^N,\ K \in \mathcal{T}_0. \tag{10}$$

In these expressions, the coefficient τ_e has the usual value $\tau_e = \frac{\tau_{Ke}\tau_{Le}}{\tau_{Ke}+\tau_{Le}}$ if $e = K|L \in \mathcal{E}^\star \cup \mathcal{E}^\Sigma$, and $\tau_e = \tau_{Ke}$ if $e = K| \in \mathcal{E}^D$, where $\tau_{Ke} = \frac{\sigma_i |e|}{d_{Ke}}$ for all $e \in \mathcal{E}_K$ and $K \in \mathcal{T}_i$ $(i = 0, 1)$. The quantities $|e|$ and d_{Ke} are the measure of the interface e and the Euclidean distance $d_{Ke} = \mathrm{d}(x_K, e)$. At last, we take $g^{D,n}_e = g^D(t^n, x_e)$ and $g^{N,n}_e |e| = \int_e g^N(t^n, \cdot)$.

Expressions (7), (9), and (10) are standard, but expression (8) is obtained on an interface $e = K|L$ between $K \in \mathcal{T}_0$ and $L \in \mathcal{T}_1$ after introducing two auxiliary unknowns, named $u_{K,e}$ and $u_{L,e}$, approximating $u_0(t^n, x_e)$ and $u_1(t^n, x_e)$, and flux expressions $F^n_{Ke} = \tau_{Ke}(u_{K,e} - u_K)$ and $F^n_{Le} = \tau_{Le}(u_{L,e} - u_L)$. The auxiliary unknowns are eliminated with the conservation and jump conditions:

$$\tau_{Ke}(u_{K,e} - u^n_K) + \tau_{Le}(u_{L,e} - u^n_L) = 0, \quad u_{L,e} - u_{K,e} = v^n_e. \tag{11}$$

Instead of the usual solution, we find that $u_{K,e} = \frac{\tau_{Ke}u^n_K + \tau_{Le}u^n_L - \tau_{Le}v^n_e}{\tau_{Ke}+\tau_{Le}}$, and $u_{L,e} = \frac{\tau_{Ke}u^n_K + \tau_{Le}u^n_L + \tau_{Ke}v^n_e}{\tau_{Ke}+\tau_{Le}}$, and then Formula (8) for the flux.

3 Discrete Norms, Discrete Problem, Coercivity

In this section, we establish existence and uniqueness of the discrete solution.

First, from Eqs. (5) and (6), and the definitions (7)–(10), we observe that the scheme writes as the following block linear system, at each time step

$$\begin{pmatrix} A & B \\ B^\top & C + \frac{c_m}{\Delta t} D \end{pmatrix} \begin{pmatrix} u^n_{\mathcal{T}} \\ v^n_{\mathcal{E}^\Sigma} \end{pmatrix} = \begin{pmatrix} G^n \\ -\left(-\frac{c_m}{\Delta t} v^{n-1}_{\mathcal{E}^\Sigma} + f(v^{n-1}_{\mathcal{E}^\Sigma})\right) |e| \end{pmatrix}, \tag{12}$$

where $A \in \mathbb{R}^{N_\mathcal{T} \times N_\mathcal{T}}$ is the usual TPFA matrix, with nonzero entries $a_{KL} = -\tau_e$ if $e = K|L$, and $a_{KK} = \sum_{e=K|L \in \mathcal{E}_K} \tau_e$. The matrix $B \in \mathbb{R}^{N_\mathcal{T} \times N_\Sigma}$ has nonzero entries $b_{Ke} = \tau_e$, and $b_{Le} = -\tau_e$ (for $e \in \mathcal{E}^\Sigma$), and the matrices $C,\ D \in \mathbb{R}^{N_\Sigma \times N_\Sigma}$ are diagonal, with $c_{ee} = \tau_e$ and $d_{ee} = |e|$ (for $e \in \mathcal{E}^\Sigma$). The vector $G^n \in \mathbb{R}^{N_\mathcal{T}}$ gathers the contributions of the boundary data.

We multiply scalarly Eq. (12) by the unknown vector $\left(u^n_\mathcal{T}\ v^n_{\mathcal{E}^\Sigma}\right)^\top$, or equivalently, we multiply Eq. (5) by u^n_K and sum over $K \in \mathcal{T}$, multiply Eq. (6) by v^n_e and sum over $e \in \mathcal{E}^\Sigma$. After reordering the summation over the set of edges, it yields

$$\sum_{e\in\mathcal{E}^\star} \tau_e|u_L^n - u_K^n|^2 + \sum_{e\in\mathcal{E}^\Sigma} \tau_e|u_L^n - u_K^n - v_e^n|^2 + \sum_{e\in\mathcal{E}^D} \tau_e|u_K^n|^2 + \sum_{e\in\mathcal{E}^\Sigma} \frac{c_m}{\Delta t}\left|v_e^n\right|^2 |e|$$
$$= \sum_{e\in\mathcal{E}^D} \tau_e g_e^{D,n} u_K^n - \sum_{e\in\mathcal{E}^N} g_e^{N,n} u_K^n\,|e| + \sum_{e\in\mathcal{E}^\Sigma} \frac{c_m}{\Delta t}\left(v_e^{n-1} - f(v_e^{n-1})\right) v_e^n\,|e|\,. \quad (13)$$

Equation (13) shows that the linear system (12) is symmetric and positive-definite. In addition, we can introduce the semi-norm of a vector $(u_\mathcal{T}, v_{\mathcal{E}^\Sigma}) = ((u_K)_{K\in\mathcal{T}}, (v_e)_{e\in\mathcal{E}^\Sigma})$ by

$$|(u_\mathcal{T}, v_{\mathcal{E}^\Sigma})|_{1,\mathcal{T}}$$
$$:= \left(\sum_{e\in\mathcal{E}^\star} \tau_e|u_L - u_K|^2 + \sum_{e\in\mathcal{E}^\Sigma} \tau_e|u_L - u_K - v_e|^2 + \sum_{e\in\mathcal{E}^D} \tau_e|u_K|^2\right)^{1/2}, \quad (14)$$

This latter formula defines a norm, since if $|(u_\mathcal{T}, v_{\mathcal{E}^\Sigma})|_{1,\mathcal{T}} = 0$, one can observes that $u_K = 0$ for all $K \in \mathcal{T}_0$, and there exists $u \in \mathbb{R}$ such that $u_L = u$ for all $L \in \mathcal{T}_1$, and $v_e = u$ for all $e \in \mathcal{E}^\Sigma$. Consequently, if $|(u_\mathcal{T}, v_{\mathcal{E}^\Sigma})|_{1,\mathcal{T}} = 0$, there exists $u \in \mathbb{R}$ such that $u_K = v_e = u$ for all $K \in \mathcal{T}$ and $e \in \mathcal{E}^\Sigma$.

Using these norms, Eq. (13) rewrites as:

$$\left|(u_\mathcal{T}^n, v_{\mathcal{E}^\Sigma}^n)\right|_{1,\mathcal{T}}^2 + \frac{c_m}{\Delta t}\|v_{\mathcal{E}^\Sigma}^n\|_{0,\Sigma}^2 = \left(u_\mathcal{T}^{n\top}\ v_{\mathcal{E}^\Sigma}^{n\top}\right)\begin{pmatrix} A & B \\ B^T & C + \frac{c_m}{\Delta t}D\end{pmatrix}\begin{pmatrix} u_\mathcal{T}^n \\ v_{\mathcal{E}^\Sigma}^n\end{pmatrix}. \quad (15)$$

Finally, the FV formulation allows to write the problem as an evolution problem on Σ only, computing the Schur complement of (12) and iterating with

$$\left(C - B^\top A^{-1}B + \frac{c_m}{\Delta t}D\right) v_{\mathcal{E}^\Sigma}^n = \left(\frac{c_m}{\Delta t} v_{\mathcal{E}^\Sigma}^{n-1} - f(v_{\mathcal{E}^\Sigma}^{n-1})\right)|e| - B^\top A^{-1} G^n. \quad (16)$$

4 Convergence Analysis

In this section, a convergence analysis is carried out by proving the following error estimates theorem.

Theorem 1 *Assume that f is Lipschitz continuous in $\mathbb{R}$. Given $T > 0$, and the discrete setup from above, assume that $u_k \in C^2([0, T] \times \overline{\Omega}_k)$, and consider $\overline{u}_\mathcal{T}^n$ and $\overline{v}_{\mathcal{E}^\Sigma}^n$ defined by $\overline{u}_K^n = u_i(t^n, x_K)$ for all $K \in \mathcal{T}_i$ $(i = 0, 1)$, and $\overline{v}_e^n = u_1(t^n, x_e) - u_0(t^n, x_e)$ for all $e \in \mathcal{E}^\Sigma$, and $n = 0 \ldots N$. The discrete errors $(\epsilon_\mathcal{T}^n, \eta_{\mathcal{E}^\Sigma}^n)$ are defined by $\epsilon_\mathcal{T}^n = \overline{u}_\mathcal{T}^n - u_\mathcal{T}^n$ and $\eta_{\mathcal{E}^\Sigma}^n = \overline{v}_{\mathcal{E}^\Sigma}^n - v_{\mathcal{E}^\Sigma}^n$. We associate to $(\eta_{\mathcal{E}^\Sigma}^n)_n$ the $L^\infty(0, T; L^2(\Sigma))$ function $\eta_{\mathcal{E}^\Sigma}(t, x) = \eta_e^n$ for $t \in]t^{n-1}, t^n[$ $(n = 1 \ldots N)$, and $x \in e$ $(e \in \mathcal{E}^\Sigma)$. If the initial approximation $v_{\mathcal{E}^\Sigma}^0$ is such that $\|v^0 - v_{\mathcal{E}^\Sigma}^0\|_{0,\Sigma} \le Ch$, then there exists $C > 0$, depending only on the data, such that*

$$\|\eta_{\mathcal{E}^\Sigma}\|_{L^\infty(0,T;L^2(\Sigma))} \le C(h+\Delta t), \quad \left(\sum_{n=1}^{N} \Delta t \left|(\epsilon_{\mathcal{T}}^n, \eta_{\mathcal{E}^\Sigma}^n)\right|_{1,\mathcal{T}}\right)^{1/2} \le C(h+\Delta t), \quad (17)$$

where we have defined $h = \max_{K\in\mathcal{T}} \operatorname{diam}(K)$.

Proof We denote by $\widehat{F}_{Ke}^n := \int_e \sigma_i \nabla u_i \cdot n_{Ke}$ the exact flux out of cell $K \in \mathcal{T}_i$ ($i = 0, 1$) through the edge $e \in \mathcal{E}_K$, and by $\overline{F}_{Ke}^n$ the flux associated to $(\overline{u}_{\mathcal{T}}^n, \overline{v}_{\mathcal{E}^\Sigma}^n)$ with Formulas (7)–(10). By subtracting Eqs. (5) and (6) defining the approximation solution, from Eqs. (3) and (4) verified by the exact solution, the equations on the error vectors $(\epsilon_{\mathcal{T}}^n, \eta_{\mathcal{E}^\Sigma}^n)$ for $n = 1 \dots N$, write

$$-\sum_{e\in\mathcal{E}_K} (\overline{F}_{Ke}^n - F_{Ke}^n) = -\sum_{e\in\mathcal{E}_K} R_{Ke}^n \,|e|\,, \quad (18)$$

on all $K \in \mathcal{T}$, where the consistency error is $R_{Ke}^n = \frac{1}{|e|}\left(\overline{F}_{Ke}^n - \widehat{F}_{Ke}^n\right)$; and

$$\begin{aligned} -(\overline{F}_{Ke}^n - F_{Ke}^n) + |e|\, R_{Ke}^n = (\overline{F}_{Le}^n - F_{Le}^n) - |e|\, R_{Le}^n &= -\left(c_m T_e^n + S_e^n\right)|e| \\ &\quad -\left(\frac{c_m}{\Delta t}(\eta_e^n - \eta_e^{n-1}) + f(\overline{v}_e^{n-1}) - f(v_e^{n-1})\right)|e|\,, \quad (19) \end{aligned}$$

on all $e = K|L \in \mathcal{E}^\Sigma$, where the additional consistency errors are defined by

$$T_e^n = \frac{1}{|e|}\int_e \partial_t v(t^n,\cdot) - \frac{\overline{v}_e^n - \overline{v}_e^{n-1}}{\Delta t}, \quad S_e^n = \frac{1}{|e|}\int_e f(v(t^n,\cdot)) - f(\overline{v}_e^{n-1}). \quad (20)$$

For $n = 0$, we have $\eta_{\mathcal{E}^\Sigma}^0 = \overline{v}_{\mathcal{E}^\Sigma}^0 - v_{\mathcal{E}^\Sigma}^0$. We multiply by $\epsilon_{\mathcal{T}}^n$ and $\eta_{\mathcal{E}^\Sigma}^n$ and obtain (like we obtained (13) and the coercivity inequality (15))

$$\begin{aligned} \left|(\epsilon_{\mathcal{T}}^n, \eta_{\mathcal{E}^\Sigma}^n)\right|_{1,\mathcal{T}}^2 + \frac{c_m}{\Delta t}\|\eta_{\mathcal{E}^\Sigma}^n\|_{0,\Sigma}^2 &\le \left(\frac{c_m}{\Delta t}\eta_{\mathcal{E}^\Sigma}^{n-1} - \left(f(\overline{v}_{\mathcal{E}^\Sigma}^{n-1}) - f(v_{\mathcal{E}^\Sigma}^{n-1})\right), \eta_{\mathcal{E}^\Sigma}^n\right)_{0,\Sigma} \\ + \sum_{e=K|L\in\mathcal{E}^\star} |e|\, R_{Ke}^n(\epsilon_L^n - \epsilon_K^n) &+ \sum_{e=K|\in\mathcal{E}^D} |e|\, R_{Ke}^n(-\epsilon_K^n) + \sum_{e=K|\in\mathcal{E}^N} |e|\, R_{Ke}^n(-\epsilon_K^n) \\ &+ \sum_{e=K|L\in\mathcal{E}^\Sigma} |e|\, R_{Ke}^n(\epsilon_L^n - \epsilon_K^n - \eta_e^n) - \left(c_m T_{\mathcal{E}^\Sigma}^n + S_{\mathcal{E}^\Sigma}^n, \eta_{\mathcal{E}^\Sigma}^n\right)_{0,\Sigma}\,, \end{aligned}$$

where $(\cdot,\cdot)_{0,\Sigma}$ denotes the natural scalar product on $\mathcal{E}^\Sigma$ associated to $\|\cdot\|_{0,\Sigma}$, and $T_{\mathcal{E}^\Sigma}^n := (T_e^n)_{e\in\mathcal{E}^\Sigma} \in \mathbb{R}^{N_\Sigma}$ and $S_{\mathcal{E}^\Sigma}^n := (S_e^n)_{e\in\mathcal{E}^\Sigma} \in \mathbb{R}^{N_\Sigma}$. We have $R_{Ke}^n = 0$ for $e \in \mathcal{E}^N$ (see def. of $g_e^{N,n}$ above). We have $\left|f(\overline{v}_e^{n-1}) - f(v_e^{n-1})\right| \le \lambda \eta_e^{n-1}$ for all $e \in \mathcal{E}^\Sigma$, where $\lambda > 0$ is the Lipschitz constant for f, and then we obtain

$$\Delta t \left|(\epsilon^n_{\mathcal{T}}, \eta^n_{\mathcal{E}^\Sigma})\right|^2_{1,\mathcal{T}} + c_m \|\eta^n_{\mathcal{E}^\Sigma}\|^2_{0,\Sigma} \leq c_m \left(1 + \frac{\lambda}{c_m}\Delta t\right) \|\eta^{n-1}_{\mathcal{E}^\Sigma}\|_{0,\Sigma} \|\eta^n_{\mathcal{E}^\Sigma}\|_{0,\Sigma}$$
$$+ \Delta t R^n \left|(\epsilon^n_{\mathcal{T}}, \eta^n_{\mathcal{E}^\Sigma})\right|_{1,\mathcal{T}} + \Delta t \left(c_m \|T^n_{\mathcal{E}^\Sigma}\|_{0,\Sigma} + \|S^n_{\mathcal{E}^\Sigma}\|_{0,\Sigma}\right) \|\eta^n_{\mathcal{E}^\Sigma}\|_{0,\Sigma}, \quad (21)$$

where $R^n = R^n(u_0, u_1) = \left(\sum_{e\in\mathcal{E}^\star\cup\mathcal{E}^D\cup\mathcal{E}^\Sigma} \frac{|e|^2}{\tau_e} \left|R^n_{Ke}\right|^2\right)^{1/2}$. Using Young inequalities, we first prove that

$$\frac{\Delta t}{2} \left|(\epsilon^n_{\mathcal{T}}, \eta^n_{\mathcal{E}^\Sigma})\right|^2_{1,\mathcal{T}} + \frac{c_m}{2} \|\eta^n_{\mathcal{E}^\Sigma}\|^2_{0,\Sigma} \leq \frac{c_m}{2} \left(1 + \frac{\lambda}{c_m}\Delta t\right)^2 \|\eta^{n-1}_{\mathcal{E}^\Sigma}\|^2_{0,\Sigma}$$
$$+ \frac{\Delta t}{2} \left|R^n\right|^2 + c_m \Delta t \left(\|T^n_{\mathcal{E}^\Sigma}\|_{0,\Sigma} + \frac{1}{c_m}\|S^n_{\mathcal{E}^\Sigma}\|_{0,\Sigma}\right) \|\eta^n_{\mathcal{E}^\Sigma}\|_{0,\Sigma}. \quad (22)$$

The last term is bounded for any $\alpha > 0$ as follows:

$$c_m \Delta t \left(\|T^n_{\mathcal{E}^\Sigma}\|_{0,\Sigma} + \frac{1}{c_m}\|S^n_{\mathcal{E}^\Sigma}\|_{0,\Sigma}\right)\|\eta^n_{\mathcal{E}^\Sigma}\|_{0,\Sigma} \leq \frac{c_m\alpha}{2} \|\eta^n_{\mathcal{E}^\Sigma}\|^2_{0,\Sigma}$$
$$+\frac{c_m}{2\alpha}\Delta t^2 \left(\|T^n_{\mathcal{E}^\Sigma}\|_{0,\Sigma} + \frac{1}{c_m}\|S^n_{\mathcal{E}^\Sigma}\|_{0,\Sigma}\right)^2. \quad (23)$$

We combine (22) and (23), choosing $\alpha > 0$ such that $1 - \alpha = \frac{1}{1+\frac{\lambda}{c_m}\Delta t}$, to obtain

$$\Delta t \left|(\epsilon^n_{\mathcal{T}}, \eta^n_{\mathcal{E}^\Sigma})\right|^2_{1,\mathcal{T}} + \frac{c_m}{1 + \frac{\lambda}{c_m}\Delta t} \|\eta^n_{\mathcal{E}^\Sigma}\|^2_{0,\Sigma} \leq c_m \left(1 + \frac{\lambda}{c_m}\Delta t\right)^2 \|\eta^{n-1}_{\mathcal{E}^\Sigma}\|^2_{0,\Sigma}$$
$$+ \Delta t \left|R^n\right|^2 + \left(1 + \frac{\lambda}{c_m}\Delta t\right) \frac{2\Delta t}{\lambda} \left(c_m^2 \|T^n_{\mathcal{E}^\Sigma}\|^2_{0,\Sigma} + \|S^n_{\mathcal{E}^\Sigma}\|^2_{0,\Sigma}\right). \quad (24)$$

From (24) we extract an estimation on $\eta^n_{\mathcal{E}^\Sigma}$, and a recurrence shows that

$$\|\eta^n_{\mathcal{E}^\Sigma}\|^2_{0,\Sigma} \leq \left(1 + \frac{\lambda}{c_m}\Delta t\right)^{3n} \|\eta^0_{\mathcal{E}^\Sigma}\|^2_{0,\Sigma} + \frac{\Delta t}{c_m} \sum_{i=0}^{n-1} \left(1 + \frac{\lambda}{c_m}\Delta t\right)^{3i+1} |R^{n-i}|^2$$
$$+ \frac{2\Delta t}{\lambda c_m} \sum_{i=0}^{n-1} \left(1 + \frac{\lambda}{c_m}\Delta t\right)^{3i+2} \left(c_m^2 \|T^n_{\mathcal{E}^\Sigma}\|^2_{0,\Sigma} + \|S^n_{\mathcal{E}^\Sigma}\|^2_{0,\Sigma}\right),$$

which proves that, for all $n = 1 \ldots N$,

$$\|\eta^n_{\mathcal{E}^\Sigma}\|_{0,\Sigma} \leq \exp\left(\frac{3}{2}\frac{\lambda}{c_m}T\right) \left(\|\eta^0_{\mathcal{E}^\Sigma}\|^2_{0,\Sigma} + \frac{1}{c_m}R^2_{\mathcal{T}} + \frac{2c_m}{\lambda}T^2_{\mathcal{E}^\Sigma} + \frac{2}{\lambda c_m}S^2_{\mathcal{E}^\Sigma}\right)^{1/2},$$

with $R_{\mathcal{T}}^2 = \Delta t \sum_{n=1}^N |R^n|^2$, $S_{\mathcal{E}^\Sigma}^2 = \Delta t \sum_{n=1}^N \|S_{\mathcal{E}^\Sigma}^n\|_{0,\Sigma}^2$, $T_{\mathcal{E}^\Sigma}^2 = \Delta t \sum_{n=1}^N \|T_{\mathcal{E}^\Sigma}^n\|_{0,\Sigma}^2$.

In the second step, we start again the computation from inequality (21), and after some standard calculations, we show that

$$\Delta t \sum_{n=1}^N \left|(\epsilon_{\mathcal{T}}^n, \eta_{\mathcal{E}^\Sigma}^n)\right|_{1,\mathcal{T}}^2 \le c_m \|\eta_{\mathcal{E}^\Sigma}^0\|_{0,\Sigma}^2 + 2c_m^2 T_{\mathcal{E}^\Sigma}^2 + 2S_{\mathcal{E}^\Sigma}^2 + 2(\lambda+1)T \max_{n=1\ldots N} \|\eta_{\mathcal{E}^\Sigma}^n\|_{0,\Sigma}^2.$$

It remains to estimate the consistency errors $R_{\mathcal{T}}$, $S_{\mathcal{E}^\Sigma}$, and $T_{\mathcal{E}^\Sigma}$. Assuming C^2 regularity, and the Lipschitz continuity of f, the estimates for S_e^n and T_e^n are obtained by the usual Taylor expansions (see Eq.(20)): $\left|S_e^n\right| \le \lambda \|v\|_{2,\infty} (h + \Delta t)$, and $\left|T_e^n\right| \le C\|v\|_{2,\infty} (h + \Delta t)$, where $\|v\|_{2,\infty}$ denotes its C^2 uniform norm. Standard results (see [2]) are used to estimate the terms R_{Ke}^n in all cases excepts for $e = K|L \in \mathcal{E}^\Sigma$. In this case,

$$\begin{aligned} \widehat{F}_{Ke}^n &= \int_e \sigma_0 \nabla u_0(t^n, \cdot) \cdot n_{Ke} = |e|\, \sigma_0 \nabla u_0(t^n, x_e) \cdot n_{Ke} + |e|\, A_K \\ &= +\tau_{Ke}(u_0(t^n, x_e) - u_0(t^n, x_K)) + |e|\, (A_K + B_K) \\ &= -\tau_{Le}(u_1(t^n, x_e) - u_1(t^n, x_L)) + |e|\, (A_K - B_L), \end{aligned} \tag{25}$$

since $\sigma_0 \nabla u_0(t^n, x_e) \cdot n_{Ke} = -\sigma_1 \nabla u_1(t^n, x_e) \cdot n_{Le}$. Here, $|A_K| \le M\|u_0\|_{2,\infty} h$ and $|B_K| \le M\|u_0\|_{2,\infty} h$ (resp. $|B_L| \le M\|u_1\|_{2,\infty} h$) by Taylor expansion, with $M = \max(\sigma_0, \sigma_1)$. Instead of Eq. (11), we have that $\tau_{Ke}(u_0(t^n, x_e) - \overline{u}_K^n) + \tau_{Le}(u_1(t^n, x_e) - \overline{u}_L^n) = -\,|e|\,(B_K + B_L)$ and $\overline{v}_e^n = u_1(t^n, x_e) - u_0(t^n, x_e)$ (recall that $\overline{u}_K^n = u_0(t^n, x_K)$ and $\overline{u}_L^n = u_1(t^n, x_L)$). Hence we find that

$$u_0(t^n, x_e) = \frac{\tau_{Ke}\overline{u}_K^n + \tau_{Le}\overline{u}_L^n - \tau_{Le}\overline{v}_e^n}{\tau_{Ke} + \tau_{Le}} - \frac{|e|}{\tau_{Ke} + \tau_{Le}}(B_K + B_L). \tag{26}$$

The discrete flux associate to $\overline{u}_K^n = u_0(t^n, x_K)$ and $\overline{u}_L^n = u_1(t^n, x_L)$ is

$$\overline{F}_{Ke}^n = -\overline{F}_{Le}^n = \tau_{Ke}(\overline{u}_{K,e}^n - \overline{u}_K^n) = -\tau_{Le}(\overline{u}_{L,e}^n - \overline{u}_L^n), \tag{27}$$

where $\overline{u}_{K,e}^n$ and $\overline{u}_{L,e}^n$ are defined as in (11) (with $\overline{v}_e^n = u_1(t^n, x_e) - u_0(t^n, x_e)$). In view of Eq. (26), we have $u_0(t^n, x_e) = \overline{u}_{K,e}^n - \frac{|e|}{\tau_{Ke}+\tau_{Le}}(B_K + B_L)$.

At last, we subtract (27)–(25), and we use the previous remark, to obtain the last estimate: $R_{K,e}^n = -\frac{\tau_{Ke}}{|e|}(u_0(t^n, x_e) - \bar{u}_{K,e}^n) - (A_K + B_K) = \frac{\tau_{Ke}B_L - \tau_{Le}B_K}{\tau_{Ke}+\tau_{Le}} - A_K$, then $|R_{K,e}^n| \le \max(|B_K|, |B_L|) + |A_K| \le 2M \max(\|u_0\|_{2,\infty}, \|u_1\|_{2,\infty})h$.

5 Conclusion and Perspectives

In conclusion, the TPFA naturally generalizes to the EMI model. An error estimate was obtained on the transmembrane voltage v on the interface Σ, though under strong regularity assumptions. The convergence analysis, without error estimate, might be easily deduced from the existence proof from [1], and obtained with minimal data regularity. We believe that the extension to many cells ($N_C > 1$) is straightforward (as in [1]), and also that the consistency estimates generalize to functions $u_i \in H^2(\Omega_i)$ and $v \in H^1(0, T; L^2(\Sigma))$.

In practice, the scheme is intended to be used for simulating billions of cardiac cells, using HPC solutions developed within the EuroHPC MICROCARD consortium. In reality, the ODEs from Eq. (2) are coupled with large sets of nonlinear equations that needs time-integration specific to cardiac models. The FVM simplify the implementation of this coupling, by introducing the voltage v as an explicit unknown. Moreover, a a Dirichlet-to-Neumann operator on Σ can be introduced to rewrite the problem as an evolution equation on the membrane only. This give rise to approximation with boundary element methods. Again, the discrete Dirichlet-to-Neumann operator may be reconstructed with the FVM, as in Eq. (16).

On a short term, we plan to implement the scheme, and compare it to finite elements and a boundary elements discretizations on simple 2D and 3D test cases. Afterwards, we would like to generalize the scheme to a FV-like method that is more robust with respect to the available meshes.

Acknowledgements This project is funded by the European High-Performance Computing Joint Undertaking EuroHPC, under grant agreement No 955495.

References

1. Becue, P.E.: Modélisation et simulation de l'électrophysiologie cardiaque à l'échelle microscopique. Theses, Université de Bordeaux (2018). https://theses.hal.science/tel-02019648
2. Eymard, R., Gallouët, T., Herbin, R.: Finite volume methods. Handbook of Numer. Anal. **7**, 713–1018 (2000)
3. Haïssaguerre, M., al: Localized structural alterations underlying a subset of unexplained sudden cardiac death. Circul.: Arrhythmia Electrophysiol. **11**(7) (2018)
4. Mori, Y., Peskin, C.S.: A numerical method for cellular electrophysiology based on the electrodiffusion equations with internal boundary conditions at membranes. Commun. Appl. Math. Comput. Sci. **4**(1) (2009). https://doi.org/10.2140/camcos.2009.4.85
5. Tveito, A., Jæger, K.H., Kuchta, M., Mardal, K.A., Rognes, M.E.: A cell-based framework for numerical modeling of electrical conduction in cardiac tissue. Front. Phys. **5** (2017). https://doi.org/10.3389/fphy.2017.00048

Improved Crouzeix-Raviart Scheme for the Stokes Problem

Eric Chénier, Erell Jamelot, Christophe Le Potier, and Andrew Peitavy

Abstract The resolution of the incompressible Navier-Stokes equations is tricky, and it is well known that one of the major issue is to compute a divergence free velocity. The non-conforming Crouzeix-Raviart finite element are convenient since they induce piecewise mass conservation and satisfy the inf-sup condition. However, spurious velocities may appear and damage the approximation. In this contribution, we propose a scheme that allows one to reduce the spurious velocities by discretizing the gradient of pressure with a symmetric MPFA scheme (finite volume MultiPoint Flux Approximation) [1, 2].

Keywords Stokes problem · Crouzeix-Raviart scheme · MPFA scheme · Finite element method · Nonconforming method

1 Motivation

The TrioCFD [3] code is a computational fluid dynamics (CFD) simulation software developed at the CEA. It is dedicated to the numerical simulation of turbulent flows for scientific and industrial applications, particularly in the nuclear field. Let Ω, the domain of study, be an open connected bounded domain of $\mathbb{R}^d$, $d = 2, 3$, with a polygonal ($d = 2$) or Lipschitz polyhedral ($d = 3$) boundary Γ with constant

E. Chénier
Université Gustave Eiffel, Université Paris Est Creteil, CNRS, UMR 8208, MSME, 77454
Marne-la-Vallée, France
e-mail: eric.chenier@univ-eiffel.fr

E. Jamelot · C. Le Potier · A. Peitavy (✉)
Université Paris-Saclay, CEA, Service de Thermo-hydraulique et de Mécanique des Fluides,
91191 Gif-sur-Yvette, France
e-mail: andrew.peitavy@cea.fr

E. Jamelot
e-mail: erell.jamelot@cea.fr

C. Le Potier
e-mail: christophe.le-potier@cea.fr

E. Franck et al. (eds.), *Finite Volumes for Complex Applications X—Volume 1, Elliptic and Parabolic Problems*, Springer Proceedings in Mathematics & Statistics 432,
https://doi.org/10.1007/978-3-031-40864-9_19

physical properties. Let $T > 0$ be a simulation time. The TrioCFD code solves the incompressible Navier-Stokes equations which read:

Find $(\mathbf{u}(\mathbf{x}, t), p(\mathbf{x}, t))$ such that $\forall(\mathbf{x}, t) \in \Omega \times (0, T)$,

$$\begin{cases} \partial_t \mathbf{u} - \nu \, \Delta \, \mathbf{u} + (\mathbf{u} \cdot \mathbf{grad}\,)\mathbf{u} + \mathbf{grad}\, p = \mathbf{f}(\mathbf{x}, t), \\ \qquad\qquad\qquad\qquad\qquad \operatorname{div} \mathbf{u} = 0, \\ \qquad\qquad\qquad\qquad u(\mathbf{x}, 0) = u_0(\mathbf{x}). \end{cases} \tag{1}$$

We consider here Dirichlet boundary conditions for the velocity $\mathbf{u}$, and we impose a normalization condition for the pressure p: $\mathbf{u} = 0$ on Γ, $\int_\Omega p = 0$. The vector field $\mathbf{u}$ represents the velocity of the fluid and the scalar field p represents its pressure divided by the fluid density which is supposed to be constant. The first equation of (1) corresponds to the momentum balance equation and the second one corresponds to the mass conservation. The constant parameter $\nu > 0$ is the kinematic viscosity of the fluid. The vector field $\mathbf{f}$ represents the body force divided by the fluid density. We first consider the steady Stokes problem which reads:

$$\text{Find } (\mathbf{u}, p) \text{ such that } \forall \mathbf{x} \in \Omega : \begin{cases} -\nu\Delta\mathbf{u} + \mathbf{grad}\, p = \mathbf{f}, & \text{in } \Omega \\ \qquad\qquad \operatorname{div} \mathbf{u} = 0, & \text{in } \Omega. \end{cases} \tag{2}$$

The resolution of (2) leads to a well-posed saddle point problem [4]. In TrioCFD code, the spatial discretization of Problem (2) is based on first order nonconforming[1] Crouzeix-Raviart finite element method [5] that we call the $\mathbf{P}^1_{nc} - P^0$ scheme. The outline of this article is as follows: in Sect. 2, we provide some notations for the discretization. Next, we recall the $\mathbf{P}^1_{nc} - P^0$ scheme and an improved version implemented in TrioCFD code for simplicial meshes, that we call the $\mathbf{P}^1_{nc} - (P^0 + P^1)$ scheme. This last discretization reduces the spurious velocities in $2D$. It is also efficient in $3D$, except when the source term is a strong gradient. In order to obtain the same accuracy in $3D$ than in $2D$, one must increase the number of degrees of freedom of the discrete pressure space, which leads to a more expensive numerical scheme. Our aim is to develop a new numerical scheme that would reduce the spurious velocities both in $2D$ and $3D$, but at a lower cost. We present such a scheme in Sect. 3 and numerical illustrations in Sect. 4.

2 The $\mathbf{P}^1_{nc} - P^0$ Scheme [5]

We call $(O, (x_{d'})_{d'=1}^d)$ the Cartesian coordinates system, of orthonormal basis $(\boldsymbol{e}_{d'})_{d'=1}^d$. We denote the Γ the boundary of Ω and its outgoing normal vector $\mathbf{n}_\Gamma$. Consider $(\mathcal{T}_h)_h$ a simplicial triangulation sequence of Ω, we use the following index sets:

[1] Let $\mathbf{v}_h$, the discrete velocity obtained with the Crouzeix-Raviart finite element, then $\mathbf{v}_h \notin \mathbf{H}(\operatorname{div} 0, \Omega) := \{\mathbf{v} \in \mathbf{L}^2(\Omega) \,/\, \operatorname{div} \mathbf{v} = 0\}$.

- $\mathcal{I}_K$ (resp. $\mathcal{I}_F$) denotes the index set of the elements (resp. facets[2]), such that $\mathcal{T}_h := \bigcup_{\ell \in \mathcal{I}_K} K_\ell$ (resp. $\mathcal{F}_h := \bigcup_{f \in \mathcal{I}_F} F_f$) is the set of elements (resp. facets).
- $\mathcal{I}_S$ denotes the index set of the vertices, such that $(S_j)_{j \in \mathcal{I}_S}$ is the set of vertices
- $\mathcal{I}_F = \mathcal{I}_F^i \cup \mathcal{I}_F^b$ (resp. $\mathcal{I}_S = \mathcal{I}_S^i \cup \mathcal{I}_S^b$), where $\forall f \in \mathcal{I}_F^i$, $F_f \subset \Omega$ (resp. $\forall j \in \mathcal{I}_S^i$, $S_j \subset \Omega$) and $\forall f \in \mathcal{I}_F^b$, $F_f \subset \Gamma$ (resp. $\forall j \in \mathcal{I}_S^b$, $S_j \subset \Gamma$).

We denote $\mathbb{L}^2(\Omega) = [L^2(\Omega)]^{d \times d}$ and $L^2_{zmv}(\Omega) := \{q \in L^2(\Omega) \mid \int_\Omega q = 0\}$. Let's introduce spaces of piecewise regular elements:
We set $\mathcal{P}_h H^1 = \left\{v \in L^2(\Omega)\,; \quad \forall \ell \in \mathcal{I}_K,\ v_{|K_\ell} \in H^1(K_\ell)\right\}$ and $\mathcal{P}_h \mathbf{H}^1 = [\mathcal{P}_h H^1]^d$, endowed with the scalar product:

$$(\mathbf{v}, \mathbf{w})_h := \sum_{\ell \in \mathcal{I}_K} (\mathbf{Grad}\,\mathbf{v}, \mathbf{Grad}\,\mathbf{w})_{\mathbb{L}^2(K_\ell)} \quad \|\mathbf{v}\|_h^2 = \sum_{\ell \in \mathcal{I}_K} \|\mathbf{Grad}\,\mathbf{v}\|^2_{\mathbb{L}^2(K_\ell)}.$$

Let $f \in \mathcal{I}_F^i$ such that $F_f = \partial K_L \cap \partial K_R$ and let $\mathbf{n}_f$ the unit normal that is outward K_L oriented. The jump of a function $v \in \mathcal{P}_h H^1$ across the facet F_f, in $\mathbf{n}_f$ direction, is defined as follows: $[v]_{F_f} := v_{|K_L} - v_{|K_R}$. For $f \in \mathcal{I}_F^b$, we set: $[v]_{F_f} := v_{|F_f}$.
We also define the operator $\mathrm{div_h}$ such that:

$$\forall \mathbf{v} \in \mathcal{P}_h \mathbf{H}^1,\ \forall q \in L^2(\Omega), \quad (\mathrm{div_h}\,\mathbf{v}, q) = \sum_{\ell \in \mathcal{I}_K} (\mathrm{div}\,\mathbf{v}, q)_{L^2(K_\ell)}.$$

For all $D \subset \mathbb{R}^d$, and $k \in \mathbb{N}^*$, we call $P^k(D)$ the set of order k polynomials on D, $\mathbf{P}^k(D) = (P^k(D))^d$, and we consider the space of the broken polynomials:

$$P^k_{disc}(\mathcal{T}_h) = \left\{q \in L^2(\Omega); \quad \forall \ell \in \mathcal{I}_K,\ q_{|K_\ell} \in P^k(K_\ell)\right\}, \quad \mathbf{P}^k_{disc}(\mathcal{T}_h) := (P^k_{disc}(\mathcal{T}_h))^d.$$

We let $P^0(\mathcal{T}_h)$ be the space of piecewise constant functions on $\mathcal{T}_h$.

$$\forall k \in \mathbb{N}, \quad Q_{k,h} := P^k(\mathcal{T}_h) \cap L^2_{zmv}(\Omega). \tag{3}$$

We will now describe three numerical schemes to solve (2) for which the components of the velocity is discretized with the first order nonconforming Crouzeix-Raviart finite element method [5, Sect. 5, Example 4]. For simplicity, we suppose now that $\mathbf{f} \in \mathbf{L}^2(\Omega)$.

Let us consider X_h (resp. $X_{0,h}$), the space of nonconforming approximation of $H^1(\Omega)$ (resp. $H^1_0(\Omega)$) of order 1:

$$X_h = \left\{v_h \in P^1_{disc}(\mathcal{T}_h)\,;\ \forall f \in \mathcal{I}_F^i,\ \int_{F_f} [v_h] = 0\right\}, \tag{4}$$

[2] The term facet stands for face (resp. edge) when $d = 3$ (resp. $d = 2$).

$$X_{0,h} = \left\{ v_h \in X_h \, ; \, \forall f \in \mathcal{I}_F^b, \; \int_{F_f} [v_h] = 0 \right\}. \tag{5}$$

Let us set $Q_h = Q_{0,h}$ and $\mathbf{X}_{0,h} = (X_{0,h})^d$. We now define the following bilinear forms:

$$a_{\nu,h} : \begin{cases} \mathbf{X}_{0,h} \times \mathbf{X}_{0,h} \to \mathbb{R} \\ \quad (\mathbf{u}'_h, \mathbf{v}_h) \mapsto \nu \, (\mathbf{u}'_h, \mathbf{v}_h)_h \end{cases} \text{and } b_h : \begin{cases} \mathbf{X}_{0,h} \times Q_h \to \mathbb{R} \\ \quad (\mathbf{v}_h, q_h) \mapsto -(\mathrm{div_h}\, \mathbf{v}_h, q_h) \end{cases}. \tag{6}$$

The discretization of variational formulation of problem (2) reads:

$$\text{Find } (\mathbf{u}_h, p_h) \in \mathbf{X}_{0,h} \times Q_h \mid \begin{cases} a_{\nu,h}(\mathbf{u}_h, \mathbf{v}_h)_h + b_h(\mathbf{v}_h, p) = (\mathbf{f}, \mathbf{v}_h)_{L^2(\Omega)} & \forall \mathbf{v}_h \in \mathbf{X}_{0,h}, \\ \qquad b_h(\mathbf{u}_h, q_h) = 0 & \forall q_h \in Q_h. \end{cases} \tag{7}$$

Suppose there exists $\phi \in H^1(\Omega) \cap L^2_{zmv}(\Omega)$ such that $\mathbf{f} = \mathbf{grad}\, \phi$. In that case, the solution to Problem (2) is $(\mathbf{u}, p) = (0, \phi)$. By integrating by parts, since $\mathbf{v}_h \notin \mathbf{H}^1(\Omega)$, we have:

$$\forall \mathbf{v}_h \in \mathbf{X}_{0,h}, \quad (\mathbf{f}, \mathbf{v}_h)_{L^2(\Omega)} = -(\mathrm{div_h}\, \mathbf{v}_h, \phi) + \sum_{f \in \mathcal{I}_F^i} \int_{F_f} [\mathbf{v}_h \cdot \mathbf{n}_f]\, \phi. \tag{8}$$

The jump term in (8), acts as a numerical source, which numerical influence is proportional to $1/\nu$. Hence, we cannot obtain exactly $\mathbf{u}_h = 0$, this velocity is called spurious velocity. There are different strategies to cure this well-known problem:

- Projecting the test-function on the right hand side of (8) on a discrete subspace of $\mathbf{H}(\mathrm{div}\, ; \Omega)$[3] [6]. This method makes the jump term in (8) vanish. The main advantage of this method is that it improves the velocity approximation when ν is small. But, when solving Navier-Stokes, one must modify the source term, the mass matrix and the convection term. While the implementation of the source term is straightforward, changing the mass matrix may slow down the pressure solver.
- Increasing the space of the discrete pressure [7, 8].

This last method consists of adding degrees of freedom for the pressure discretization on the vertices. The resulted discretization is called the $\mathbf{P}^1_{nc} - (P^0 + P^1)$ scheme. The space of discrete pressure is then defined as $\widetilde{Q}_h = Q_{0,h} \oplus Q_{1,h}$. It has been shown that the scheme is well posed and give a good approximation for the gradient of pressure, and it is the scheme currently used in TrioCFD code. Compared to $\mathbf{P}^1_{nc} - P^0$ scheme, the $\mathbf{P}^1_{nc} - (P^0 + P^1)$ scheme gives a better approximation of the velocity in the sense that the discrete mass conservation equation is strengthened. Indeed, for any $\widetilde{q}_h \in \widetilde{Q}_h$, we write: $\widetilde{q}_h = q_{0,h} + q_{1,h}$, where $q_{0,h} \in Q_{0,h}$ and $q_{1,h} \in Q_{1,h}$. Let us consider the bilinear form used to define the $\mathbf{P}^1_{nc} - (P^0 + P^1)$ scheme:

[3] $\mathbf{H}(\mathrm{div}\, ; \Omega) := \{\mathbf{v} \in \mathbf{L}^2(\Omega) / \mathrm{div}\, \mathbf{v} \in \mathbf{L}^2(\Omega)\}$.

$$\widetilde{b}_h : \begin{cases} \mathbf{X} \times \widetilde{Q}_h \to \mathbb{R} \\ (\mathbf{v}_h, \widetilde{q}_h) \mapsto -(\mathrm{div_h}\, \mathbf{v}_h, q_{0,h}) + (\mathbf{v}_h, \mathbf{grad}\, q_{1,h})_{\mathbf{L}^2(\Omega)} \end{cases} . \tag{9}$$

Then, one can show, for $d = 2$ that [8, Theorem 4.3.2]:

Property 1 Let $\mathbf{v}_h \in \mathbf{V}_h := \{\mathbf{w}_h \in \mathbf{X}_h \mid \ \forall q_h \in \widetilde{Q}_h, \quad \widetilde{b}_h(\mathbf{w}_h, q_h) = 0\}$. Then for $d = 2$, we have: for all $q_{2,h} \in Q_{2,h}$ (defined by (3)),

$$\widetilde{b}_h(\mathbf{v}_h, q_{2,h}) = (\mathbf{grad}\, q_{2,h}, \mathbf{v}_h)_{\mathbf{L}^2(\Omega)} = 0.$$

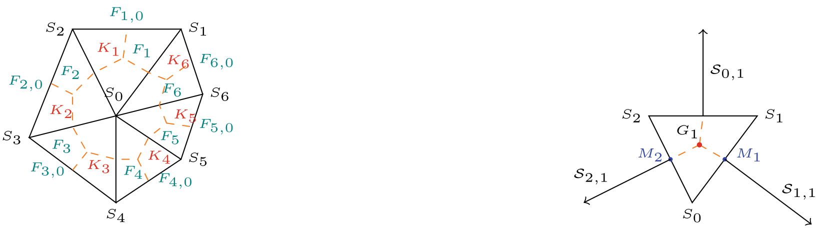
Even if $q_{2,h} \notin \widetilde{Q}_h$, we show that $\widetilde{b}_h(\mathbf{v}_h, q_{2,h}) = 0$. This case of “superconvergence” allows to obtain a better robustness with respect to the viscosity. This is illustrated with numerical experiments in the Sect. 4. The proof of Property 1 relies on a $2D$ quadrature formula which uses the degrees of freedom of the discrete pressure and cannot be extended in $3D$ with the same degrees of freedom. To recover Property 1 in $3D$, we must introduce P^2 discrete pressure degrees of freedom, located on the edges of the mesh. This increases the number of unknowns by the number of edges, which leads to an expensive linear system. Hence, we look for a numerical scheme which could be as precise in $3D$ than in $2D$, but at a lower cost. In the next Section, we propose a new strategy, which relies on the multi-points flux approximation to discretize the pressure gradient term in (2).

3 The $\mathbf{P}^1_{nc} - P^0_{Mps}$ Scheme

Here, we use the symmetric MPFA scheme [2] (where MPFA stands for *multi-points flux approximation*) to discretize the pressure gradient term in (2), in the case of a simplicial mesh. This scheme is part of the gradient scheme formalism and the resulting diffusion operator converge with the hypothesis given by [9, Theorem 12.5]. The discrete pressure space remains $Q_h = Q_{0,h}$. We call this new scheme the $\mathbf{P}^1_{nc} - P^0_{Mps}$ scheme. Let us consider the $2D$ case. We start by splitting the triangles into three quadrangles, connecting the barycentre of the triangle to the midpoint of each edges. Considering some $q_h \in Q_h$, we will calculate an affine approximation of q_h on each quadrangle. To do so, we need to add temporary auxiliary unknowns located at one third of the edges (see Fig. 2b) to approximate the gradients of the affine pressures in exact ways.

Let $j \in \mathcal{I}_S$. We denote $N_{K,j}$ the number of triangles with S_j as vertex and $N_{S,j}$ the number of neighbouring vertices. Notice that in $2D$, $N_{K,j} = N_{S,j}$. We define the macro-element $\mathcal{M}_j$ such that $\overline{\mathcal{M}_j} := \bigcup_{\ell \in \mathcal{I}_{K,j}} \overline{K_\ell}$. Let's renumber the vertices so that: $S_0 = S_j$, $\mathcal{I}_{S,0} = \{1, \cdots, N_{S,0}\}$ and for all $i \in \mathcal{I}_{S,0}$, $S_i S_{i+1} \subset \mathcal{F}_h$ (setting $S_{N_{S,0}+1} = S_1$). For $i \in \mathcal{I}_{S,0}$ we denote by:

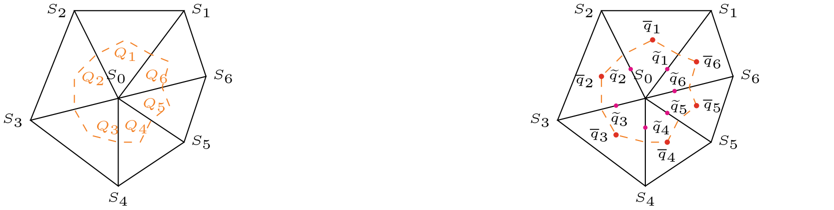
- K_i the triangle of vertices $S_0 S_i S_{i+1}$, and we call its barycentre G_i.
- F_i the edge such that $F_i = S_0 S_i$, and we call M_i its midpoint.



(a) Macro-element $\mathcal{M}_0 = S_1\, S_2\, S_3\, S_4\, S_5\, S_6$.

(b) Triangle $K_1 = S_0\, S_1\, S_2$.

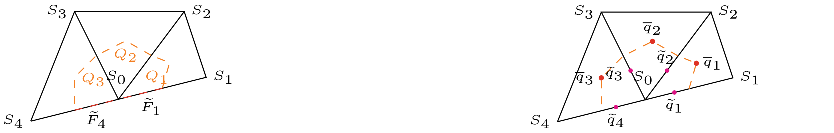
Fig. 1 Notations in case $N_{S,0} = 6$ and $j \in \mathcal{I}_S^i$



(a) Quadrangles $(Q_i)_{i=1}^{N_{S,0}}$.

(b) Discrete pressures $(\overline{q}_i, \widetilde{q}_i)_{i=1}^{N_{S,0}}$.

Fig. 2 MPFA Scheme for $j \in \mathcal{I}_S^i$ and $N_{S,0} = 6$



(a) Quadrangles $(Q_i)_{i=1}^{N_{S,0}}$.

(b) Discrete pressures $(\overline{q}_i, \widetilde{q}_i)_{i=1}^{N_{S,0}}$.

Fig. 3 MPFA Scheme for $j \in \mathcal{I}_S^b$ and $N_{S,0} = 4$

- $F_{i,0}$ the edge opposite to S_0 in K_i.
- $\widetilde{F}_i$ the half-edges defined by S_0 and the midpoint of F_i.
- Q_i the quadrangle of vertices $S_0\, M_i\, G_i\, M_{i+1}$ (Fig. 2a for $S_0 \subset \Omega$ and Fig. 3a for $S_0 \subset \Gamma$).

For $i,\ j \in \mathcal{I}_{S,0}$, we denote by $\mathcal{S}_{i,j}$ the normal vector outgoing of K_j at F_i and of norm $|F_i|$. For $i \in \mathcal{I}_{S,0}$, we call $\mathcal{S}_{0,i}$ the normal vector outgoing of K_i at $F_{i,0}$. On Fig. 1a, we represent $\mathcal{M}_0$ in case $S_0 \subset \Omega$ and $N_{S,0} = 6$. On Fig. 1b, we represent the triangle K_1 with the vectors $(\mathcal{S}_{j,1})_{j=0}^2$ and its barycentre G_1.

Let $q_h \in Q_h$. We set $q_{h|K_\ell} := \overline{q}_\ell$. Consider $S_0 \subset \Omega$ (Fig. 2). Let us build a piecewise affine approximation of q_h on each quadrangle $(Q_i)_{i=1}^{N_{S,0}}$ (see Fig. 2a). We call this approximation $\widetilde{q}_h$. We first introduce auxiliary discrete pressure values $(\widetilde{q}_i)_{i=1}^{N_{S,0}}$ on the thirds of the inner edges of $\mathcal{M}_0$ (see Fig. 2b). For all $j \in \mathcal{I}_{S,i}$, we define $\mathcal{G}_i(q_h) := \mathbf{grad}\, \widetilde{q}_{h|Q_i}$, using an integration by part as it is done in [2, Sect. 3]:

$$|Q_i|\mathcal{G}_i = \int_{Q_i} \mathcal{G}_i(q_h) = \int_{\partial Q_i} \widetilde{q}_h \mathbf{n}_\Gamma = \widetilde{q}_i \frac{\mathcal{S}_{i,i}}{d} + \widetilde{q}_{i+1} \frac{\mathcal{S}_{i+1,i}}{d} + \overline{q}_i (-\frac{\mathcal{S}_{i,i}}{d} - \frac{\mathcal{S}_{i+1,i}}{d}).$$

Hence, noticing that $|Q_i| = \frac{|T_i|}{d+1}$, we have:

$$\mathcal{G}_i(q_h) = \frac{1}{|Q_i|}\left((\widetilde{q}_i - \overline{q}_i)\,\frac{\mathcal{S}_{i,i}}{d} + (\widetilde{q}_{i+1} - \overline{q}_i)\,\frac{\mathcal{S}_{i+1,i}}{d} \right) = \frac{d+1}{d\,|T_i|}\left(\widetilde{q}_i\, \mathcal{S}_{i,i} + \widetilde{q}_{i+1}\, \mathcal{S}_{i+1,i} + \overline{q}_i\, \mathcal{S}_{0,i} \right). \quad (10)$$

In order to preserve the flux across the inner edges of $\mathcal{M}_0$, we write that:

$$\forall i \in \mathcal{I}_{S,0}, \quad \mathcal{G}_i(q_h) \cdot \mathcal{S}_{i+1,i} + \mathcal{G}_{i+1}(q_h) \cdot \mathcal{S}_{i+1,i+1} = 0. \quad (11)$$

These $N_{S,0}$ equations with $N_{S,0}$ unknowns (the auxiliary discrete pressure values $(\widetilde{q}_i)_{i=1}^{N_{S,0}}$) lead to a well posed linear system. Thus, we can evaluate the auxiliary discrete pressure values $(\widetilde{q}_i)_{i=1}^{N_{S,0}}$ with the data $(\overline{q}_i)_{i=1}^{N_{S,0}}$. Therefore, we can explicitly express the pressure gradients $(\mathcal{G}_i(q_h))_{i=1}^{N_{S,0}}$ (10).

Consider now $S_0 \subset \Gamma$ (see Fig. 3). According to [4, proof of Proposition IV.3.7], if $\mathbf{f} \in \mathbf{H}^1(\Omega)$, the solution $(\mathbf{u}, p)$ to Problem (2) is such that:

$$\mathbf{grad}\, p_{|\partial\Omega} \cdot \mathbf{n}_{|\partial\Omega} = \mathbf{f} \cdot \mathbf{n}_{|\partial\Omega} - \nu \Delta \mathbf{u} \cdot \mathbf{n}_{|\partial\Omega}, \quad (12)$$

where $\mathbf{n}_{|\Gamma}$ is the unit outward normal vector at Γ.

In our numerical experiments, we make explicit the auxiliary discrete pressure values located on Γ (i.e. $\widetilde{q}_1$ and $\widetilde{q}_4$ on Fig. 3-(b)) by imposing that for all $i \in \mathcal{I}_{S,0}$ such that $F_i \subset \Gamma$:

$$\int_{\widetilde{F}_i} \mathcal{G}_i(q_h) \cdot \mathbf{n}_{|\Gamma} = \int_{\widetilde{F}_i} \mathbf{f} \cdot \mathbf{n}_{|\Gamma}. \quad (13)$$

This approximation gives good numerical results, as will be shown later in the numerical section. Again, the auxiliary discrete pressure values solve a well posed linear system. They can be written with the data $(\overline{q}_i)_{i=1}^{N_{S,0}}$, and we can explicitly express $\mathcal{G}_i(q_h)$.

For $i \in \mathcal{I}_S$, we let $(Q_{i,j})_j \in \mathcal{I}_{S,i}$ be the set of quadrangles built around S_i, and we call $\mathcal{Q}_h$ the mesh of all the quadrangles $\mathcal{Q}_h := (\,(Q_{i,j})_{j \in \mathcal{I}_{S,i}})_{i \in \mathcal{I}_S}$. Let $q_h \in Q_h$. Let $i \in \mathcal{I}_S$. In the macro-element $\mathcal{M}_i$, we call $\mathcal{G}_{i,j}(q_h)$ the local reconstructed gradient of q_h. We now define the MPFA gradient reconstruction as the operator $\mathcal{G}_h$:

$$\mathcal{G}_h : \begin{cases} Q_h \to \mathbf{P}^0(\mathcal{Q}_h) \\ q_h \mapsto \mathcal{G}_h(q_h) \end{cases} \mid \ \forall i \in \mathcal{I}_S,\ \forall j \in \mathcal{I}_{S,i}, \quad \mathcal{G}_h(q_h)_{|Q_{i,j}} = \mathcal{G}_{i,j}(q_{h|\mathcal{M}_i}). \quad (14)$$

If the data $\mathbf{f}$ is of low regularity, one can enhance the space of discrete pressure, adding the auxiliary unknowns on the boundary as degrees of freedom.

Let $g_h(\cdot, \cdot)$ be the following bilinear form:

$$g_h : \begin{cases} \mathbf{X}_h \times Q_h \to \mathbb{R} \\ (\mathbf{v}_h, q_h) \mapsto (\mathcal{G}_h(q_h), \mathbf{v}_h)_{\mathbf{L}^2(\Omega)} \end{cases} . \tag{15}$$

The discretization of (2) using the MPFA scheme for the pressure gradient reads:

$$\text{Find } (\mathbf{u}, p) \in \mathbf{X}_{0,h} \times Q_h \mid \quad \begin{cases} a_{\nu,h}(\mathbf{u}_h, \mathbf{v}_h) + g_h(\mathbf{v}_h, p_h) = (\mathbf{f}, \mathbf{v}_h)_{\mathbf{L}^2(\Omega)} & \forall \mathbf{v}_h \in \mathbf{X}_h \\ b_h(\mathbf{u}_h, q_h) = 0 & \forall q_h \in Q_h \end{cases} , \tag{16}$$

where the bilinear forms $a_{\nu,h}(\cdot,\cdot)$ and $b_h(\cdot,\cdot)$ are defined by (6). Notice that the linear system related to variational formulation (16) is not symmetric.

4 Numerical Results on the Stokes Problem

In this Section, we give some $2D$ numerical results which compare the $\mathbf{P}^1_{nc} - P^0_{Mps}$ scheme to the $\mathbf{P}^1_{nc} - P^0$ and $\mathbf{P}^1_{nc} - (P^0 + P^1)$ schemes. More numerical results (including on the Navier-Stokes problem) of the $\mathbf{P}^1_{nc} - P^0_{Mps}$ are available in [10]. Consider Problem (2) with prescribed solution such that: $(\mathbf{u}, p) = (\mathbf{0}, \varphi)$. When φ is some affine function, then both $\mathbf{P}^1_{nc} - (P^0 + P^1)$ and $\mathbf{P}^1_{nc} - P^0_{Mps}$ schemes give exactly $\mathbf{u}_h = \mathbf{0}$. When φ is some quadratic function, then $\mathbf{P}^1_{nc} - (P^0 + P^1)$ scheme gives exactly $\mathbf{u}_h = \mathbf{0}$, as a consequence of Property 1.

We notice that, compared to the $\mathbf{P}^1_{nc} - P^0$ scheme, the spurious velocities are greatly reduced by $\mathbf{P}^1_{nc} - (P^0 + P^1)$ and $\mathbf{P}^1_{nc} - P^0_{Mps}$ schemes. These schemes mitigate the amplitude of spurious velocities and therefore provide a less viscosity-dependent error simulation. This is illustrated by the resolution of (2) with $(\mathbf{u}, p)$ defined by (17). The errors resulted for $h = 0.1$ and $h = 0.0125$ are given by Tables 1 and 2. In these tables, we see that the $\mathbf{P}^1_{nc} - P^0_{Mps}$ scheme gives intermediate results. Also, we notice that the spurious velocities errors become overriding when:

- $\nu \leq 10^0$ with $h = 0.1$ and $\nu \leq 10^0$ with $h = 0.0125$ for the $\mathbf{P}^1_{nc} - P^0$.
- $\nu \leq 10^{-2}$ with $h = 0.1$ and $\nu \leq 10^{-3}$ with $h = 0.0125$ for the $\mathbf{P}^1_{nc} - P^0_{Mps}$.
- $\nu \leq 10^{-3}$ with $h = 0.1$ and $\nu \leq 10^{-5}$ with $h = 0.0125$ for the $\mathbf{P}^1_{nc} - (P^0 + P^1)$.

The tipping viscosity point, where the spurious velocities errors become dominant, depends on the velocity error generated by the gradient approximation and therefore the mesh size. As these schemes converge with different orders when $\mathbf{u} = 0$, it can be seen that decreasing the mesh size reduces the viscosity at which this point is reached more or less depending on the order.

$$(\mathbf{u}, p) = \left(\begin{matrix} (\cos(2\pi x) - 1)\ \sin(2\pi y) \\ -(\cos(2\pi y) - 1)\ \sin(2\pi x) \end{matrix}, \sin(2\pi x)\sin(2\pi y) \right) \tag{17}$$

Table 1 Velocity and pressure errors for $(\mathbf{u}, p)$ in (17) for $h = 0.1$

ν	$\|\mathbf{u} - \mathbf{u}_h^{CR}\|_0$	$\|\mathbf{u} - \mathbf{u}_h^{\mathrm{Trio}}\|_0$	$\|\mathbf{u} - \mathbf{u}_h^{Mps}\|_0$	$\|p - p_h^{CR}\|_0$	$\|p - p_h^{\mathrm{Trio}}\|_0$	$\|p - p_h^{Mps}\|_0$
1.00×10^{-2}	2.47×10^{-1}	2.45×10^{-2}	3.23×10^{-2}	1.88×10^{-1}	2.73×10^{-2}	2.11×10^{-2}
1.00×10^{-3}	2.46×10^{0}	2.58×10^{-2}	1.33×10^{-1}	1.88×10^{-1}	2.48×10^{-2}	1.95×10^{-2}
1.00×10^{-4}	2.46×10^{1}	8.79×10^{-2}	1.30×10^{0}	1.88×10^{-1}	2.48×10^{-2}	1.95×10^{-2}
1.00×10^{-5}	2.46×10^{2}	8.46×10^{-1}	1.30×10^{1}	1.88×10^{-1}	2.48×10^{-2}	1.95×10^{-2}

Table 2 Velocity and pressure errors for $(\mathbf{u}, p)$ in (17) for $h = 0.0125$

ν	$\|\mathbf{u} - \mathbf{u}_h^{CR}\|_0$	$\|\mathbf{u} - \mathbf{u}_h^{\mathrm{Trio}}\|_0$	$\|\mathbf{u} - \mathbf{u}_h^{Mps}\|_0$	$\|p - p_h^{CR}\|_0$	$\|p - p_h^{\mathrm{Trio}}\|_0$	$\|p - p_h^{Mps}\|_0$
1.00×10^{-2}	3.41×10^{-3}	3.65×10^{-4}	4.33×10^{-4}	2.22×10^{-2}	1.49×10^{-3}	1.11×10^{-3}
1.00×10^{-3}	3.38×10^{-2}	3.65×10^{-4}	5.42×10^{-4}	2.22×10^{-2}	3.79×10^{-4}	2.98×10^{-4}
1.00×10^{-4}	3.38×10^{-1}	3.66×10^{-4}	3.23×10^{-3}	2.22×10^{-2}	3.50×10^{-4}	2.78×10^{-4}
1.00×10^{-5}	3.38×10^{0}	4.21×10^{-4}	3.19×10^{-2}	2.22×10^{-2}	3.50×10^{-4}	2.78×10^{-4}

References

1. Agelas, L., Masson, R.: Convergence of the finite volume MPFA O scheme for heterogeneous anisotropic diffusion problems on general meshes. Acad. Sci. Paris, Ser. I 346 (2008)
2. Le Potier, C.: A finite volume method for the approximation of highly anisotropic diffusion operators on unstructured meshes. In: FVCA IV, Marrakesh, Marocco (2005)
3. Angeli, P.-E., Puscas, M.-A., Fauchet, G., Cartalade, A.: FVCA8 benchmark for the Stokes and Navier-Stokes equations with the TrioCFD code. In: FVCA VIII - Methods and Theoretical Aspects, vol. 199. Springer Proceedings in Mathematics & Statistics, pp. 181–302 (2017)
4. Boyer, F., Fabrie, P.: Mathematical Tools for the Study of the Incompressible Navier-Stokes Equations and Related Models. Applied Mathematical Sciences, . Springer, New York (2012)
5. Crouzeix, M., Raviart, P.-A.: Conforming and nonconforming finite element methods for solving the stationary Stokes equations. RAIRO, Sér. Anal. Numer. **33** (1973)
6. 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**(1), 782–800 (2014)
7. Heib, S.: Nouvelles discrétisations non structurées pour des écoulements de fluides á incompressibilité renforcée". PhD thesis. Université Paris 6 (2003)
8. Fortin, T.: Une méthode éléments finis á décompositoin L2 d'ordre élevé motivée par la simulation d'écoulement diphasique bas Mach. Ph.D. thesis. Université Pierre et Marie Curie – Paris VI (2006)
9. Droniou, J., Eymard, R., Gallouet, T., Guichard, C., Herbin, R.: The multi-point flux approximation MPFA-O scheme. In: The Gradient Discretisation Method, pp. 343–351 (2018)
10. Chénier, E., Jamelot, E., Le Potier, C., Peitavy, A.: Improved Crouzeix-Raviart scheme for the Stokes and Navier-Stokes problem (2023). cea-04033455

Towards a Finite Volume Discretization of the Atmospheric Surface Layer Consistent with Physical Theory

Simon Clément, Florian Lemarié, and Eric Blayo

Abstract We study an atmospheric column and its discretization. Because of numerical considerations, the column must be divided into two parts: (1) a surface layer, excluded from the computational domain and parameterized, and (2) the rest of the column, which reacts more slowly to variations in surface conditions. A usual practice in atmospheric models is to parameterize the surface layer without excluding it from the computational domain, leading to possible consistency issues. We propose here to unify the two representations in a Finite Volume discretization. In order to do so, the reconstruction inside the first grid cell is performed using the particular functions involved in the parameterizations and not only with polynomials. Using a consistency criterion, surface layer management strategies are compared in different physical situations.

Keywords Finite volume · Monin-Obukhov theory · Surface flux scheme

1 Introduction

A common difficulty for atmospheric models is to represent the surface layer (SL), i.e. the area directly and almost instantaneously influenced by the presence of the ground or the ocean. The scales in the SL (approximately the first 10 m of the air column) are so small that the resolution needed for a numerical model to represent the phenomena correctly is out of reach. However, the Monin-Obukhov (MO) theory, which generalises the wall law to density-stratified fluids, provides under certain simple hypotheses (quasi-stationarity, horizontal homogeneity, etc.) an analytical formulation of the solution in the SL and the expression of the fluxes (heat, momentum) exchanged with the atmosphere above it. At the discrete level, the present treatment of this SL in numerical models is inconsistent: it is both treated like the rest of the atmosphere column by a standard numerical scheme (polynomial profile)

S. Clément (✉) · F. Lemarié · E. Blayo
University Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France
e-mail: simon.clement2@univ-grenoble-alpes.fr

E. Franck et al. (eds.), *Finite Volumes for Complex Applications X—Volume 1, Elliptic and Parabolic Problems*, Springer Proceedings in Mathematics & Statistics 432,
https://doi.org/10.1007/978-3-031-40864-9_20

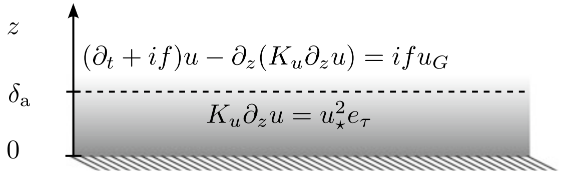

z
$(\partial_t + if)u - \partial_z(K_u \partial_z u) = ifu_G$
δ_a
$K_u \partial_z u = u_\star^2 e_\tau$
0

Fig. 1 Continuous equations in the computational domain (δ_a, $+\infty$) and constant flux in the SL (0, δ_a). The MO theory specifies the complex constant $u_\star e_\tau$ and a range of possible values for δ_a

when discretising the equations, and in a parameterised form (MO profile, which is a perturbation of a logarithmic profile, see e.g. [1]) for the calculation of fluxes. The consequences of this inconsistency are still poorly assessed in the context of atmospheric modeling, but we can mention for instance that in the context of combustion, it is mentioned in [2] that the way the wall law is implemented in a given code and the way it interacts with the numerical methods used (in particular the turbulence scheme) can influence the numerical results as much as a particular choice of wall law. In this paper, we will address this inconsistency and propose a new finite volume formulation to remedy it.

The turbulent Ekman layer model. Our approach is derived hereafter in the case of the 1D vertical Ekman layer model [3] in the neutral case. It includes the Coriolis effect with a constant parameter f and a vertical turbulent flux term $\langle w'u' \rangle$:

$$\partial_t u + ifu + \partial_z \langle w'u' \rangle = ifu_G \tag{1}$$

where i is the imaginary unit. The constant nudging term u_G pulls the solution towards the geostrophic equilibrium (a large-scale solution where the pressure gradient balances the Coriolis force). The horizontal wind u (in m.s^{-1}) is a complex variable accounting for both orientation and speed of the wind. The so-called Boussinesq hypothesis states that the turbulent flux is proportional to the gradient of u: $\langle w'u' \rangle = -K_u \partial_z u$ where $K_u \geq 0$ is the turbulent viscosity. In the SL, the MO theory states that this turbulent flux is constant along the vertical axis and provides its analytical expression. We thus obtain the system of equations given in Fig. 1.

Usual approach in atmospheric models. The usual choice made in current atmospheric models is to consider that the SL extends from the wall to the center of the first cell (let us note $z_{1/2}$ this altitude). In practice this means that compatibility constraints should apply at $z = z_{1/2}$ to correctly connect the profile as parameterized in the SL with the upper profile obtained using the numerical model. However the usual practice is to integrate the region $z \leq z_{1/2}$ corresponding to the SL into the computational domain and then use the MO parameterization only to predict a surface flux at $z = 0$. The impact of this approximation is poorly documented to date: (i) the extent of the SL is fixed for purely numerical reasons and not for physical reasons, which does not guarantee that the solution will converge with the resolution; (ii) the solution in the area $z \leq z_{1/2}$ is both parameterized and computed by the model, without ensuring the consistency of these two profiles. The coupling between the SL and the rest of the model is thus weak: the model provides the flow information at $z = z_{1/2}$ to the SL scheme and the latter provides in exchange a surface flux to the model at $z = 0$. In general no other interaction exists. For example, with this kind of coupling, the SL structures cannot really interact with the rest of the flow. For more details and numerous references, see [4].

A few studies address this issue: several alternatives for implementing a wall law in a Large Eddy Simulation solver are proposed in [2]; a first step toward a proper Finite Volume (FV) approach is proposed by [5], where the authors extend the SL to the entire first cell $(0, z_1)$ and design a scheme adapted to FV.

In this paper, we propose to implement directly in the FV discretization the existing assumptions underlying the MO theory. In order to do so, the reconstruction inside the first grid cell is performed using the analytical functions involved in the wall laws and not only with polynomials. This approach also allows to relax the artificial assumption $\delta_a = z_{1/2}$ and to extend the height of the SL beyond the first grid point if necessary (note that the log-layer mismatch, a well known numerical problem in Large Eddy Simulations, comes from a too thin SL). By being able to choose the thickness of the SL based on physical—and not only numerical—criteria, the consistency of the schemes is improved. The vertical resolution can thus be refined without changing the continuous equations solved by the discretization, thus answering the issues raised in [6]. Numerical experiments with a 1D Ekman layer model are performed, and SL management strategies are compared for different types of stratification.

2 The Finite Volume Scheme

Spline reconstruction of solutions. The space domain is divided into M cells delimited by heights $(z_0 = 0, \dots, z_m, \dots, z_M)$. The size of the mth cell is $h_{m+\frac{1}{2}} = z_{m+1} - z_m$ and the average of $u(z)$ over this cell is noted $\overline{u}_{m+\frac{1}{2}} = \frac{1}{h_{m+\frac{1}{2}}} \int_{z_m}^{z_{m+1}} u(z)dz$. The space derivative of u at z_m is noted ϕ_m. Figure 2 summarizes these notations. Averaging the evolution equation over a cell gives the semi-discrete equation

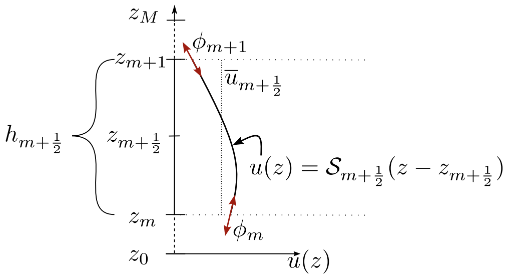

z_M
ϕ_{m+1}
z_{m+1}
$\overline{u}_{m+\frac{1}{2}}$
$h_{m+\frac{1}{2}}$
$z_{m+\frac{1}{2}}$
$u(z) = \mathcal{S}_{m+\frac{1}{2}}(z - z_{m+\frac{1}{2}})$
z_m
ϕ_m
z_0
$u(z)$

Fig. 2 Summary of the notations related to the discretisation

$$(\partial_t + if)\overline{u}_{m+\frac{1}{2}} - \frac{K_{u,m+1}\phi_{m+1} - K_{u,m}\phi_m}{h_{m+\frac{1}{2}}} = ifu_G \qquad (2)$$

The reconstruction of $u(z) = \mathcal{S}_{m+\frac{1}{2}}(z - z_{m+\frac{1}{2}})$ is chosen to be a quadratic polynomial (higher order schemes can be similarly derived, see [1]). The continuity of $u(z)$ and its space derivative ϕ between cells yields the relation:

$$\frac{h_{m-1/2}}{6}\phi_{m-1} + \frac{h_{m+1/2} + h_{m-1/2}}{3}\phi_m + \frac{h_{m+1/2}}{6}\phi_{m+1} = \bar{u}_{m+\frac{1}{2}} - \bar{u}_{m-\frac{1}{2}} \qquad (3)$$

which is a FV approximation used in fourth-order compact schemes for the first derivative $\partial_z u$ and second-order for $\partial_z^2 u$ (e.g. [7]).

Usual treatment of the SL with Finite Volume methods. The typical treatment of the SL in atmospheric models is to use the evolution equation in the first cell (z_0, z_1) to compute $\overline{u}_{\frac{1}{2}}$ and then assume that this averaged value is the wind speed at the center of the cell in the Monin-Obukhov theory applied with $\delta_a = z_1$. The corresponding bottom boundary condition is then

$$\underbrace{K_{u,0}\phi_0^{n+1}}_{\text{Surface flux}} = u_\star^2 e_\tau \quad \text{with } u_\star = \text{BULK}(\underbrace{\overline{u}_{\frac{1}{2}}^n}_{\text{Average around } z_{\frac{1}{2}}}) \qquad (4)$$

where BULK is a routine based on the Monin-Obukhov theory, $e_\tau = \frac{\overline{u}_{\frac{1}{2}}^{n+1}}{||\overline{u}_{\frac{1}{2}}^n||}$, and n denotes the time step. This method has several drawbacks:

- The value at the center of the cell is systematically larger than the average value because of the concavity of Monin-Obukhov profiles. This leads to a systematic underestimation of the surface flux by the SL scheme. A specific SL scheme was designed in [5] to prevent this bias.
- The evolution equation is not compatible with the constant flux hypothesis that defines the SL, as introduced in Sect. 1.
- δ_a, the height of the SL, is driven only by the space step and does not take into account any physical consideration.

On the incompatibility. According to the wall law, $K_{u,0}$ should be equal to the (very small) molecular viscosity $K_{mol} \approx 10^{-5}$ m^2.s^{-1}. However, the boundary condition $K_{u,0}\phi_0 = u_\star^2 e_\tau$ does not have the same influence depending on the numerical scheme used to discretize (2):

- **Finite Differences**: Injecting the boundary condition in the evolution equation at the first grid level gives

$$(\partial_t + if)u_{1/2} = \frac{1}{h_{1/2}}\left(K_{u,1}\frac{u_{3/2} - u_{1/2}}{h_1} - u_\star^2 e_\tau\right) \qquad (5)$$

where one can see that the value $K_{u,0}$ does not intervene in the equation.

- **Finite Volumes**: applying $(\partial_t + if)$ to (3) and using the polynomial reconstruction and the equations of Fig. 1, one can see that the FV scheme implicitly uses

$$(\partial_t + if)u(z_1) = \frac{K_{u,1}\phi_1 - u_\star^2 e_\tau}{h_{1/2}} + (\partial_t + if)\left(\frac{\phi_1}{3} + \frac{u_\star^2 e_\tau}{6K_{u,0}}\right)h_{1/2} \tag{6}$$

The (small) value of $K_{u,0}$ directly appears when we assume the parabolic profile inside the first grid cell. As a result, $u(z_1)$ scales with $\frac{1}{K_0}$ and exhibits unreasonable values. To obtain physically plausible profiles, one can replace $K_{u,0}$ by $K_{u,\delta}$: the wall law is then denied and $(\partial_z u)(z_0)$ is multiplied by $\frac{K_{mol}}{K_{u,\delta}}$. Note that this problem would not occur if the simple FV approximation $h_m \phi_m \approx \overline{u}_{m+\frac{1}{2}} - \overline{u}_{m-\frac{1}{2}}$ was used instead of (3).

Toward a Finite Volume scheme coherent with the physical theory. To address the drawbacks of the usual method presented above, we now construct a numerical boundary condition that is coherent with the continuous model with a free value of δ_a, named "FV free":

$$\underbrace{K_{u,\delta}\,\phi_\delta^{n+1}}_{\text{Flux at }\delta_\mathrm{a}} = u_\star^2\, e_\tau^{\text{free}} \quad \text{with } u_\star = \text{BULK}(\quad \underbrace{u^n(\delta_\mathrm{a})}_{\text{Reconstruction at }\delta_\mathrm{a}} \quad) \tag{7}$$

where $e_\tau^{\text{free}} = \frac{u^{n+1}(\delta_\mathrm{a})}{||u^n(\delta_\mathrm{a})||}$ is the orientation of $u(\delta_\mathrm{a})$ obtained with the spline reconstruction. For the sake of simplicity, we assume in the following that $\delta_\mathrm{a} < z_1$ (this hypothesis being easily relaxed by using the Monin-Obukhov profiles as the reconstruction in cells entirely contained in the SL). In the first grid cell, we assume that the constant flux hypothesis applies for $z < \delta_\mathrm{a}$ and we separate this cell into two parts: the surface layer $(0, \delta_\mathrm{a})$ and the "sub-cell" (δ_a, z_1). This split corresponds to the change of governing equations in Fig. 1. Let $\widetilde{h} = z_1 - \delta_\mathrm{a}$ be the size of the upper sub-cell (δ_a, z_1) and $\widetilde{u} = \frac{1}{\widetilde{h}} \int_{\delta_\mathrm{a}}^{z_1} u(z)dz$ be the corresponding averaged value of u. The following subgrid reconstruction is used:

$$u(z) = \begin{cases} \mathcal{S}_{1/2}\left(z - \dfrac{z_1 + \delta_\mathrm{a}}{2}\right), & z \geq \delta_\mathrm{a} \\ \displaystyle\int_0^z \frac{u_\star^2\, e_\tau^{\text{free}}}{K_{u,z'}}\,\mathrm{d}z', & z < \delta_\mathrm{a} \end{cases} \tag{8}$$

where a closed-form of the integral for $z < \delta_\mathrm{a}$ is given by MO theory. The quadratic spline $\mathcal{S}_{1/2}$ used for reconstruction is computed with the averaged value $\widetilde{u}$, the size of the sub-cell $\widetilde{h}$ and the fluxes at the extremities ϕ_δ and ϕ_1: its definition $\mathcal{S}_{1/2}(\xi) = \widetilde{u} + \frac{\phi_1 + \phi_\delta}{2}\xi + \frac{\phi_1 - \phi_\delta}{2\widetilde{h}}\left(\xi^2 - \frac{\widetilde{h}^2}{12}\right)$ is thus similar to the one in the other cells.

3 Numerical Experiments

The strategies to handle the SL are now compared through a test of consistency: for several strategies, the differences between a low-resolution and a high-resolution simulations are compared. The smaller the difference between the low-resolution and the high-resolution simulations, the better the consistency of the scheme. The proposed strategy “FV free” is compared with “FV1” (the typical current practice with Finite Volumes), “FV2” (an intermediate between “FV free” and “FV1”: similar to “FV free” but where the height of the surface layer δ_a is set to z_1) and “FD” (a Finite Difference reference).

- The turbulent viscosity is parameterized with a one-equation turbulence closure based on turbulent kinetic energy. The code is available at [8] and an in-depth description in [1]. An Euler implicit time scheme integrates the model over a full day of simulation.
- Parameters are $\Delta t = 30$ s, $u_G = 8$ m.s^{-1}, $f = 10^{-4}$ s^{-1}
- For “FV free” the same δ_a is used in both low- and high-resolution simulations, whereas the resolution imposes δ_a in the other configurations.
- The vertical levels of the low resolution simulation are taken as the 25 first of the 137-level configuration of the atmospheric model *Integrated Forecasting System* at ECMWF (European Centre for Medium-Range Weather Forecasts). The high-resolution simulation has 3 times more cells: two grid levels are added in each of the low-resolution grid cells. The usual SL strategies are designed for low-resolution configurations: the latter can hence be considered as reference solutions, compared through the sensitivity to the resolution.

Neutral case: In the neutral case (constant density), the difference between low and high resolution of the “FV free” scheme is small at low altitude (see Fig. 3). This is mainly due to two factors:

- $\delta_\text{a} = z_{\frac{1}{2}}^{\text{low-res}}$ is the same for both low and high resolutions, whereas for the other surface flux schemes the continuous equations change with δ_a.
- The initial relative difference for $u_\star$ (Fig. 3, left panel) is already much smaller than with the other schemes. This is a consequence of the imposed wall law: at initialization, there is already a logarithmic profile in the surface layer, instead of evolving toward a kind of compromise between the parameterized and the modeled values.

Stratified case: We now focus on a stratified model [3] that includes more of the physical behavior of atmospheric models: the turbulence closure depends on the density ρ such that $\partial_z \rho \propto -\partial_z \theta$ where θ is the potential temperature. We designed two cases:

1. A stable stratification, obtained with an initial temperature increasing with the altitude, and a surface temperature decreasing with time. The initial potential temperature is 265 K in the first 100 m of the atmosphere and then gains 1°C

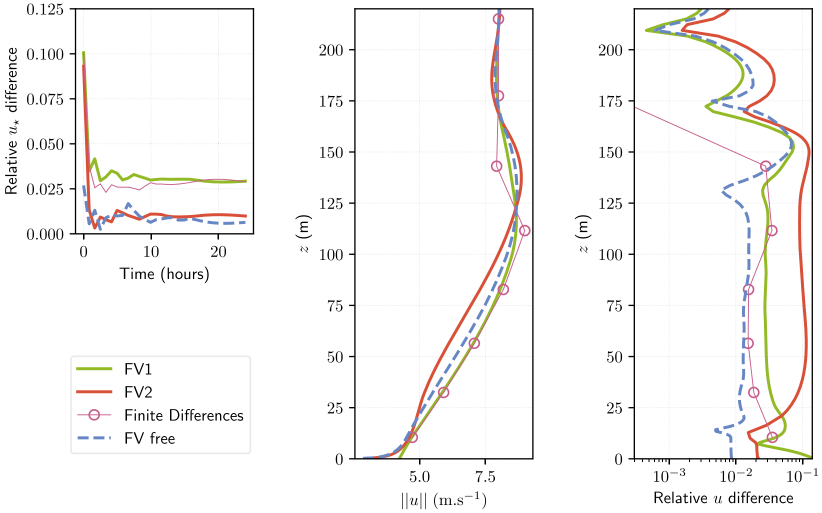


Fig. 3 Relative difference between low-resolution and high-resolution simulations with several strategies for the SL handling. Left: Relative difference in $u_\star$ as a function of time. Center: vertical profiles of the wind speed at the end of the simulation. Right: Relative difference of the wind speed between low- and high-resolution along the vertical (note the log scale)

every 100 m; the surface temperature starts at 265 K and loses 1°C every ten hours. The "low resolution" uses 15 grid points in the 400 m column and the "high resolution" uses 45 grid points.

2. An unstable stratification, obtained with a surface temperature following a daily oscillation between 279 and 281 K, and initial profiles of temperature and wind set to constant values of 280 K and 8 m $\cdot$ s^{-1} respectively. The "low resolution" is composed of 50 grid levels of 10 m each; 15 additional stretched levels between 500 and 1080 m make sure that the upper boundary condition is not involved in the results. The "high resolution" divides every space cell into 3 new space cells of equal sizes.

The differences between the two simulations are displayed in Fig. 4. In the stable case, the difference between the high resolution and the low resolution results does not significantly change with the surface flux schemes. The "FV2" scheme is not very consistent because it tries to follow the continuous model but with δ_a changing with the resolution. The Finite Difference or the "FV1" methods suffer less from this problem because, even if δ_a changes, it is assumed that the evolution equation is integrated inside the surface layer. In [9], authors also find that the sensitivity of their Large Eddy Simulation model to the grid spacing is "*more likely related to under-resolved near-surface gradients and turbulent mixing at the boundary-layer top, to the [sub-grid scale] model formulation, and/or to numerical issues, and not to deficiencies due to the use of improper surface boundary conditions*".

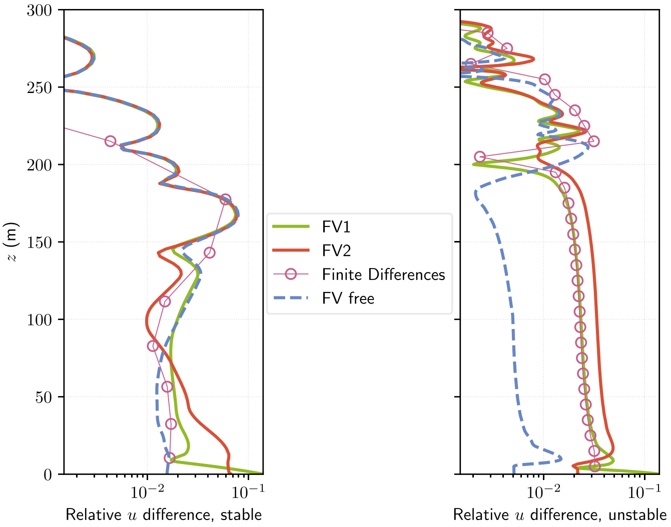


Fig. 4 Relative difference of the wind speed between low- and high-resolution simulations for several SL strategies. left: stable stratification. Right: unstable stratification

In the unstable case, the "FV free" scheme (with $\delta_{\mathrm{a}} = z_{\frac{1}{2}}^{\mathrm{low-res}}$) seems much more robust than the other schemes in the first 200 m (remind that the height of the SL is approximately 10 m). Above this height the differences between the high resolution and the low resolution simulations are not clearly influenced by the SL treatment. Note also that, as in the stable case, the "FV2" scheme is also less consistent: enforcing the MO theory in the first cell increases the sensitivity of the solution to δ_{a} because the SL is then tightly coupled with the computational domain. Finally, the "FV free" scheme combines good consistency properties with a SL scheme coherent with the physical theory.

References

1. Clement, S.: Numerical analysis for a combined space-time discretization of air-sea exchanges and their parameterizations. Ph.D. thesis, Université Grenoble Alpes (2022), (tel-04066324)
2. Jaegle, F., Cabrit, O., Mendez, S., Poinsot, T.: Implementation methods of wall functions in cell-vertex numerical solvers. Flow Turbul. Combust. **85**, 245–272 (2010). https://doi.org/10.1007/s10494-010-9276-1
3. McWilliams, J.C., Huckle, E., Shchepetkin, A.F.: Buoyancy effects in a stratified Ekman layer. J. Phys. Oceanogr. **39**, 2581–2599 (2009). https://doi.org/10.1175/2009JPO4130.1
4. Larsson, J., Kawai, S., Bodart, J., Bermejo-Moreno, I.: Large eddy simulation with modeled wall-stress: recent progress and future directions. Mech. Eng. Rev. **3** (2016). https://doi.org/10.1299/mer.15-00418
5. Nishizawa, S., Kitamura, Y.: A surface flux scheme based on the Monin-Obukhov similarity for finite volume models. J. Adv. Model. Earth Syst. **12**, 3159–3175 (2018). https://doi.org/10.1029/2018MS001534

6. Basu, S., Lacser, A.: A cautionary note on the use of Monin-Obukhov similarity theory in very high-resolution Large-Eddy Simulations. Bound.-Layer Meteorol. **163**, 351–355 (2017). https://doi.org/10.1007/s10546-016-0225-y
7. Piller, M., Stalio, E.: Finite-volume compact schemes on staggered grids. J. Comput. Phys. **197**(1), 299–340 (2004). https://doi.org/10.1016/j.jcp.2003.10.037
8. Clement, S.: Code for Ph.D. thesis. Zenodo (2022). https://doi.org/10.5281/zenodo.7092357
9. Maronga, B., Knigge, C., Raasch, S.: An improved surface boundary condition for large-eddy simulations based on Monin-Obukhov similarity theory: evaluation and consequences for grid convergence in neutral and stable conditions. Bound.-Layer Meteorol. **174**(2), 297–325 (2020). https://doi.org/10.1007/s10546-019-00485-w

Thermodynamically Consistent Discretisation of a Thermo-Hydro-Mechanical Model

Jérome Droniou, Mohamed Laaziri, and Roland Masson

Abstract We consider in this work a Thermo-Hydro-Mechanical (THM) model coupling the non-isothermal single phase flow in the porous rock and the linear thermo-poro-elasticity. This type of models plays an important role in several applications such as e.g. the hydraulic stimulation of deep geothermal systems, or the risk assessment of induced seismicity in CO2 storages. Compared with the isothermal case, the thermal coupling induces additional difficulties related in particular to the nonlinear convection term. Starting from the pioneer work of Coussy [2], we introduce a thermodynamically consistent discretisation of the THM coupled model which naturally leads to a discrete energy estimate. Our approach applies to a large class of Finite Volume schemes for the flow and energy equations but to fix ideas we consider the Hybrid Finite Volume (HFV) discretisation [3]. It is combined with a conforming Galerkin approximation of the mechanics. Our methodology accounts for a wide range of thermodynamical single phase fluid model and of thermo-poro-elastic parameters, as well as for diffusive or convective dominated energy transport. The efficiency of our approach is assessed on a 2D analytical test case using the HFV scheme for the non-isothermal flow and a $\mathbb{P}_2$ Finite Element method for the mechanics.

Keywords Thermo-poro-mechanics · Energy estimates · Thermodynamically consistent discretization · Finite volume

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

M. Laaziri · R. Masson
Université Côte d'Azur, Inria, CNRS, Laboratoire J.A. Dieudonné, Parc Valrose, 06108, Nice, France
e-mail: mohamed.laaziri@univ-cotedazur.fr

R. Masson
e-mail: roland.masson@univ-cotedazur.fr

E. Franck et al. (eds.), *Finite Volumes for Complex Applications X—Volume 1, Elliptic and Parabolic Problems*, Springer Proceedings in Mathematics & Statistics 432,
https://doi.org/10.1007/978-3-031-40864-9_21

1 Continuous Model

We consider a Thermo-Poro-Mechanical (THM) model under the hypothesis of small perturbations for the skeleton accounting for small transformations, displacement and variations of porosity [2]. Linear isotropic thermo-poro-elastic constitutive laws are considered for the skeleton assuming small variations of temperature around the reference temperature T_0. The Darcy law is used for the fluid velocity and the Fourier law for the thermal conduction. Thermal equilibrium is assumed between the fluid and the skeleton, and the fluid dissipation is neglected on an assumption of small Darcy velocities. The mechanical inertial term is modelled using the freezed specific average fluid—rock density m_0. Following [2], the resulting THM model is

$$\partial_t(\varrho\overline{\phi}) + \mathrm{div}(\varrho\overline{\mathbf{V}}) = h_m \quad \mathrm{in}(0,\tau)\times\Omega, \tag{1a}$$

$$\partial_t(\overline{S}_s + \varrho\overline{\phi}\overline{s}) + \mathrm{div}(\varrho\overline{s}\overline{\mathbf{V}}) + \frac{1}{T_0}\mathrm{div}\overline{\mathbf{q}} = \frac{h_e}{\overline{T}} \quad \text{in } (0,\tau)\times\Omega, \tag{1b}$$

$$m_0\partial_t^2\overline{\mathbf{u}} - \mathrm{div}(\sigma(\overline{\mathbf{u}},\overline{p},\overline{T})) = \mathbf{h} \quad \text{in } (0,\tau)\times\Omega, \tag{1c}$$

with

$$\overline{\mathbf{V}} = -\frac{\mathbb{K}}{\mu}\nabla\overline{p}, \quad \overline{\mathbf{q}} = -\lambda\nabla\overline{T}, \tag{1d}$$

$$\partial_t\overline{\phi} = b\partial_t(\mathrm{div}\overline{\mathbf{u}}) - 3\alpha_\phi\partial_t\overline{T} + \frac{1}{N}\partial_t\overline{p}, \tag{1e}$$

$$\partial_t\overline{S}_s = 3\alpha_s K_s\partial_t(\mathrm{div}\overline{\mathbf{u}}) - 3\alpha_\phi\partial_t\overline{p} + \frac{C_s}{T_0}\partial_t\overline{T}, \tag{1f}$$

$$\sigma(\overline{\mathbf{u}},\overline{p},\overline{T}) = \sigma^e(\overline{\mathbf{u}}) - b\overline{p}\,\mathbb{I} - 3\alpha_s K_s(\overline{T}-T_0)\mathbb{I}, \tag{1g}$$

$$\sigma^e(\overline{\mathbf{u}}) = \frac{E}{1+\nu}\left(\epsilon(\overline{\mathbf{u}}) + \frac{\nu}{1-2\nu}(\mathrm{div}\overline{\mathbf{u}})\mathbb{I}\right). \tag{1h}$$

The primary unknowns of the model (1) are the fluid pressure $\overline{p}$, the fluid temperature $\overline{T}$ and the skeleton displacement field $\overline{\mathbf{u}}$. They are solutions of the nonlinear system of PDEs coupling the fluid mass conservation equation (1a), the total entropy conservation equation (1b), and the skeleton momentum balance equation (1c). The closure laws (1d)–(1h) define the Darcy velocity $\overline{\mathbf{V}}$, the conductive heat flux $\overline{\mathbf{q}}$, and account for the linear thermo-poro-elastic constitutive laws defining the porosity $\overline{\phi}$, the volumetric skeleton entropy $\overline{S}_s$ and the total stress tensor σ (from the effective stress tensor σ^e). The parameters E and ν are the effective Young modulus and Poisson coefficient, N is the Biot modulus, b the Biot coefficient, $K_s = \frac{(1+(d-2)\nu)E}{d(1+\nu)(1-2\nu)}$ is the bulk modulus, d the space dimension, $3\alpha_s$ is the volumetric skeleton thermal dilation coefficient, $3\alpha_\phi$ is the volumetric thermal dilation coefficient related to the porosity, C_s is the skeleton volumetric heat capacity, and m_0 is the average fluid skeleton specific density considered freezed at its initial value.

To simplify the presentation, the fluid is assumed incompressible with a constant specific density $\varrho > 0$, a constant dynamic viscosity $\mu > 0$, and the gravity terms are not considered. The fluid specific entropy $\overline{s}$ and internal energy $\overline{e}$ depend only on the temperature $\overline{T}$ and are such that $d\overline{e} = \overline{T} d\overline{s}$. Note that the methodology presented below readily extends to the case with gravity terms, general fluid thermodynamics, and $\overline{p}$, $\overline{T}$ dependent viscosity.

To prepare the discretisation, we need to recast (1b). We have

$$\begin{aligned}\partial_t(\varrho\overline{\phi}\overline{s}) + \operatorname{div}(\varrho\overline{s}\overline{\mathbf{V}}) &= \varrho\overline{\phi}\partial_t\overline{s} + \varrho\overline{\mathbf{V}}\cdot\nabla\overline{s} + \overline{s}\,\underbrace{(\partial_t(\varrho\overline{\phi}) + \operatorname{div}(\varrho\overline{\mathbf{V}}))}_{=h_m \text{ by } (1)} \\ &= \frac{\varrho\overline{\phi}}{\overline{T}}\partial_t\overline{e} + \frac{1}{\overline{T}}\varrho\overline{\mathbf{V}}\cdot\nabla\overline{e} + \overline{s}h_m,\end{aligned}$$

where we have used the relation $d\overline{e} = \overline{T} d\overline{s}$ in the second line. This leads to replacing (1b) with

$$\partial_t\overline{S}_s + \frac{\varrho\overline{\phi}}{\overline{T}}\partial_t\overline{e} + \frac{1}{\overline{T}}\varrho\overline{\mathbf{V}}\cdot\nabla\overline{e} + \frac{1}{T_0}\operatorname{div}\overline{\mathbf{q}} = \frac{h_e}{\overline{T}} - \overline{s}h_m. \tag{2}$$

To keep the presentation simple, we consider no-flow, no-energy flux and no-displacement boundary conditions.

2 Discretisation

Let $\mathcal{M}$ denote the set of cells, and $\mathcal{F}$ the set of faces of the mesh, with internal faces gathered in $\mathcal{F}^{\text{int}}$ and boundary faces in $\mathcal{F}^{\text{ext}}$. The subset $\mathcal{F}_K \subset \mathcal{F}$ denotes the set of faces of the cell $K \in \mathcal{M}$, and we denote by $\sigma = K|L$ the face between two cells K, L; the notation $\sigma = K|\cdot$ is used for a face $\sigma \in \mathcal{F}_K \cap \mathcal{F}^{\text{ext}}$. For the pressure and temperature discretisation, we define the vector space of discrete unknowns

$$X_\mathcal{D} = \{v = ((v_K)_{K\in\mathcal{M}}, (v_\sigma)_{\sigma\in\mathcal{F}}) \,:\, v_K \in \mathbb{R} \text{ for all } K \in \mathcal{M}, v_\sigma \in \mathbb{R} \text{ for all } \sigma \in \mathcal{F}\}.$$

We let $\nabla_\mathcal{D} : X_\mathcal{D} \to L^\infty(\Omega)^d$ be the HFV gradient reconstruction operator, and the cellwise constant function reconstruction operator $\Pi_\mathcal{D} : X_\mathcal{D} \to L^2(\Omega)$ is such that for all $v \in X_\mathcal{D}$ and all $K \in \mathcal{M}$, $(\Pi_\mathcal{D} v)_{|K} = v_K$. For the displacement field discretisation, we denote by $\mathbf{U}_\mathcal{D}$ a finite-dimensional subspace of $H_0^1(\Omega)^d$.

The HFV Darcy and Fourier fluxes are denoted respectively by $V_{K,\sigma}$ and $G_{K,\sigma}$ defined from $X_\mathcal{D}$ to $\mathbb{R}$ such that, for all $v, w \in X_\mathcal{D}$, all $K \in \mathcal{M}$ and $\sigma \in \mathcal{F}_K$,

$$\int_K \frac{\mathbb{K}}{\mu}\nabla_\mathcal{D} v \cdot \nabla_\mathcal{D} w = \sum_{\sigma\in\mathcal{F}_K} V_{K,\sigma}(v)(w_K - w_\sigma),$$

and
$$\int_K \frac{\lambda}{T_0}\nabla_{\mathcal{D}} v \cdot \nabla_{\mathcal{D}} w = \sum_{\sigma\in\mathcal{F}_K} G_{K,\sigma}(v)(w_K - w_\sigma).$$

Let the index $\sigma,+$ denote either σ if no upwinding is used, or an upwind choice between K and L if $\sigma = K|L \in \mathcal{F}^{\rm int}$ and an upwind choice between K and σ if $\sigma = K|\cdot \in \mathcal{F}^{\rm ext}$. A key ingredient of the spatial discretisation that enables an energy estimate is the following discretisation of $\varrho\overline{\mathbf{V}}\cdot\nabla\overline{e} = {\rm div}(\varrho\overline{e}\overline{\mathbf{V}}) - \overline{e}\ {\rm div}(\varrho\overline{\mathbf{V}})$ on K, which includes a possible upwinding $e_{\sigma,+}$ of the discrete internal energy: for $p \in X_{\mathcal{D}}$ and $e \in X_{\mathcal{D}}$,

$$\sum_{\sigma\in\mathcal{F}_K} e_{\sigma,+}\varrho V_{K,\sigma}(p) - e_K \sum_{\sigma\in\mathcal{F}_K} \rho V_{K,\sigma}(p) = \sum_{\sigma\in\mathcal{F}_K} \rho V_{K,\sigma}(p)(e_{\sigma,+} - e_K).$$

We consider a time discretisation $(t^n)_{n=0,\ldots,N}$ of the time interval $(0,\tau)$ with $t^0 = 0$ and $t^N = \tau$, and denote by $\delta t^{(n+\frac{1}{2})} = t^{n+1} - t^n$ the time step n. If $f = (f^n)_{n=0,\ldots,N}$ is a family of functions, the discrete time derivative of f is defined as

$$\delta_t^{(n+\frac{1}{2})} f = \frac{f^{n+1} - f^n}{\delta t^{(n+\frac{1}{2})}}.$$

We also set $\dot{\mathbf{u}} = (\dot{\mathbf{u}}^n)_{n=0,\ldots,N}$ with $\dot{\mathbf{u}}^n = \frac{\mathbf{u}^n - \mathbf{u}^{n-1}}{t^n - t^{n-1}}$ and we write $\delta_t^{(n+\frac{1}{2})}\dot{\mathbf{u}} = 2\frac{\dot{\mathbf{u}}^{n+1} - \dot{\mathbf{u}}^n}{t^{n+1} - t^{n-1}}$. For given discrete pressures $p \in (X_{\mathcal{D}})^{N+1}$, temperatures $T \in (X_{\mathcal{D}})^{N+1}$ and displacement field $\mathbf{u} \in (\mathbf{U}_{\mathcal{D}})^{N+1}$, the discrete porosity $\phi = (\phi^n)_{n=0,\ldots,N}$ and skeleton entropy $S_s = (S_s^n)_{n=0,\ldots,N}$ are families of cellwise constant functions $\Omega \to \mathbb{R}$ on $\mathcal{M}$ such that ϕ^0, S_s^0 are given (e.g., as projections of the continuous initial porosity and entropy) and, for all $n = 0,\ldots,N-1$,

$$\begin{aligned}
\delta_t^{(n+\frac{1}{2})}\phi^n &= b\pi_{\mathcal{M}}(\delta_t^{(n+\frac{1}{2})}{\rm div}\mathbf{u}) - 3\alpha_\phi\delta_t^{(n+\frac{1}{2})}\Pi_{\mathcal{D}}T + \frac{1}{N}\delta_t^{(n+\frac{1}{2})}\Pi_{\mathcal{D}}p,\\
\delta_t^{(n+\frac{1}{2})}S_{s,\mathcal{D}} &= 3\alpha_s K_s\pi_{\mathcal{M}}(\delta_t^{(n+\frac{1}{2})}{\rm div}\mathbf{u}) - 3\alpha_\phi\delta_t^{(n+\frac{1}{2})}\Pi_{\mathcal{D}}p + \frac{C_s}{T_0}\delta_t^{(n+\frac{1}{2})}\Pi_{\mathcal{D}}T,
\end{aligned}$$

where $\pi_{\mathcal{M}}$ is the projection on piecewise constant functions on $\mathcal{M}$, that is, $(\pi_{\mathcal{M}}f)_{|K} = \frac{1}{|K|}\int_K f$ for all $K \in \mathcal{M}$.

We also define, for $\bullet = \{m, e\}$ and $n = 0,\ldots,N-1$, the function $\widehat{h}_\bullet^{n+1} : \Omega \to \mathbb{R}$ as the piecewise constant function on $\mathcal{M}$ equal on $K \in \mathcal{M}$ to the average $\widehat{h}_{\bullet,K}^{n+1}$ of $h_\bullet$ on $(t^n, t^{n+1}) \times K$. The function $\widehat{\mathbf{h}}^{n+1} : \Omega \to \mathbb{R}^d$ is defined in the same way from $\mathbf{h}$.

Setting $\xi_K^n = \overline{\xi}(T_K^n)$ and $\xi_\sigma^n = \overline{\xi}(T_\sigma^n)$ for $\xi \in \{e, s\}$, the time stepping is defined by the discrete system:

$$\varrho|K|\delta_t^{(n+\frac{1}{2})}\phi_K + \sum_{\sigma\in\mathcal{F}_K} \varrho V_{K,\sigma}(p^{n+1}) = |K|\widehat{h}_{m,K}^{n+1} \qquad \forall K \in \mathcal{M}, \tag{3a}$$

$$\begin{aligned} &V_{K,\sigma}(p^{n+1}) + V_{L,\sigma}(p^{n+1}) = 0 \qquad \forall \sigma = K|L \in \mathcal{F}^{\mathrm{int}}, \\ &V_{K,\sigma}(p^{n+1}) = 0 \qquad \forall \sigma \in \mathcal{F}^{\mathrm{ext}}. \end{aligned} \tag{3b}$$

$$\begin{aligned} &|K|\Big(\delta_t^{(n+\frac{1}{2})} S_{s,K} + \varrho \frac{\phi_K^n}{T_K^{n+1}} \delta_t^{(n+\frac{1}{2})} e_K\Big) + \frac{1}{T_K^{n+1}} \sum_{\sigma \in \mathcal{F}_K} \varrho V_{K,\sigma}(p^{n+1})(e_{\sigma,+}^{n+1} - e_K^{n+1}) \\ &+ \sum_{\sigma \in \mathcal{F}_K} G_{K,\sigma}(T^{n+1}) = |K| \left(\frac{\widehat{h}_{e,K}^{n+1}}{T_K^{n+1}} - \widehat{h}_{m,K}^{n+1} s_K^{n+1} \right) \qquad \forall K \in \mathcal{M}. \end{aligned} \tag{3c}$$

$$\begin{aligned} &G_{K,\sigma}(T^{n+1}) + G_{L,\sigma}(T^{n+1}) = 0 \qquad \forall \sigma = K|L \in \mathcal{F}^{\mathrm{int}}, \\ &G_{K,\sigma}(T^{n+1}) = 0 \qquad \forall \sigma \in \mathcal{F}^{\mathrm{ext}}. \end{aligned} \tag{3d}$$

$$\begin{aligned} &\int_\Omega m_0(\delta_t^{(n+\frac{1}{2})} \dot{\mathbf{u}}) \cdot \mathbf{v} + \int_\Omega \sigma^e(\mathbf{u}^{n+1}) : \epsilon(\mathbf{v}) \\ &- \int_\Omega \Big(b \Pi_{\mathcal{D}} p^{n+1} + 3\alpha_s K_s \Pi_{\mathcal{D}} T^{n+1}\Big) \mathrm{div}(\mathbf{v}) = \int_\Omega \widehat{\mathbf{h}}^{n+1} \cdot \mathbf{v} \qquad \forall \mathbf{v} \in \mathbf{U}_{\mathcal{D}}. \end{aligned} \tag{3e}$$

Here, (3a)–(3b) discretise the mass conservation (1a), (3c)–(3d) discretise the entropy equation (2), and the mechanical equation (1c) is discretised by (3e). We note that, combining the mass and energy equations, the discretization of the term $\partial_t(\varrho\overline{\phi}\overline{e}) + \mathrm{div}(\varrho\overline{e}\overline{\mathbf{V}})$ is conservative. Note also that the discretization $\delta_t^{(n+\frac{1}{2})}\dot{\mathbf{u}}$ defined above is a natural extension of the classical formula $\frac{\mathbf{u}^{n+1} - 2\mathbf{u}^n + \mathbf{u}^{n-1}}{(\delta t)^2}$ in the case of a constant time step δt.

3 Energy Estimate

Let us define the discrete Darcy and Fourier diffusive terms by

$$\mathfrak{D}(p^{n+1}) = \int_\Omega \frac{\mathbb{K}}{\mu} \nabla_{\mathcal{D}} p^{n+1} \cdot \nabla_{\mathcal{D}} p^{n+1}, \qquad \mathfrak{F}(T^{n+1}) = \int_\Omega \frac{\lambda}{T_0} |\nabla_{\mathcal{D}} T^{n+1}|^2,$$

and we assume that

$$M := \begin{bmatrix} \frac{1}{N} & -3\alpha_\phi \\ -3\alpha_\phi & \frac{C_s}{T_0} \end{bmatrix} \text{ is definite positive.}$$

The discrete energy is $\mathfrak{E} = (\mathfrak{E}^n)_{n=0,\ldots,N}$ with $\mathfrak{E}^n : \Omega \to \mathbb{R}$ given by

$$\mathfrak{E}^n = \frac{1}{2}\left[\Pi_{\mathcal{D}}p^n \; \Pi_{\mathcal{D}}T^n\right] M \begin{bmatrix}\Pi_{\mathcal{D}}p^n \\ \Pi_{\mathcal{D}}T^n\end{bmatrix} + \varrho\phi^n \Pi_{\mathcal{D}}e^n + \frac{1}{2}\frac{E}{1+\nu}\left(|\epsilon(\mathbf{u})|^2 + \frac{\nu}{1-2\nu}(\mathrm{div}\mathbf{u})^2\right).$$

Let us also set the discrete specific free enthalpy of the fluid as

$$g^{n+1} = \Pi_{\mathcal{D}}e^{n+1} + \frac{\Pi_{\mathcal{D}}p^{n+1}}{\varrho} - \Pi_{\mathcal{D}}T^{n+1}\Pi_{\mathcal{D}}s^{n+1}.$$

Then, any solution of the discrete system (3) satisfies the following discrete energy estimate: for all $n = 0, \ldots, N-1$,

$$\int_\Omega \frac{m_0}{2}\delta_t^{(n+\frac{1}{2})}|\dot{\mathbf{u}}|^2 + \int_\Omega \delta_t^{(n+\frac{1}{2})}\mathfrak{E} + \mathfrak{D}(p^{n+1}) + \mathfrak{F}(T^{n+1}) \le \int_\Omega \left(\widehat{h}_e^{n+1} + g^{n+1}\widehat{h}_m^{n+1}\right) + \int_\Omega \widehat{\mathbf{h}}^{n+1}\cdot\dot{\mathbf{u}}^{n+1}.$$

To deduce a control on the primary discrete unknowns, we make the following assumptions.

- Throughout the simulation, $\phi \ge \phi_* > 0$; the model itself does not contain any mechanism that ensures that the continuous porosity remains positive, so this assumption is mandatory (see the introduction of [1]), but can also easily be checked during the simulation.
- The energy and entropy laws satisfy: $\overline{e}(\overline{T}) \ge 0$ for all $\overline{T}$, and $\overline{e} - \overline{T}\overline{s}$ is sub-quadratic in the sense that $\lim_{|\overline{T}|\to\infty} \frac{\overline{e}(\overline{T})-\overline{T}\overline{s}(\overline{T})}{|\overline{T}|^2} = 0$.
- The mass, energy and momentum source terms h_m, h_e, $\mathbf{h}$ are bounded.
- The thermo-poro-elastic parameters satisfy $\frac{1}{N} > 0$, $C_s > 0$, $\alpha_\phi \ge 0$, $E > 0$, $\nu \in (0, \frac{1}{2})$.
- The specific average density satisfies $m_0 \ge m_* > 0$.

Then, using a discrete Gronwall lemma we can show that there exists C, depending only on the data, such that for all small enough maximum time step (such a condition is only needed if $\mathbf{h} \neq 0$), one has

$$\|\dot{\mathbf{u}}\|_{L^\infty(0,\tau;L^2(\Omega))} + \|\Pi_{\mathcal{D}}p\|_{L^\infty(0,\tau;L^2(\Omega))} + \|\Pi_{\mathcal{D}}T\|_{L^\infty(0,\tau;L^2(\Omega))} + \|\mathbf{u}\|_{L^\infty(0,\tau;H^1(\Omega))} + \left\|\nabla_{\mathcal{D}}p\right\|_{L^2(0,\tau;L^2(\Omega))} + \|\nabla_{\mathcal{D}}T\|_{L^2(0,\tau;L^2(\Omega))} + \|\Pi_{\mathcal{D}}e\|_{L^\infty(0,\tau;L^1(\Omega))} \le C.$$

4 Numerical Validation

We investigate in this section the numerical convergence of the scheme based on the following analytical solution

Table 1 Material properties

Symbol	Quantity	Value	Unit
E	Young modulus	2.5	Pa
ν	Poisson's coefficient	0.25	–
N	Biot's modulus	0.25	Pa^{-1}
b	Biot's coefficient	1.0	–
K	Bulk modulus	2.0	Pa
μ	Fluid viscosity	1.0	Pa s
ϕ^0	Initial porosity	4	–
λ	Effective thermal conductivity	0.1	$\text{W m}^{-1}\text{ K}^{-1}$
ϱ	The fluid specific density	1	Kg m^{-3}
$3\ \alpha_s$	The volumetric skeleton thermal dilation coefficient	1	K^{-1}
$3\ \alpha_\phi$	The volumetric thermal dilation coefficient related to the porosity	1	K^{-1}
T_0	Reference temperature	1	K
m_0	Average fluid skeleton specific density	0	Kg m^{-3}
C_s	The skeleton volumetric heat capacity	0.5	$\text{J m}^{-3}\text{ K}^{-1}$

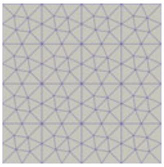


Fig. 1 Square domain Ω with its triangular mesh $i = 2$ using 56×4 cells

$$\mathbf{u}(\mathbf{x}, t) = 10^{-1}\mathrm{e}^{-t}\begin{pmatrix} x^2y^2 \\ -x^2y^2 \end{pmatrix},$$

$$p(\mathbf{x}, t) = \mathrm{e}^{-t}\sin(x)\sin(y), \quad T(\mathbf{x}, t) = \mathrm{e}^{-t}(2 - \sin(x)\sin(y)),$$

on the domain $\Omega = (0, 1)^2$ and time interval $(0, \tau)$ with $\tau = 1$. The fluid internal energy and entropy are defined by $\overline{e}(\overline{T}) = \overline{T}$ and $\overline{s}(\overline{T}) = \log(\frac{\overline{T}}{T_0})$. Dirichlet boundary conditions are imposed for p, T and $\mathbf{u}$ on $(0, \tau) \times \partial\Omega$ and the source terms h_m, h_e and $\mathbf{h}$ are computed based on the data set defined in Table 1. The domain Ω is discretized using the first family of triangular meshes from [4] as illustrated in Fig. 1. Each mesh indexed by $i \in \{1, 2, 3, 4\}$ includes $\#\mathcal{M} = 56 \times 4^{i-1}$ triangles. The HFV discretisation [3] of the flow and energy equations is combined with the $\mathbb{P}_2$ conforming Finite Element method for the mechanics to ensure the inf-sup condition and avoid potential oscillations of the pressure field at short times in the undrained regime. We consider a uniform time stepping of $(0, \tau)$ with time step $\Delta t = 10^{-4}$ chosen small enough to reduce the error due to the time discretization and focus on the convergence in space.

Table 2 Errors and convergence rates obtained with the centered scheme using $\mathbb{K} = \mathbb{I}$

Mesh	P		∇P		T		∇T		U		∇U	
	Error	Rate	Error	Rate	Error	Rate	Error	Rate	Error	Rate	Error	Rate
1	3.41E-03	–	7.40E-02	–	5.10E-04	–	7.40E-02	–	1.30E-02	–	5.17E-02	–
2	8.67E-04	1.97	3.68E-02	1.01	1.29E-04	1.98	3.68E-02	1.01	3.89E-03	1.74	2.41E-02	1.10
3	2.19E-04	1.99	1.84E-02	1.00	3.49E-05	1.89	1.84E-02	1.00	1.07E-03	1.86	1.15E-02	1.06
4	5.50E-05	1.99	9.17E-03	1.00	1.41E-05	1.31	9.18E-03	1.00	2.81E-04	1.93	5.62E-03	1.03

Table 3 Errors and convergence rates obtained with the upwind scheme using $\mathbb{K} = \mathbb{I}$

Mesh	P		∇P		T		∇T		U		∇U	
	Error	Rate	Error	Rate	Error	Rate	Error	Rate	Error	Rate	Error	Rate
1	3.44E-03	–	7.40E-02	–	9.84E-04	–	7.93E-02	–	1.33E-02	–	5.11E-02	–
2	8.79E-04	1.97	3.68E-02	1.01	3.18E-04	1.63	3.93E-02	1.01	3.93E-03	1.76	2.39E-02	1.09
3	2.26E-04	1.96	1.84E-02	1.00	1.28E-04	1.31	1.96E-02	1.00	1.07E-03	1.87	1.15E-02	1.06
4	5.96E-05	1.92	9.17E-03	1.00	6.25E-05	1.03	9.81E-03	1.00	2.81E-04	1.93	5.62E-03	1.03

The coupled nonlinear system is solved at each time step using a fixed-point method on cell pressures p and temperatures T accelerated by a Newton-Krylov algorithm [1]. At each iteration of the Newton-Krylov algorithm, the p, T sub-system is solved using a Newton-Raphson algorithm and the contact mechanics is solved using a semi-smooth Newton method.

The L^2 space time errors for p, T, $\mathbf{u}$ and their gradients are exhibited in Tables 2, 3 and 4 as functions of the mesh number i. Both the upwind and centered schemes are considered for the thermal convection as well as the two values of the permeability $\mathbb{K} = \mathbb{I}$ and $\mathbb{K} = 100\ \mathbb{I}$, respectively corresponding to equilibrated and convection dominated regimes. We first note that, due to the instability of the centered scheme in the convection dominated regime, this scheme fails to provide a solution for $\mathbb{K} = 100\,\mathbb{I}$ as a result of a failure of the nonlinear algorithm used to solve the scheme.

Regarding the displacement field, second and first order convergence rates are observed in all cases for respectively $\mathbf{u}$ and $\nabla\mathbf{u}$. This is in accordance with the cellwise constant reconstruction $\Pi_{\mathcal{D}}$ of the pressure and temperature in the displacement field variational formulation (3e). The convergence rates for p and ∇p are respectively roughly 2 and 1 in all cases as could be expected. On the other hand, the converge rates for T and ∇T depend on the approximation of the convection term and on the convection diffusion regime. For equilibrated convection and diffusion, the centered scheme provides a higher convergence rate for T (order 2, except on mesh 4 where

Table 4 Errors and convergence rates obtained with the upwind scheme using $\mathbb{K} = 100\,\mathbb{I}$

Mesh	P		∇P		T		∇T		U		∇U	
	Error	Rate	Error	Rate	Error	Rate	Error	Rate	Error	Rate	Error	Rate
1	3.49E-03	–	7.40E-02	–	3.15E-02	–	6.24E-01	–	1.94E-02	–	8.12E-02	–
2	9.08E-04	1.94	3.68E-02	1.01	1.45E-02	1.12	3.84E-01	0.70	5.18E-03	1.91	3.64E-02	1.16
3	2.47E-04	1.88	1.83E-02	1.00	6.66E-03	1.12	2.38E-01	0.69	1.30E-03	1.99	1.63E-02	1.16
4	7.45E-05	1.73	9.17E-03	1.00	2.97E-03	1.16	1.48E-01	0.69	3.14E-04	2.05	7.16E-03	1.19

the time error starts to dominate) than the upwind scheme (order between 1 and 2). An order 1 is observed on ∇T for both schemes. In the convection dominated regime, the upwind scheme exhibits a convergence rate slightly better than 1 for T and an order roughly equal to 0.7 for ∇T.

Acknowledgements The authors would like to thank BRGM and Andra for partially supporting this work and authorizing its publication.

References

1. Bonaldi, F., Brenner, K., Droniou, J., Masson, R., Pasteau, A., Trenty, L.: Gradient discretization of two-phase poro-mechanical models with discontinuous pressures at matrix fracture interfaces. ESAIM: Math. Model. Numer. Anal. **55**(5), 1741–1777 (2021)
2. Coussy, O.: Poromechanics. Wiley (2004)
3. 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**(02), 265–295 (2010)
4. Herbin, R., Hubert, F.: Benchmark on discretization schemes for anisotropic diffusion problems on general grids. In: The Proceedings of the Conference Finite Volumes for Complex Applications V, Jun 2008, pp. 659–692. France (2008)

Justification of Generalized Interface Conditions for Stokes–Darcy Problems

Elissa Eggenweiler, Joscha Nickl, and Iryna Rybak

Abstract For accurate modeling and numerical simulation of free-flow and porous-medium flow systems, the correct choice of coupling conditions at the common interface is essential. Most of the interface conditions available in the literature are limited to flows parallel or perpendicular to the porous layer. This significantly limits the number of applications that can be modeled in a physically meaningful way. Recently, generalized coupling conditions for arbitrary flow directions to the fluid–porous interface have been developed using homogenization and boundary layer theory. These conditions were validated numerically, however, their justification via error estimates was up to now an open question. In this work, we derive new interface conditions that extend the generalized coupling conditions by some higher-order boundary layer correctors. Under additional regularity and boundedness assumptions, we obtain error estimates that justify our newly developed coupling conditions.

Keywords Free flow · Porous medium · Interface conditions · Homogenization · Boundary layer theory

1 Introduction

Coupled systems of free flow and porous-medium flow play an important role in many fields of biological, environmental, and technical applications. Examples include blood flow through vessels and body tissues, surface water/groundwater flows, or water management in fuel cells. To investigate such coupled flow systems different spatial scales can be employed. At the pore scale, the detailed pore geometry is

E. Eggenweiler (✉) · I. Rybak
University of Stuttgart, Pfaffenwaldring 57, 70569 Stuttgart, Germany
e-mail: elissa.eggenweiler@ians.uni-stuttgart.de

I. Rybak
e-mail: iryna.rybak@ians.uni-stuttgart.de

J. Nickl
Aix-Marseille Université, 39 rue Frédéric Joliot-Curie, 13453 Marseille, France

E. Franck et al. (eds.), *Finite Volumes for Complex Applications X—Volume 1, Elliptic and Parabolic Problems*, Springer Proceedings in Mathematics & Statistics 432,
https://doi.org/10.1007/978-3-031-40864-9_22

resolved and the flow in the entire fluid domain is described by one system of partial differential equations. However, computation of the pore-scale flow field is often infeasible for practical applications, and thus macroscale models are preferred. At the macroscale, the free-flow and porous-medium regions are modeled as two different continua separated by a fluid–porous interface. Since the flow in coupled systems is highly influenced by complex interface-driven processes, the correct choice of interface conditions to couple the two flow models is crucial for physically consistent modeling and accurate numerical simulation.

Depending on the application of interest, there exists a variety of macroscale models describing fluid flows in coupled systems, e.g., [3, 4]. In this work, we are interested in steady-state, single-phase fluid flows at low Reynolds numbers where the porous medium is fully saturated with the fluid occupying the free-flow domain. From the macroscale perspective, such systems are typically described by the Stokes equations in the free-flow region, Darcy's law in the porous medium, and appropriate coupling conditions on the fluid–porous interface. Traditional coupling concepts [1, 8] for the Stokes–Darcy problem are based on the Beavers–Joseph condition which is valid only for fluid flows parallel to the porous medium. Hence, these traditional coupling conditions are not applicable to general filtration problems where the flow contacts the fluid–porous interface at an arbitrary angle [5].

Recently, alternative interface conditions have been developed in [6] accounting for arbitrary fluid flows in Stokes–Darcy systems. These conditions are derived using periodic homogenization and boundary layer theory and are confirmed, so far, only numerically. In this paper, we extend the work done in [6] and derive a set of generalized coupling conditions including additional higher-order terms in comparison to the ones in [6]. These terms lead to a more accurate description of coupled flow problems. In addition, we provide error estimates that justify the newly derived coupling conditions.

The paper is organized as follows. In Sect. 2, we formulate the problem setting and present the mathematical flow models including the interface conditions developed in this work. Section 3 is devoted to the derivation of the new interface conditions and error estimates. We discuss the presented results and give an outlook on future work in Sect. 4.

2 Problem Setting and Flow Models

In this section, we first introduce the geometrical setting and present the assumptions on the flow system. Then, we provide the microscopic and the macroscopic flow models. The latter includes the new interface conditions derived in Sect. 3.

At the macroscale, the coupled flow domain $\Omega \subset \mathbb{R}^2$ consists of the free-flow region $\Omega_{\text{ff}} = (0, L) \times (0, h)$ and the porous medium $\Omega_{\text{pm}} = (0, L) \times (0, -H)$, that are separated by a sharp fluid–porous interface Σ, i.e., $\Omega = \Omega_{\text{ff}} \cup \Sigma \cup \Omega_{\text{pm}}$ (Fig. 1, left). In this work, we consider the horizontal interface $\Sigma = (0, L) \times \{0\}$, where the unit normal vector on Σ is $\mathbf{n} = \mathbf{e}_2$ and the unit tangential vector is $\boldsymbol{\tau} = \mathbf{e}_1$. At

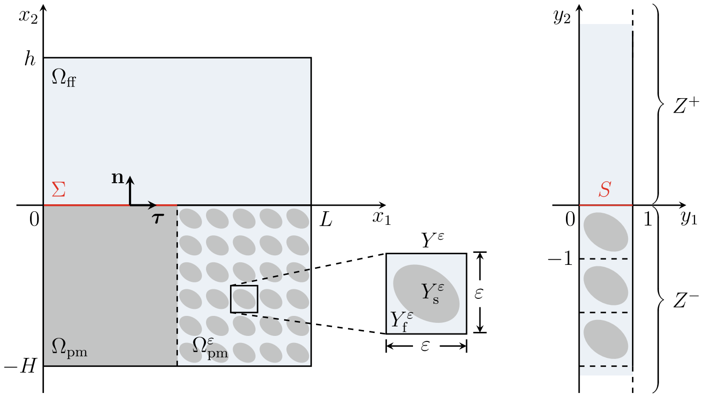


Fig. 1 Coupled flow domain at the macroscale (left part of domain) and at the pore scale (right part of domain) with a scaled unit cell $Y^\varepsilon = Y^\varepsilon_{\mathrm{f}} \cup Y^\varepsilon_{\mathrm{s}}$ (left). Boundary layer stripe $Z^{\mathrm{bl}} = Z^+ \cup S \cup Z^-$ (right)

the pore-scale, the entire flow domain $\Omega^\varepsilon \subset \mathbb{R}^2$ comprises the free-flow region Ω_{ff}, the interface Σ, and the pore space $\Omega^\varepsilon_{\mathrm{pm}} \subset \mathbb{R}^2$ of the porous medium, i.e., $\Omega^\varepsilon = \Omega_{\mathrm{ff}} \cup \Sigma \cup \Omega^\varepsilon_{\mathrm{pm}}$. We assume that the porous medium is constructed by the periodic repetition of the scaled unit cell $Y^\varepsilon = (0, \varepsilon) \times (0, \varepsilon)$ consisting of a fluid part $Y^\varepsilon_{\mathrm{f}}$ and a solid part $Y^\varepsilon_{\mathrm{s}}$ (Fig. 1, left). Here, ε denotes the characteristic pore size with $L/\varepsilon,\ H/\varepsilon \in \mathbb{N}$. For further details on the construction of the periodic porous medium, we refer to [6, 7].

We study steady-state, laminar ($Re \ll 1$), single-phase flows of an incompressible fluid that has constant viscosity. We assume that the fluid contains only one chemical species and fully saturates the pore space of the porous medium. The solid inclusions are supposed to be impermeable, rigid, and non-deformable. Moreover, the coupled system is considered to be isothermal.

2.1 Microscopic Flow Model

Under the prescribed assumptions, the fluid flow in the entire flow domain Ω^ε is described by the non-dimensional Stokes equations

$$\begin{aligned} -\Delta \mathbf{v}^\varepsilon + \nabla p^\varepsilon = \mathbf{0}\,, \quad \nabla\boldsymbol{\cdot}\mathbf{v}^\varepsilon = 0 \quad \text{in } \Omega^\varepsilon\,, \quad \int_{\Omega_{\mathrm{ff}}} p^\varepsilon \,\mathrm{d}\mathbf{x} = 0\,, \\ \mathbf{v}^\varepsilon = \mathbf{0} \quad \text{on } \partial\Omega^\varepsilon \setminus \partial\Omega\,, \quad \{\mathbf{v}^\varepsilon, p^\varepsilon\} \text{ is } L\text{-periodic in } x_1\,, \\ \mathbf{v}^\varepsilon = (v_1^{\mathrm{in}}(x_1), 0) \quad \text{on } (0,L)\times\{h\}\,, \quad v_2^\varepsilon = \frac{\partial v_1^\varepsilon}{\partial x_2} = 0 \quad \text{on } (0,L)\times\{-H\}\,. \end{aligned} \tag{1}$$

Here, $\mathbf{v}^\varepsilon = (v_1^\varepsilon, v_2^\varepsilon)$ and p^ε denote the fluid velocity and pressure, v_1^{in} is a prescribed inflow velocity. Problem (1) is also considered in [6], on which our work is based. It describes a coupled flow system where the flow direction is arbitrary to the porous region.

2.2 Macroscopic Flow Model

At the macroscale, the fluid flow in the free-flow region Ω_{ff} is modeled by the non-dimensional Stokes equations

$$\begin{gathered}
-\Delta \mathbf{v}^{\text{ff}} + \nabla p^{\text{ff}} = \mathbf{0}\,, \quad \nabla\cdot\mathbf{v}^{\text{ff}} = 0 \quad \text{in } \Omega_{\text{ff}}\,, \quad \textstyle\int_{\Omega_{\text{ff}}} p^{\text{ff}}\,\mathrm{d}\mathbf{x} = 0\,, \\
\{\mathbf{v}^{\text{ff}}, p^{\text{ff}}\} \text{ is } L\text{-periodic in } x_1\,, \quad \mathbf{v}^{\text{ff}} = (v_1^{\text{in}}(x_1), 0) \quad \text{on } (0,L)\times\{h\}\,,
\end{gathered} \tag{2}$$

and in the porous-medium domain Ω_{pm} by the non-dimensional Darcy equations

$$\begin{gathered}
\mathbf{v}^{\text{pm}} = -\mathbf{K}^{\varepsilon}\nabla p^{\text{pm}}\,, \quad \nabla\cdot\mathbf{v}^{\text{pm}} = 0 \quad \text{in } \Omega_{\text{pm}}\,, \\
p^{\text{pm}} \text{ is } L\text{-periodic in } x_1\,, \quad v_2^{\text{pm}} = 0 \quad \text{on } (0,L)\times\{-H\}\,.
\end{gathered} \tag{3}$$

Here, $\mathbf{v}^{\text{ff}}$, p^{ff}, $\mathbf{v}^{\text{pm}}$ and p^{pm} are the fluid velocity and pressure in the free-flow and porous-medium region, respectively, and $\mathbf{K}^{\varepsilon} = \varepsilon^2\mathbf{K}$ is the permeability tensor. The entries of $\mathbf{K}$ are defined in a standard way, e.g., [7, Eq. (1.17)].

In order to obtain a complete macroscale model formulation, i.e., to couple equations (2) and (3), conditions on the fluid–porous interface Σ need to be specified. In Sect. 3, we derive a new set of interface conditions for the Stokes–Darcy problem (2)–(3) that read

$$v_1^{\text{ff}} = -\varepsilon N_1^{\text{bl}} \frac{\partial v_1^{\text{ff}}}{\partial x_2}\bigg|_{\Sigma} + \varepsilon^2 \sum_{j=1}^{2} M_1^{j,\text{bl}} \frac{\partial p^{\text{pm}}}{\partial x_j}\bigg|_{\Sigma} - \varepsilon^2 \left(E_1^{\text{bl}} + L_1^{\text{bl}}\right) \frac{\partial}{\partial x_1}\frac{\partial v_1^{\text{ff}}}{\partial x_2}\bigg|_{\Sigma}\,, \tag{4}$$

$$v_2^{\text{ff}} = v_2^{\text{pm}} - \varepsilon^2 W^{\text{bl}} \frac{\partial}{\partial x_1}\frac{\partial v_1^{\text{ff}}}{\partial x_2}\bigg|_{\Sigma}\,, \tag{5}$$

$$\begin{aligned}
p^{\text{pm}} = p^{\text{ff}} - \frac{\partial v_2^{\text{ff}}}{\partial x_2}\bigg|_{\Sigma} + N_s^{\text{bl}} \frac{\partial v_1^{\text{ff}}}{\partial x_2}\bigg|_{\Sigma} - \varepsilon \sum_{j=1}^{2} M_{\omega}^{j,\text{bl}} \frac{\partial p^{\text{pm}}}{\partial x_j}\bigg|_{\Sigma} \\
+\varepsilon\left(L_{\eta}^{\text{bl}} + E_b^{\text{bl}} + N_1^{\text{bl}}\right) \frac{\partial}{\partial x_1}\frac{\partial v_1^{\text{ff}}}{\partial x_2}\bigg|_{\Sigma}\,.
\end{aligned} \tag{6}$$

Constants N_1^{bl}, $M_1^{j,\text{bl}}$, E_1^{bl}, L_1^{bl}, W^{bl}, N_s^{bl}, $M_{\omega}^{j,\text{bl}}$, L_{η}^{bl} and E_b^{bl} for $j = 1, 2$ appearing in conditions (4)–(6) are obtained from solutions to boundary layer problems defined on the boundary layer stripe $Z^{\text{bl}} = Z^+ \cup S \cup Z^-$ (Fig. 1, right). These constants are given by (14), (19), [6, Eqs. (3.15b), (3.15c), (3.26b), (3.26c)] and by

$$W^{\text{bl}} = -\int_{Z^-} t_1^{\text{bl}}\,\mathrm{d}\mathbf{y}\,, \tag{7}$$

where t_1^{bl} is the solution to [6, Eq. (3.14)]. All constants can be computed numerically based on the pore geometry near the fluid–porous interface. Due to lack of space, we do not provide the values of the constants for specific pore geometries in this paper (for N_1^{bl}, $M_1^{j,\mathrm{bl}}$, N_s^{bl} and $M_\omega^{j,\mathrm{bl}}$, see [6]).

Remark 1 Coupling conditions (4)–(6) have the same form as the ones proposed in [10], however, the constants appearing in these conditions are not exactly the same. A thorough comparison of interface conditions (4)–(6) with the conditions from [10] will be presented at the conference.

Remark 2 The set of interface conditions (4)–(6) extends the conditions derived in [6] by additional higher-order terms which take into account the variation of shear stress along the interface.

3 Derivation of Interface Conditions

In this section, we present the main steps for the derivation of the higher-order generalized interface conditions (4)–(6), and provide error estimates for the model approximation. Since our work is an extension of [6], we follow the procedure proposed in [6, Sect. 3.2] and use the same notations introduced therein. The goal is to construct an accurate approximation $\{\mathbf{v}^\varepsilon_{\mathrm{approx}}, p^\varepsilon_{\mathrm{approx}}\}$ of the pore-scale solution $\{\mathbf{v}^\varepsilon, p^\varepsilon\}$ such that the error of the model approximation $\{\mathbf{v}^\varepsilon - \mathbf{v}^\varepsilon_{\mathrm{approx}}, p^\varepsilon - p^\varepsilon_{\mathrm{approx}}\}$ is sufficiently small. Then, interface conditions can be formulated based on the derived model approximation.

In order to find an approximation of the pore-scale solution we apply homogenization and boundary layer theory. We expand the pore-scale velocity and pressure as follows

$$\begin{aligned}\mathbf{v}^\varepsilon(\mathbf{x}) &= \mathbf{v}_0(\mathbf{x},\mathbf{y}) + \varepsilon\mathbf{v}_1(\mathbf{x},\mathbf{y}) + \varepsilon^2\mathbf{v}_2(\mathbf{x},\mathbf{y}) + \mathcal{O}_{L^2(\Omega^\varepsilon)^2}(\varepsilon^3)\\ p^\varepsilon(\mathbf{x}) &= p_0(\mathbf{x},\mathbf{y}) + \varepsilon p_1(\mathbf{x},\mathbf{y}) + \mathcal{O}_{L^2(\Omega^\varepsilon)}(\varepsilon^{3/2}),\end{aligned} \tag{8}$$

where $\mathbf{y} = \mathbf{x}/\varepsilon$ and $\mathbf{v}_i$, p_i are 1-periodic functions in y_1 for $i \in \mathbb{N}_0$. The final approximation $\{\mathbf{v}^\varepsilon_{\mathrm{approx}}, p^\varepsilon_{\mathrm{approx}}\}$ is constructed based on (8) during several steps as in [6]. We tag the first approximation of the pore-scale solution by the superscript 0, i.e., $\{\mathbf{v}^{0,\varepsilon}_{\mathrm{approx}}, p^{0,\varepsilon}_{\mathrm{approx}}\}$. This approximation is then improved by adding boundary layer correctors. The resulting new approximation is indicated by a rising superscript, i.e., $\{\mathbf{v}^{1,\varepsilon}_{\mathrm{approx}}, p^{1,\varepsilon}_{\mathrm{approx}}\}$. In an analogous way $\{\mathbf{v}^{n,\varepsilon}_{\mathrm{approx}}, p^{n,\varepsilon}_{\mathrm{approx}}\}$ is constructed for the index $n \in \mathbb{N}$. We introduce the error functions $\mathbf{U}^{n,\varepsilon} = \mathbf{v}^\varepsilon - \mathbf{v}^{n,\varepsilon}_{\mathrm{approx}}$ and $P^{n,\varepsilon} = p^\varepsilon - p^{n,\varepsilon}_{\mathrm{approx}}$ corresponding to the approximation $\{\mathbf{v}^{n,\varepsilon}_{\mathrm{approx}}, p^{n,\varepsilon}_{\mathrm{approx}}\}$. The error functions according to the final model approximation $\{\mathbf{v}^\varepsilon_{\mathrm{approx}}, p^\varepsilon_{\mathrm{approx}}\}$ are denoted by $\mathbf{U}^\varepsilon$ and P^ε.

3.1 Model Approximation

In this section, we construct an approximation of the pore-scale solution $\{\mathbf{v}^\varepsilon, p^\varepsilon\}$ based on the work in [6]. We explain why additional terms of order ε are needed in the approximation and we provide the corresponding boundary layer problems.

We found out that for the pore-scale model approximation $\{\mathbf{v}^{6,\varepsilon}_{\text{approx}}, p^{6,\varepsilon}_{\text{approx}}\}$ derived in [6, Sect. 3.2.6] the error estimates that can be obtained are not sufficient. Since $\mathbf{v}^{6,\varepsilon}_{\text{approx}}$ is of order ε^2, we expect that for the velocity error it holds $\|\mathbf{U}^{6,\varepsilon}\|_{L^2(\Omega^\varepsilon)^2} \le C\varepsilon^i$ with $i > 2$. However, we realized that such an estimate is not possible when the approximation $\{\mathbf{v}^{6,\varepsilon}_{\text{approx}}, p^{6,\varepsilon}_{\text{approx}}\}$ is used. Thus, we need to improve this approximation. Careful examination of [2, 6] leads us to the following conclusion. For better error estimates we need to improve the result from [6, Corollary 3.5]: instead of the factors $\varepsilon^{3/2}$ and $\varepsilon^{5/2}$ we need $\varepsilon^{5/2}$ and $\varepsilon^{7/2}$, respectively.

At this stage, we remark that the solution $\{\mathbf{v}^{\text{cf}}, p^{\text{cf}}\}$ to [6, Eqs. (3.44), (3.45)] should not be used for the approximation $\{\mathbf{v}^{6,\varepsilon}_{\text{approx}}, p^{6,\varepsilon}_{\text{approx}}\}$ since this leads to an undesirable contribution of the velocity error function $\mathbf{U}^{6,\varepsilon}$ on the top boundary. Thus, instead of $\{\mathbf{v}^{6,\varepsilon}_{\text{approx}}, p^{6,\varepsilon}_{\text{approx}}\}$, we use

$$\mathbf{v}^{6,\varepsilon,\text{mod}}_{\text{approx}} = \mathbf{v}^{6,\varepsilon}_{\text{approx}} + \varepsilon^2\mathcal{H}(x_2)\mathbf{v}^{\text{cf}}, \quad p^{6,\varepsilon,\text{mod}}_{\text{approx}} = p^{6,\varepsilon}_{\text{approx}} + \varepsilon^2\mathcal{H}(x_2)p^{\text{cf}}, \tag{9}$$

as a basis for our work in this paper.

Improvement of Approximation from [6]. To find out how approximation (9) can be improved we study the corresponding weak formulation given by [6, Eq. (3.53)]. We identify the integral terms that are sources for a low estimation order, i.e., terms that are bounded by $C\varepsilon^{3/2}$ or $C\varepsilon^2$ under the assumptions in Remark 3. These terms are given in [6, Eqs. (3.11), (3.20), (3.21), (3.24), estimates on p. 743]. To improve the pore-scale approximation (9) we add boundary layer correctors such that in the weak formulation according to the new approximation, the integral terms of low order are eliminated. In this way, we obtain sufficiently good estimates for the errors $\mathbf{U}^\varepsilon$ and P^ε (Theorem 1).

In the following, we provide details on the correction of approximation (9) w.r.t. the integral terms in [6, Eq. (3.20), (3.21), (3.24), estimates on p. 743] since the other corrections are of order ε^3 and do not appear in the new interface conditions (4)–(6). To eliminate the integral term given in [6, Eq. (3.21), (3.24), estimates on p. 743], we define two boundary layer problems. The first one reads

$$\Delta_{\mathbf{y}}\mathbf{c}^{\text{bl}} - \nabla_{\mathbf{y}}b^{\text{bl}} = \Delta_{\mathbf{y}}\boldsymbol{\zeta}^{\text{bl}} \qquad \text{in } Z^+ \cup Z^-, \tag{10}$$

$$\nabla_{\mathbf{y}}\cdot\mathbf{c}^{\text{bl}} = 0 \qquad \text{in } Z^+ \cup Z^-, \tag{11}$$

$$[\![\mathbf{c}^{\text{bl}}]\!]_S = [\![\nabla_{\mathbf{y}}\mathbf{c}^{\text{bl}} - b^{\text{bl}}\mathbf{I}]\!]_S\mathbf{e}_2 = \mathbf{0} \qquad \text{on } S, \tag{12}$$

$$\mathbf{c}^{\text{bl}} = \mathbf{0} \quad \text{on} \quad \cup_{k=1}^{\infty}(\partial Y_{\text{s}} - (0,k)), \quad \{\mathbf{c}^{\text{bl}}, b^{\text{bl}}\} \text{ is 1-periodic in } y_1\,. \tag{13}$$

Since there exists $\gamma \in (0,1)$ such that $e^{\gamma|y_2|}\Delta_{\mathbf{y}}\boldsymbol{\zeta}^{\text{bl}} \in L^2(Z^{\text{bl}})^2$, where $\boldsymbol{\zeta}^{\text{bl}}$ is the solution to [6, Eq. (3.44)], system (10)–(13) is a boundary layer problem after [7].

Thus, we know that there exists a unique solution to (10)–(13) and that the velocity $\mathbf{c}^{\mathrm{bl}}$ and the pressure b^{bl} stabilize exponentially (in the sense of [7, Eqs. (3.33), (3.34), (3.38), (3.39)]) to constants in the free-flow region

$$\mathbf{E}^{\mathrm{bl}} = \left(\int_0^1 c_1^{\mathrm{bl}}\,\mathrm{d}y_1, 0\right), \qquad E_b^{\mathrm{bl}} = \int_0^1 b^{\mathrm{bl}}\,\mathrm{d}y_1\,, \tag{14}$$

and to zero in the porous medium. We extend the velocity $\mathbf{c}^{\mathrm{bl}}$ to zero in $\Omega \setminus \Omega^\varepsilon$, and set $\mathbf{c}^{\mathrm{bl},\varepsilon}(\mathbf{x}) = \mathbf{c}^{\mathrm{bl}}(\mathbf{x}/\varepsilon)$ and $b^{\mathrm{bl},\varepsilon}(\mathbf{x}) = b^{\mathrm{bl}}(\mathbf{x}/\varepsilon)$.

The second boundary layer problem is given by

$$\Delta_{\mathbf{y}}\boldsymbol{\xi}^{\mathrm{bl}} - \nabla_{\mathbf{y}}\eta^{\mathrm{bl}} = -\left(2\frac{\partial \mathbf{t}^{\mathrm{bl}}}{\partial y_1} - s^{\mathrm{bl}}\mathbf{e}_1 + \mathcal{H}(y_2)N_s^{\mathrm{bl}}\mathbf{e}_1\right) \qquad \text{in } Z^+ \cup Z^-\,, \tag{15}$$

$$\nabla_{\mathbf{y}}\cdot\boldsymbol{\xi}^{\mathrm{bl}} = 0 \qquad \text{in } Z^+ \cup Z^-\,, \tag{16}$$

$$[\![\boldsymbol{\xi}^{\mathrm{bl}}]\!]_S = [\![\nabla_{\mathbf{y}}\boldsymbol{\xi}^{\mathrm{bl}} - \eta^{\mathrm{bl}}\mathbf{I}]\!]_S\mathbf{e}_2 = \mathbf{0} \qquad \text{on } S\,, \tag{17}$$

$$\boldsymbol{\xi}^{\mathrm{bl}} = \mathbf{0} \text{ on } \cup_{k=1}^{\infty}(\partial Y_{\mathrm{s}} - (0,k)), \quad \{\boldsymbol{\xi}^{\mathrm{bl}}, \eta^{\mathrm{bl}}\} \text{ is 1-periodic in } y_1\,. \tag{18}$$

We know that there exists $\gamma \in (0,1)$ such that $e^{\gamma|y_2|}\nabla_{\mathbf{y}}\mathbf{t}^{\mathrm{bl}} \in L^2(Z^{\mathrm{bl}})^2$ and $e^{\gamma|y_2|}\left(s^{\mathrm{bl}} - \mathcal{H}(y_2)N_s^{\mathrm{bl}}\right) \in L^2(Z^{\mathrm{bl}})$. Thus, problem (15)–(18) fits in the form of the AUX problem from [7]. Therefore, we know that there exists a solution to (15)–(18) which is unique. Furthermore, we know that the boundary layer velocity $\boldsymbol{\xi}^{\mathrm{bl}}$ and pressure η^{bl} stabilize exponentially to boundary layer constants in the free-flow region and to zero in the porous medium. These boundary layer constants are given by

$$\mathbf{L}^{\mathrm{bl}} = \left(\int_0^1 \xi_1^{\mathrm{bl}}\,\mathrm{d}y_1, 0\right), \qquad L_\eta^{\mathrm{bl}} = \int_0^1 \eta^{\mathrm{bl}}\,\mathrm{d}y_1\,. \tag{19}$$

The boundary layer velocity $\boldsymbol{\xi}^{\mathrm{bl}}$ is extended to zero in $\Omega \setminus \Omega^\varepsilon$ and we define $\boldsymbol{\xi}^{\mathrm{bl},\varepsilon}(\mathbf{x}) = \boldsymbol{\xi}^{\mathrm{bl}}(\mathbf{x}/\varepsilon)$ and $\eta^{\mathrm{bl},\varepsilon}(\mathbf{x}) = \eta^{\mathrm{bl}}(\mathbf{x}/\varepsilon)$.

The improved approximations of the pore-scale velocity and pressure are

$$\mathbf{v}_{\mathrm{approx}}^{7,\varepsilon} = \mathbf{v}_{\mathrm{approx}}^{6,\varepsilon,\mathrm{mod}} - \varepsilon^2\left(\mathbf{c}^{\mathrm{bl},\varepsilon} - \mathcal{H}(x_2)\mathbf{E}^{\mathrm{bl}} + \boldsymbol{\xi}^{\mathrm{bl},\varepsilon} - \mathcal{H}(x_2)\mathbf{L}^{\mathrm{bl}}\right)\frac{\partial}{\partial x_1}\frac{\partial v_1^{\mathrm{ff}}}{\partial x_2}\bigg|_\Sigma\,,$$

$$p_{\mathrm{approx}}^{7,\varepsilon} = p_{\mathrm{approx}}^{6,\varepsilon,\mathrm{mod}} - \varepsilon\left(b^{\mathrm{bl},\varepsilon} - \mathcal{H}(x_2)E_b^{\mathrm{bl}} + \eta^{\mathrm{bl},\varepsilon} - \mathcal{H}(x_2)L_\eta^{\mathrm{bl}}\right)\frac{\partial}{\partial x_1}\frac{\partial v_1^{\mathrm{ff}}}{\partial x_2}\bigg|_\Sigma\,.$$

At this stage, corrections of the compressibility effects using boundary layer problems are needed similar to [6, Sect. 3.2.5] in order to obtain accurate error estimates (Theorem 1). Note that all further corrections of approximation $\{\mathbf{v}_{\mathrm{approx}}^{7,\varepsilon}, p_{\mathrm{approx}}^{7,\varepsilon}\}$ are of higher order and do not appear in the derived conditions (4)–(6). Thus, we do not provide details on these corrections here.

Final Approximation. We obtain the following final error functions

$$\mathbf{U}^\varepsilon = \mathbf{v}^\varepsilon - \mathbf{v}^{7,\varepsilon}_{\text{approx}} + \mathcal{O}_{L^2(\Omega^\varepsilon)^2}(\varepsilon^3)\,, \quad P^\varepsilon = p^\varepsilon - p^{7,\varepsilon}_{\text{approx}} + \mathcal{O}_{L^2(\Omega^\varepsilon)}(\varepsilon^2)\,. \tag{20}$$

Note that higher-order terms due to corrections of compressibility effects are included in $\mathcal{O}_{L^2(\Omega^\varepsilon)^2}(\varepsilon^3)$ for the velocity and in $\mathcal{O}_{L^2(\Omega^\varepsilon)}(\varepsilon^2)$ for the pressure.

We formulate interface condition (6) in such a way that all integral terms over Σ appearing in the weak formulation vanish. Interface conditions (4) and (5) are derived due to the requirement $[\![\mathbf{U}^\varepsilon]\!]_\Sigma = \mathbf{0}$ necessary for $\mathbf{U}^\varepsilon \in H^1(\Omega^\varepsilon)^2$. With the constructed error functions (20) and under the assumptions given in Remark 3, we obtain the following result.

Corollary 1 *For* $\mathbf{U}^\varepsilon$ *and* P^ε *defined in (20) the following estimate holds*

$$\|\nabla \mathbf{U}^\varepsilon\|_{L^2(\Omega^\varepsilon)^{2\times 2}} \leq C\varepsilon^{7/2}\|P^\varepsilon\|_{L^2(\Omega^\varepsilon)} + C\varepsilon^{5/2}\|\nabla \mathbf{U}^\varepsilon\|_{L^2(\Omega^\varepsilon)^{2\times 2}}\,. \tag{21}$$

Estimate (21) improves the result from Corollary 3.5 in [6].

Remark 3 For the proof of Corollary 1 we assume that the derivatives $\frac{\partial^2 v_k^{\text{ff}}}{\partial x_i \partial x_j}, \frac{\partial^2 p^{\text{pm}}}{\partial x_i \partial x_j}$ exist and are continuous for $i, j, k = 1, 2$ satisfying the following uniform bounds w.r.t. $\varepsilon > 0$ for a constant $C > 0$:

$$\left\| \frac{\partial^2 v_k^{\text{ff}}}{\partial x_i \partial x_j} \right\|_{C^0(\Omega_{\text{ff}})} \leq C\,, \qquad \left\| \frac{\partial^2 p^{\text{pm}}}{\partial x_i \partial x_j} \right\|_{C^0(\Omega^\varepsilon_{\text{pm}})} \leq C\,.$$

3.2 *Error Estimates for Model Approximation*

In this section, we provide the error estimates of the model approximation that justify the homogenization ansatz (8) on which the derivation of the interface conditions (4)–(6) is based. Under the assumptions given in Remark 3 and using the theory of very weak solutions [9], we obtain the following result.

Theorem 1 *Let us suppose the geometry as described in Sect. 2 and the velocity and pressure error function given by (20). Then, for a fixed* $\tilde{\varepsilon} > 0$ *there exists a constant* $C > 0$ *such that the following estimates hold for all* $0 < \varepsilon < \tilde{\varepsilon}$*:*

$$\|\nabla \mathbf{U}^\varepsilon\|_{L^2(\Omega)^{2\times 2}} \leq C\varepsilon^{5/2}\,, \quad \|\mathbf{U}^\varepsilon\|_{L^2(\Omega_{\text{pm}})^2} \leq C\varepsilon^{7/2}\,, \quad \|\mathbf{U}^\varepsilon\|_{L^2(\Omega_{\text{ff}})^2} \leq C\varepsilon^3\,,$$
$$\|\mathbf{U}^\varepsilon\|_{L^2(\Sigma)^2} \leq C\varepsilon^3\,, \qquad \|P^\varepsilon\|_{L^2(\Omega^\varepsilon)} \leq C\varepsilon^{3/2}\,.$$

4 Discussion and Future Work

In this work, we derived new generalized coupling conditions for arbitrary flows to the fluid–porous interface in Stokes–Darcy systems using homogenization and boundary layer theory. These conditions are an extension of the ones proposed in [6] by additional higher-order terms. To justify the new conditions we derived error estimates of the model approximation based on regularity and uniform boundedness assumptions for the free-flow velocity and the porous-medium pressure. Validation of the developed conditions by comparison of pore-scale resolved to macroscale numerical simulations will be presented at the conference. Moreover, the derived conditions will be compared to existing interface conditions in the literature, e.g., [6, 10].

Acknowledgements The work is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) Project number 490872182, Project number 327154368 - SFB 1313 and the French National Research Agency (ANR-PRCI) Identification number ANR-21-CE40-0018-01.

References

1. Beavers, G.S., Joseph, D.D.: Boundary conditions at a naturally permeable wall. J. Fluid Mech. **30**, 197–207 (1967). https://doi.org/10.1017/S0022112067001375
2. 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). https://doi.org/10.1016/j.cma.2014.10.050
3. Dawson, C.: A continuous/discontinuous Galerkin framework for modeling coupled subsurface and surface water flow. Comput. Geosci. **12**, 451–472 (2008). https://doi.org/10.1007/s10596-008-9085-y
4. Discacciati, M., Quarteroni, A.: Navier-Stokes/Darcy coupling: modeling, analysis, and numerical approximation. Rev. Mat. Complut. **22**, 315–426 (2009). https://doi.org/10.5209/rev_REMA.2009.v22.n2.16263
5. Eggenweiler, E., Rybak, I.: Unsuitability of the Beavers-Joseph interface condition for filtration problems. J. Fluid Mech. **892**, A10 (2020). https://doi.org/10.1017/jfm.2020.194
6. Eggenweiler, E., Rybak, I.: Effective coupling conditions for arbitrary flows in Stokes-Darcy systems. Multiscale Model. Simul. **19**, 731–757 (2021). https://doi.org/10.1137/20M1346638
7. 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)
8. Jäger, W., Mikelić, A., Neuss, N.: Asymptotic analysis of the laminar viscous flow over a porous bed. SIAM J. Sci. Comput. **2**, 2006–2028 (2001). https://doi.org/10.1137/S1064827599360339
9. Marciniak-Czochra, A., Mikelić, A.: Effective pressure interface law for transport phenomena between an unconfined fluid and a porous medium using homogenization. Multiscale Model. Simul. **10**, 285–305 (2012). https://doi.org/10.1137/110838248
10. Sudhakar, Y., Lācis, U., Pasche, S., Bagheri, S.: Higher-order homogenized boundary conditions for flows over rough and porous surfaces. Transp. Porous Med. **136**, 1–42 (2021). https://doi.org/10.1007/s11242-020-01495-w

Two Entropic Finite Volume Schemes for a Nernst–Planck–Poisson System with Ion Volume Constraints

Jürgen Fuhrmann, Benoît Gaudeul, and Christine Keller

Abstract Modeling and simulation of ion transport in electrolytes is an important tool to investigate electrochemical devices as well as biological systems at the cell scale. Well designed models follow first principles of non-equilibrium thermodynamics and include the fact that ions have a finite size. It is highly desirable that these properties are valid as well for discretized models. In this contribution, we present two numerical fluxes for two-point flux finite volume schemes which fulfill these requirements. We review recent results on entropic behavior and convergence. Concluding, we present first simulation results for biological ion channels.

Keywords Finite volume methods · Drift-diffusion equations · Generalized Nernst–Planck–Poisson system · Finite size effects · Ion channels

1 Introduction

Consider a bounded connected polytopal domain $\Omega \subset \mathbb{R}^d$, and finite simulation horizon $T > 0$. We model the evolution of the concentration c_0 of a solvent and N dissolved species: c_i, $i \in [\![1, N]\!]$. Due to finite particle sizes, the mixture satisfies a volume filling constraint $\sum_{i=0}^{N} v_i c_i = 1$, where v_i are the molar volumes of a species. We will use this constraint using ratios of molar volumes $k_i = \frac{v_i}{v_0}$: $\sum_{i=0}^{N} k_i c_i = \frac{1}{v_0}$. The coefficients $(k_1, \ldots, k_N)$ are parameters of the problem and k_0 is by defini-

J. Fuhrmann (✉) · C. Keller
Weierstrass Institute (WIAS), Mohrenstr. 39, 10117 Berlin, Germany
e-mail: juergen.fuhrmann@wias-berlin.de

C. Keller
e-mail: christine.keller@wias-berlin.de

B. Gaudeul
Université Paris-Saclay, CNRS, Laboratoire de mathématiques d'Orsay, 91405 Orsay, France
e-mail: benoit.gaudeul@universite-paris-saclay.fr

E. Franck et al. (eds.), *Finite Volumes for Complex Applications X—Volume 1, Elliptic and Parabolic Problems*, Springer Proceedings in Mathematics & Statistics 432,
https://doi.org/10.1007/978-3-031-40864-9_23

tion equal to 1. As the molar volumes are not the same, the total concentration $\overline{c} := \sum_{i=0}^{N} c_i$ is not uniform. The set of positive concentrations c_i, $i \in [\![1, N]\!]$ such that c_0 is positive is denoted by

$$\mathcal{A} = \{(c_1, ..., c_N) \in (0, +\infty)^N | c_0 := \frac{1}{v_0} - \sum_{i=1}^{N} k_i c_i > 0\}.$$

For the sake of clarity, we will let $C = (c_1, ..., c_N) \in \mathcal{A}$ and consider c_0 and $\overline{c}$ as functions of C without clearly expressing the dependency. The dissolved species follow a conservation equation:

$$\partial_t c_i - \operatorname{div} D_i \mathcal{N}_i = 0, \qquad \mathcal{N}_i = c_i \nabla \left(h_i(C) + z_i \Phi\right) \qquad \forall i \in [\![1, N]\!]. \tag{1}$$

where z_i the charge number and $D_i > 0$ the diffusion coefficient are parameters of the problem, while $h_i(C)$ the chemical potential depends on all the concentrations through:

$$h_i(C) = \log \frac{c_i}{\overline{c}} - k_i \log \frac{c_0}{\overline{c}} \qquad \forall i \in [\![1, N]\!]. \tag{2}$$

This system is supplemented with a Poisson equation for the potential:

$$-\Delta \Phi = \sum_{i=1}^{N} z_i c_i, \tag{3}$$

Note that we have assumed that the solvent carries no charge, in other word that $z_0 = 0$. As in [1], we consider a Dirichlet boundary condition for the potential on a non-negligible part of the boundary $\Gamma_{\mathrm{D}} \subset \partial\Omega$ and homogeneous Neumann boundary condition on $\Gamma_{\mathrm{N}} = \partial\Omega \setminus \Gamma_{\mathrm{D}}$:

$$\Phi = \Phi^D \quad \text{on } (0, T) \times \Gamma_{\mathrm{D}}, \qquad \nabla\Phi \cdot n = 0 \quad \text{on } (0, T) \times \Gamma_{\mathrm{N}},$$

where Φ^D is in $H^1(\Omega) \cap L^\infty(\Omega)$ and assumed to be constant in time.

The system is supplemented with the following no flux boundary conditions for the concentrations:

$$c_i \nabla \left(h_i(C) + z_i \Phi\right) \cdot n = 0 \quad \text{on } (0, T) \times \partial\Omega, \text{ for all } i \in [\![1, N]\!],$$

and with an initial condition C^0 satisfying:

$$C^0 \in L^\infty(\Omega, \bar{\mathcal{A}}) \qquad \text{and} \qquad \int_\Omega c_i^0(x)\, \mathrm{d}x > 0 \quad \forall i \in [\![0, N]\!]. \tag{4}$$

Notice that, for any $t \in [0, T]$, $i \in [\![1, N]\!]$:

$$\int_\Omega c_i(0,x)\,\mathrm{d}x = \int_\Omega c_i(t,x)\,\mathrm{d}x,$$

so that the mass is preserved. Another key property of the system is the dissipation of a free energy. In this case, the chemical free energy density $H(C)$ is defined as follows:

$$H(C) := \sum_{i=0}^{N} c_i \log\left(\frac{c_i}{\overline{c}}\right) = \sum_{i=0}^{N} c_i \log c_i - \overline{c}\log\overline{c}.$$

The total free energy is formed by the integral of the chemical free energy density and electrical terms:

$$E(C,\Phi) = \int_\Omega H(C) + \frac{|\nabla\Phi|^2}{2}\,\mathrm{d}x - \int_{\Gamma_\mathrm{D}} \Phi_D \nabla\Phi\cdot n\,\mathrm{d}\gamma. \tag{5}$$

Proposition 1 *Let (C,Φ) be smooth solutions of* (1)–(4) *such that $C(t,x)\in\mathcal{A}$. For such solutions, E is a convex Lyapunov functional. Moreover, we have:*

$$\partial_t E + \int_\Omega \sum_{i=1}^{n} D_i c_i |\nabla h_i(C) + z_i\Phi|^2\,\mathrm{d}x = 0.$$

The proof is found in [2], Proposition 1.1.

Our notion of weak solution relies on a reformulation of the fluxes:

$$\mathcal{N}_i = \nabla c_i + c_i\nabla\Big(-k_i\log c_0 + (k_i-1)\log\overline{c} + z_i\Phi\Big), \tag{6}$$

the space of H^1 functions satisfying the Dirichlet boundary conditions $\mathcal{H}_{\Gamma_\mathrm{D}} = \{f\in H^1(\Omega),\ f_{|\Gamma_\mathrm{D}} = 0\}$ and the cylinder $Q_T = (0,T)\times\Omega$. More precisely:

Definition 1 A couple (C,Φ) is a *weak solution of* (1)–(4) if $C\in L^\infty(Q_T;\overline{\mathcal{A}})$ with $\log(c_0)\in L^2((0,T);H^1(\Omega))$, and $\Phi-\Phi^D\in L^\infty((0,T),\mathcal{H}_{\Gamma_\mathrm{D}})$ and they satisfy

1. for all $\varphi\in C_c^\infty([0,T)\times\overline{\Omega})^N$, $i\in[\![1,N]\!]$

$$\iint_{Q_T} c_i\partial_t\varphi_i\,\mathrm{d}x\,\mathrm{d}t + \int_\Omega c_i^0\varphi_i(0,x)\,\mathrm{d}x - D_i\iint_{Q_T}\nabla c_i\cdot\nabla\varphi_i$$
$$-D_i\iint_{Q_T} c_i\nabla\big(-k_i\log c_0 + (k_i-1)\log\overline{c} + z_i\Phi\big)\cdot\nabla\varphi_i\,\mathrm{d}x\,\mathrm{d}t = 0;$$

2. for all $\psi\in\mathcal{H}_{\Gamma_\mathrm{D}}$ and almost all $t\in(0,T)$,

$$\lambda^2\int_\Omega \nabla\Phi(t,x)\cdot\nabla\psi(x)\,\mathrm{d}x = \int_\Omega \psi(x)\sum_{i=1}^{N} z_i c_i(t,x)\,\mathrm{d}x.$$

2 Two Point Flux Finite Volume Approximations

For the space discretization, we use the standard notation of an admissible finite volume mesh $\left(\mathcal{T}, \mathcal{E}, (\mathbf{x}_K)_{K\in\mathcal{T}}\right)$, see [2]. Control volumes are denoted by $K \in \mathcal{T}$ with respective measures m_K, edges are denoted by $\sigma \in \mathcal{E}$, and their $(d-1)$-dimensional measure by m_σ. Since our method relies on a two-point flux approximation (TPFA), we suppose that the mesh satisfies the classical orthogonality condition [3], Chap. 9. For the time discretization, we consider an increasing finite family of times $0 = t_0 < t_1 < \cdots < t_{N_T} = T$. We denote by $\Delta t_n = t_n - t_{n-1}$ for $1 \le n \le N_T$, by $\Delta\mathbf{t} = (\Delta t_n)_{1\le n\le N_T}$, and by $h_{\Delta\mathbf{t}} = \max_{1\le n\le N} \Delta t_n$.

We will use boldface notations for vectors whose number of components is dependent on the mesh while keeping the uppercase notation $\mathbf{C}$ when we also consider different species.

The initial data C^0 is discretized into $\left(C_K^0\right)_{K\in\mathcal{T}} \in \bar{\mathcal{A}}^{\mathcal{T}}$ by setting

$$c_{K,i}^0 = \frac{1}{|K|}\int_K c_i^0(x)\,\mathrm{d}x \qquad \forall K \in \mathcal{T}, i \in [\![1, N]\!]. \tag{7}$$

Assume that $\mathbf{C}^{n-1} = \left(C_K^{n-1}\right)_{K\in\mathcal{T}} \in \bar{\mathcal{A}}^{\mathcal{T}}$ is given for some $n > 0$. We define how to compute $(\mathbf{C}^n, \Phi^n) = \left(c_K^n, \Phi_K^n\right)_{K\in\mathcal{T}}$. To that extent, for all $K \in \mathcal{T}$ and all $\sigma \in \mathcal{E}_K = \mathcal{E}_{\text{int}} \cup \mathcal{E}_{\text{ext}}$, the set of interior and exterior edges, we define the mirror values $C_{K\sigma}^n$ (resp. $\Phi_{K\sigma}^n$) of C_K^n (resp. Φ_K^n) across σ by setting :

$$C_{K\sigma}^n = \begin{cases} C_L^n & \text{if } \sigma = K|L \in \mathcal{E}_{\text{int}}, \\ C_K^n & \text{if } \sigma \in \mathcal{E}_{\text{ext}}, \end{cases} \qquad \Phi_{K\sigma}^n = \begin{cases} \Phi_L^n & \text{if } \sigma = K|L \in \mathcal{E}_{\text{int}}, \\ \Phi_K^n & \text{if } \sigma \in \mathcal{E}^N, \\ \frac{1}{m_\sigma}\int_\sigma \Phi^D\,\mathrm{d}\gamma & \text{if } \sigma \in \mathcal{E}^D. \end{cases}$$

For $\sigma \in \mathcal{E}$, we set $d_\sigma = |\mathbf{x}_K - \mathbf{x}_L|$ if $\sigma = K|L \in \mathcal{E}_{\text{int}}$, $d_\sigma = |\mathbf{x}_K - \mathbf{x}_\sigma|$ if $\sigma \in \mathcal{E}_{\text{ext}}$, and $\tau_\sigma = \frac{m_\sigma}{d_\sigma}$. Given $\mathbf{u} = (u_K)_{K\in\mathcal{T}} \in \mathbb{R}^{\mathcal{T}}$, we define the oriented and absolute jumps of $\mathbf{u}$ across $\sigma \in \mathcal{E}_K$ by $D_{K\sigma}\mathbf{u} = u_{K\sigma} - u_K$, and $D_\sigma\mathbf{u} = |D_{K\sigma}\mathbf{u}|$.

Both 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:

$$-\sum_{\sigma\in\mathcal{E}_K} \tau_\sigma D_{K\sigma}\Phi^n = m_K \sum_{i=1}^N z_i c_{K,i}^n, \qquad \forall K \in \mathcal{T}, \tag{8a}$$

$$m_K \frac{c_{K,i}^n - c_{K,i}^{n-1}}{\Delta t_n} + \sum_{\sigma\in\mathcal{E}_{K,\text{int}}} F_{K\sigma,i}^n = 0, \qquad \forall K \in \mathcal{T}, i \in [\![1, N]\!]. \tag{8b}$$

$$c^n_{K,0} = \frac{1}{v_0} - \sum_{i=1}^{N} k_i c^n_{K,i}, \qquad \forall K \in \mathcal{T}. \tag{8c}$$

To close the system (8), all that remains is to define the numerical fluxes $F^n_{K\sigma}$. They are defined with functions $\mathcal{F}_i$ of the primary unknowns $(C^n_K, C^n_L, \Phi^n_K, \Phi^n_L)$:

$$F^n_{K\sigma,i} = \begin{cases} 0 & \text{if } \sigma \in \mathcal{E}_{\text{ext}} \\ \tau_\sigma D_i \mathcal{F}_i(C_K, C_L, \Phi_K, \Phi_L) & \text{if } \sigma = K|L \in \mathcal{E}_{\text{int}} \end{cases} \tag{9}$$

The different schemes considered in this contribution correspond to different choices of $\mathcal{F}$. All of them satisfy $\mathcal{F}(C_K, C_L, \Phi_K, \Phi_L) = -\mathcal{F}(C_L, C_K, \Phi_L, \Phi_K)$, so that the numerical fluxes are locally conservative. Both schemes are extensions of the schemes studied in [1], and one of them is based on the Scharfetter-Gummel scheme [4] and features the Bernoulli function $B(u) = \frac{u}{e^u - 1}$.

The **centered flux** is derived from (1), which suggests the following definition:

$$\mathcal{F}_i(C_K, C_L, \Phi_K, \Phi_L) = -\frac{c_{K,i} + c_{L,i}}{2} D_{K\sigma}\big(h_i(\mathbf{C}) + z_i \Phi\big), \qquad \forall i \in [\![1, N]\!]. \tag{C}$$

The associate flux can be seen as a particular case in the TPFA context of the fluxes introduced in [5].

The **Sedan flux** is named after the SEDAN III simulator [6] which inspired this discretization approach. It was independently introduced as well in [7]. Formula (6) for the flux $\mathcal{N}_i$ suggests to use a classical Scharfetter-Gummel scheme, but for a modified potential $\Phi + \nu_i(C)$ instead of only Φ, where $\nu_i(C) = h_i(C) - \log c_i$. Thus, for all $i \in [\![1, N]\!]$, we let:

$$\mathcal{F}_i(C_K, C_L, \Phi_K, \Phi_L) = B\big(D_{K\sigma}(\nu_i(\mathbf{C}) + z_i\Phi)\big)c_{K,i} - B\big(D_{L\sigma}(\nu_i(\mathbf{C}) + z_i\Phi)\big)c_{L,i}. \tag{S}$$

2.1 Main Results

Energy decay is one of the key properties of the continuous model, see Proposition 1. This property is transposed to the discrete setting by both discretizations considered. The discrete energy functional $E_\mathcal{T}$ has to be thought of as a discrete counterpart of the continuous energy functional E, see (5). It is defined by:

$$E_\mathcal{T}(\mathbf{C}^n, \Phi^n) = \sum_{K \in \mathcal{T}} m_K H(C^n_K) + \frac{1}{2}\sum_{\sigma \in \mathcal{E}} \tau_\sigma \left(D_\sigma \Phi^n\right)^2 - \sum_{K \in \mathcal{T}} \sum_{\sigma \in \mathcal{E}^D \cap \mathcal{E}_K} \tau_\sigma \Phi^D_\sigma D_{K\sigma} \Phi^n.$$

Our first result focuses on a fixed mesh analysis. 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* $(\mathcal{T}, \mathcal{E}, (x_K)_{K\in\mathcal{T}})$ *be an admissible mesh and let* $\mathbf{C}^0$ *be defined by* (7)*. Then, for all* $1 \leq n \leq N_T$*, the nonlinear system of equations* (8)–(9)*, supplemented either with* (C) *or* (S)*, has a solution* $(\mathbf{C}^n, \Phi^n) \in \mathcal{A}^{\mathcal{T}} \times \mathbb{R}^{\mathcal{T}}$*. Moreover, the solution to the scheme satisfies, for all* $1 \leq n \leq N_T$*,*

$$E_{\mathcal{T}}(\mathbf{C}^n, \Phi^n) - E_{\mathcal{T}}(\mathbf{C}^{n-1}, \Phi^{n-1}) \leq \Delta t_n \sum_{i=1}^{N} \sum_{\sigma\in\mathcal{E}} F^n_{K\sigma,i} D_{K\sigma}(h_i(\mathbf{C}^n) + z_i \Phi^n),$$

and

$$\sum_{K\in\mathcal{T}} c_{K,i} m_K = \int_\Omega c_i^0(x)\,\mathrm{d}x \qquad \forall i \in [\![0, N]\!].$$

In [2], this result is stated in Theorem 2.1 and proven in Sect. 3 using the convexity of $H(C)$ and a topological invariant on a three stage homotopy.

Once a discrete solution to the scheme $(\mathbf{C}^n, \Phi^n)_{1\leq n\leq N}$ is at hand, we can define an approximate solution $(C_{\mathcal{T},\Delta\mathbf{t}}, \Phi_{\mathcal{T},\Delta\mathbf{t}})$. It is the piecewise constant function defined almost everywhere by

$$C_{\mathcal{T},\Delta\mathbf{t}}(t, \mathbf{x}) = C_K^n, \quad \Phi_{\mathcal{T},\Delta\mathbf{t}}(t, \mathbf{x}) = \Phi_K^n \quad \text{if } (t, \mathbf{x}) \in (t_{n-1}, t_n] \times K.$$

Let $\left(\mathcal{T}_m, \mathcal{E}_m, (\mathbf{x}_K)_{K\in\mathcal{T}_m}\right)_{m\geq 1}$ be a sequence of admissible meshes such that $h_{\mathcal{T}_m}, h_{\Delta\mathbf{t}_m} \underset{m\to\infty}{\longrightarrow} 0$ while the mesh regularity remains bounded (see [1] for the definition of the regularity of the mesh). A natural question is the convergence of the associated sequence of approximate solutions $(C_{\mathcal{T}_m,\Delta\mathbf{t}_m}, \Phi_{\mathcal{T}_m,\Delta\mathbf{t}_m})_{m\geq 1}$ towards a weak solution to the continuous problem. Our second result focuses on convergence and is stated in Theorem 2 (Theorem 2.2 in [2]). Its proof is detailed in [2], Sect. 4. The proof is based on compactness arguments and the identification of the limit requires a non-degeneracy assumption.

Theorem 2 *For the two schemes under study, a sequence of approximate solutions* $(\mathbf{C}_{\mathcal{T}_m,\Delta\mathbf{t}_m}, \Phi_{\mathcal{T}_m,\Delta\mathbf{t}_m})_{m\geq 1}$ *satisfies, up to a subsequence:*

$$C_{\mathcal{T}_m,\Delta\mathbf{t}_m} \underset{m\to\infty}{\longrightarrow} C \quad in\ L^2(Q_T)^{N+1}, \qquad \Phi_{\mathcal{T}_m,\Delta\mathbf{t}_m} \underset{m\to\infty}{\longrightarrow} \Phi \quad in\ L^2(Q_T).$$

Moreover if $\inf_{\substack{mesh\ m\\ n\in[\![1,N_{T,m}]\!]\\ K\in\mathcal{T}_m}} c^n_{m,K,0} > 0$*,* (C, Φ) *is a weak solution in the sense of Definition* 1*.*

3 Towards Simulation of Charge Transport in Ion Channels

Ion channels are pore-forming proteins in the membrane of biological cells that control a large part of biological processes. They are important targets for the development of medications and effective therapies. The behavior of ion channels can be studied by measuring the current response to a time-dependent voltage difference. The interpretation of the measured current-voltage (IV) relation is therefore of great importance for biology, physiology, and medicine.

We provide first results towards the numerical simulation on a calcium (Ca^{2+}) selective ion channel using the Sedan scheme introduced in this paper.

For this purpose, the Nernst-Planck-Poisson system consisting of (1)–(2)–(3) is adapted to the physical situation by removing the simplifications introduced for the theoretical investigations, e.g. in the left hand side of (3), the electrostatic permeability is introduced, and the right hand side is multiplied by the Faraday constant. Calling for further investigation is the introduction of the pressure p which remains from the consideration of the Stokes equation for electrolyte flow in mechanical equilibrium [8]. After [9], define p by

$$-\Delta p = \nabla \cdot q \nabla \phi \tag{10}$$

and modify the chemical potentials to

$$h_i(C, p) = \log \frac{c_i}{\overline{c}} - k_i \log \frac{c_0}{\overline{c}} + (v_i - \kappa_i v_0)(p - p_{ref}) \qquad \forall i \in [\![1, N]\!].$$

We define a rotational symmetric approximate geometry of an ion channel with charged walls connecting the extracellular and the intracellular space. Simulations were reduced to the appropriate 2D case. The model was implemented in Julia on top of the VoronoiFVM.jl [10] package.

We investigate the stationary transport of ions. Dirichlet boundary conditions fix the concentrations in both extracellular and intracellular domains: $c_i = [\mathrm{C}_i]_{\text{intra.}}$ on $\Gamma_{\text{intra.}}$ and $c_i = [\mathrm{C}_i]_{\text{extra.}}$ on $\Gamma_{\text{extra.}}$ for $i = 1, \ldots, N$. We apply a varying potential difference along the channel: $\Phi = \Phi_{\text{intra.}}$ on $\Gamma_{\text{intra.}}$ and $\Phi = \Phi_{\text{extra.}}$ on $\Gamma_{\text{extra.}}$. On the channel wall, we introduce $\mathcal{N}_i \cdot \mathbf{n} = 0$ on Γ_{wall} (impermeable wall) and $\nabla \phi \cdot \mathbf{n} = \frac{\sigma}{\varepsilon}$ on Γ_{wall} (charged wall). For the pressure, we use $(\nabla p + q \nabla \phi) \cdot \mathbf{n} = 0$.

Figure 1 shows the distribution of the concentrations for the different ion species. The parameter values used are given in Table 1. For this set of parameters, cations accumulate in the channel due to the charged wall, while the concentration of anions in the channel is the lowest. This leads to an increase of the pressure inside the channel. The distributions of pressure and potential are given in Fig. 2. We calculated the IV relation for different applied membrane potentials of $\Delta\Phi = \Phi_{\text{intra.}} - \Phi_{\text{extra.}} \in [-80, 80]$ mV. We expect an inward cation current for a membrane potential $\Delta\Phi < 20$ mV and an outward current for a membrane potential $\Delta\Phi > 20$ mV. In addition to the membrane potential we varied the channel radii, see Fig. 3 (left). We observe, that

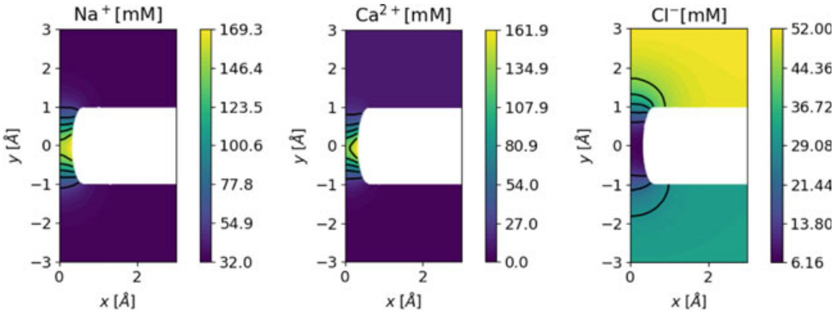


Fig. 1 Concentrations of Na^+, Ca^{2+}, Cl^- ions. The parameter values used for the simulation are given in Table 1

Table 1 Parameter values from [7] used to simulate a calcium ion channel

Symbol	Meaning	Value	Unit
$D_{Ca^{2+}}$, D_{Na^+}, D_{Cl^-}	Diffusion coeff.	$[0.79, 1.33, 2.03] \cdot 10^{-5}$	cm^2/s
z_{H_2O}, $z_{Ca^{2+}}$, z_{Na^+}, z_{Cl^-}	Ion charge nr.	0, 2, 1, –1	
M_{H_2O}, $\mathrm{M}_{Ca^{2+}}$, M_{Na^+}, M_{Cl^-}	Molar weights	18.0, 40.1, 23.0, 35.5	g/mol
v_{H_2O}, $\mathrm{v}_{Ca^{2+}}$, v_{Na^+}, v_{Cl^-}	Molar volumes	$[55.4, 26.20, 23.78, 17.39]\cdot 10^{-6}$	m^3/mol
σ/ϵ	Wall charge	$-1.6 \cdot 10^{-20}$	C/nm^2
r	Channel radius	3	Å
$[CaCl_2]_{\text{intra.}}$, $[CaCl_2]_{\text{extra.}}$	Bulk conc.	0, 10	mM
$[NaCl]_{\text{intra.}}$, $[NaCl]_{\text{extra.}}$	Bulk conc.	32, 32	mM
$\Phi_{\text{intra.}}$, $\Phi_{\text{extra.}}$	Bulk potential	–20, 0	mV
$\kappa_{Ca^{2+}}$, κ_{Na^+}, κ_{Cl^-}	Solvation nr.	10.0, 7.5, 3.9	

the current increases for wider channels. An increase in the calcium concentration in the extracellular fluid leads to an increase in the ionic current, see Fig. 3 (right).

Future research aims at the introduction of a selectivity filter within the ion channel, and the coupling to an elasticity model for the channel walls.

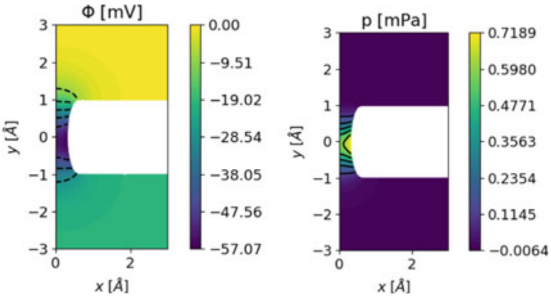


Fig. 2 Distribution of electrostatic potential (left) and pressure (right). The parameter values used for the simulation are given in Table 1

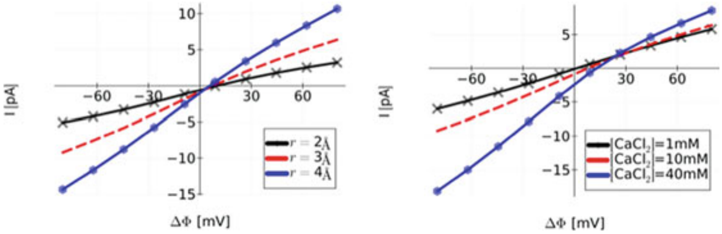


Fig. 3 IV curves for different channel radii (left) and Ca^{2+} concentrations (right). The parameter values used for the simulation are given in Table 1. A membrane potential of $\Delta\Phi = \Phi_{\text{intra.}} - \Phi_{\text{extra.}} \in [-80, 80]$ mV was applied

Acknowledgements C.K. is supported by the German Research Foundation (DFG) via the Berlin Mathematics Research Center MATH+ (project AA1-14).

References

1. 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. IMA J. Numer. Anal. **41**(1), 271–314 (2021)
2. Gaudeul, B., Fuhrmann, J.: Entropy and convergence analysis for two finite volume schemes for a Nernst-Planck-Poisson system with ion volume constraints. Numer. Math. **151**(1), 99–149 (2022)
3. Eymard, R., Gallouët, T., Herbin, R.: Finite volume methods. In: Handbook of Numerical Analysis, vol. 7, pp. 713–1018. Elsevier (2000)
4. Scharfetter, D., Gummel, H.: Large-signal analysis of a silicon Read diode oscillator. IEEE Trans. Electron Devic. **16**(1), 64–77 (1969)
5. 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)
6. Yu, Z., Dutton, R.: SEDAN III simulator (1988). http://www-tcad.stanford.edu/oldftp_sw/Sedan-III/relB.8830.tar.Z

7. Liu, J.-L., Eisenberg, B.: Molecular mean-field theory of ionic solutions: a Poisson-Nernst-Planck-Bikerman model. Entropy **22**(5), 550 (2020)
8. Dreyer, W., Guhlke, C., Landstorfer, M.: A mixture theory of electrolytes containing solvation effects. Electrochem. Commun. **43**, 75–78 (2014)
9. Fuhrmann, J.: Comparison and numerical treatment of generalised Nernst-Planck models. Comput. Phys. Commun. **196**, 166–178 (2015)
10. Fuhrmann, J.: VoronoiFVM.jl: Solver for coupled nonlinear partial differential equations based on the Voronoi finite volume method (2022). https://doi.org/10.5281/zenodo.3529808

Dimensional Reduction by Fourier Analysis of a Stokes-Darcy Fracture Model

Martin J. Gander, Julian Hennicker, Roland Masson, and Tommaso Vanzan

Abstract We consider a Stokes flow along a thin fracture coupled to a Darcy flow in the surrounding matrix domain. In order to derive a dimensionally reduced model representing the fracture as an interface coupled to the surrounding matrix, we extend the methodology based on Fourier analysis developed in [1] for a Darcy-Darcy coupling. We show that this approach not only allows us to derive error estimates between the solutions of the full and mixed-dimensional models, but also leads to a model correction term compared with what is obtained from the classical reduction technique based on integration along the fracture width combined with profile closure assumptions [2, 3].

Keywords Dimensionally reduced fracture model · Mixed-dimensional model · Discrete fracture matrix model · Stokes-Darcy coupling · Fourier analysis

M. J. Gander · J. Hennicker (✉)
Université de Genève, Geneva, Switzerland
e-mail: julian.hennicker@gmail.com

M. J. Gander
e-mail: martin.gander@unige.ch

R. Masson
Université Côte d'Azur, Inria, CNRS, Laboratoire J.A. Dieudonné, team Coffee, Parc Valrose, 06108 Nice Cedex 02, France
e-mail: roland.masson@univ-cotedazur.fr

T. Vanzan
CSQI Chair, Ecole Polytecnique Fédérale de Lausanne, Lausanne, Switzerland
e-mail: tommaso.vanzan@epfl.ch

E. Franck et al. (eds.), *Finite Volumes for Complex Applications X—Volume 1, Elliptic and Parabolic Problems*, Springer Proceedings in Mathematics & Statistics 432,
https://doi.org/10.1007/978-3-031-40864-9_24

1 Stokes-Darcy Fracture Model

Let us consider the matrix domains $\Omega_1 = (-L_1, -\delta) \times \mathbb{R}$, $\Omega_2 = (\delta, L_2) \times \mathbb{R}$ and the fracture domain $\Omega_f = (-\delta, \delta) \times \mathbb{R}$ as illustrated in Fig. 1.

We consider the following Darcy (in the matrix) Stokes (in the fracture) coupled model:

$$\begin{aligned} -\mu\Delta\mathbf{u} + \nabla p &= 0 && \text{on } \Omega_f,\\ \operatorname{div}\mathbf{u} &= 0 && \text{on } \Omega_f,\\ \operatorname{div}(\mathbf{u}_i) &= f_i && \text{on } \Omega_i, i = 1, 2,\\ \mathbf{u}_i &= -\mathbb{K}_i\nabla p_i && \text{on } \Omega_i, i = 1, 2, \end{aligned}$$

combined with the following coupling conditions on $\Gamma_1 = \{-\delta\} \times \mathbb{R}$ and $\Gamma_2 = \{\delta\} \times \mathbb{R}$:

$$\mathbf{u}_i \cdot \mathbf{n}_i = \mathbf{u} \cdot \mathbf{n}_i \qquad \text{on } \Gamma_i, i = 1, 2, \tag{1}$$

$$p_i = p - \mu(\nabla\mathbf{u}\,\mathbf{n}_i) \cdot \mathbf{n}_i \qquad \text{on } \Gamma_i, i = 1, 2, \tag{2}$$

$$\mu(\nabla\mathbf{u}\,\mathbf{n}_i) \cdot \boldsymbol{\tau} = \alpha\mathbf{u} \cdot \boldsymbol{\tau} \qquad \text{on } \Gamma_i, i = 1, 2, \tag{3}$$

where $\mathbf{n}_i$ is the unit normal vector on Γ_i, oriented outward of Ω_i, $\boldsymbol{\tau}$ is the unit vector tangent to the interfaces oriented in the positive y direction, $\mu > 0$ is the fluid kinematic viscosity, α is the Beaver-Joseph-Saffman parameter assumed to be constant for simplicity, and $\mathbb{K}_i$ is the permeability tensor in subdomain Ω_i. We also set $\mathbf{n} = \mathbf{n}_1 = -\mathbf{n}_2$ in what follows.

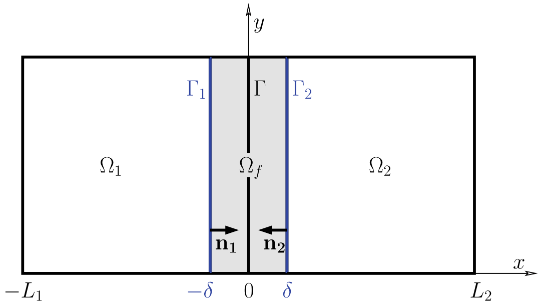


Fig. 1 Model problem geometry, with $\Omega_1 = (-L_1, -\delta) \times \Gamma$, $\Omega_2 = (\delta, L_2) \times \Gamma$, $\Gamma_1 = \{-\delta\} \times \Gamma$, $\Gamma_2 = \{\delta\} \times \Gamma$, and $\Omega_f = (-\delta, \delta) \times \Gamma$. The unit normals on Γ_j pointing outside of Ω_j are denoted by $\mathbf{n}_j$, $j = 1, 2$. Note that the Fourier analysis below will be carried out on unbounded domains by setting $\Gamma = \mathbb{R}$

2 Dimensional Reduction by Fourier Analysis

2.1 Elimination of the Fracture by Fourier Analysis

Let us set $\mathbf{u} = \begin{pmatrix} u \\ v \end{pmatrix}$, $\delta_i = (-1)^i \delta$, and take the Fourier transform in the y direction of the Stokes equations and of the transmission conditions. Setting in short

$$\hat{u}_i = \widehat{\mathbf{u}_i \cdot \mathbf{n}}(\delta_i, k), \quad \hat{p}_i = \hat{p}_i(\delta_i, k),$$

leads to the system

$$\begin{aligned}
-\mu\partial_{xx}\hat{u}(x,k) + \mu k^2\hat{u}(x,k) + \partial_x\hat{p}(x,k) &= 0 && x \in (-\delta,\delta), && (4)\\
-\mu\partial_{xx}\hat{v}(x,k) + \mu k^2\hat{v}(x,k) + ik\hat{p}(x,k) &= 0 && x \in (-\delta,\delta), && (5)\\
\partial_x\hat{u}(x,k) + ik\hat{v}(x,k) &= 0 && x \in (-\delta,\delta), && (6)\\
\hat{p}_i = \hat{p}(\delta_i,k) - \mu\partial_x\hat{u}(\delta_i,k) & && i = 1,2, && (7)\\
(-1)^{i+1}\mu\partial_x\hat{v}(\delta_i,k) = \alpha\hat{v}(\delta_i,k) & && i = 1,2, && (8)\\
\hat{u}_i = \hat{u}(\delta_i,k) & && i = 1,2. && (9)
\end{aligned}$$

Using that $\Delta p = 0$ yields the equation $\partial_{xx}\hat{p}(x,k) - k^2\hat{p}(x,k) = 0$, whose solution is $\hat{p}(x,k) = C_1(k)e^{|k|x} + C_2(k)e^{-|k|x}$. We next substitute this pressure solution $\hat{p}$ into the momentum equations (4)–(5) of the previous system yielding four additional integration constants $C_j(k)$ with $j = 3, 4, 5, 6$. These 6 integration constants can be computed using the divergence free condition (6) (providing two additional equations on these 6 constants) and the transmission conditions (7)–(8). The last two transmission conditions (9) are then used to provide the following two exact transmission conditions of the model posed on $\Omega_1 \cup \Omega_2$ eliminating the fracture model:

$$\mu|k| \begin{pmatrix} H_1^{ex}(|k|\delta) & 0 \\ 0 & H_2^{ex}(|k|\delta) \end{pmatrix} \begin{pmatrix} \hat{u}_2 + \hat{u}_2 \\ \hat{u}_1 - \hat{u}_2 \end{pmatrix} = \begin{pmatrix} \hat{p}_1 - \hat{p}_2 \\ \hat{p}_1 + \hat{p}_2 \end{pmatrix}, \tag{10}$$

where, setting $\xi := |k|\delta$,

$$\begin{aligned}
H_1^{ex}(\xi) &= \frac{-4(1 + C_\alpha\xi^2)e^{2\xi} + (2 + 3C_\alpha\xi)e^{4\xi} + (2 - 3C_\alpha\xi)}{4\xi(1 + C_\alpha)e^{2\xi} + (1 + 2C_\alpha\xi)e^{4\xi} + (2C_\alpha\xi - 1)},\\
H_2^{ex}(\xi) &= \frac{4(1 + C_\alpha\xi^2)e^{2\xi} + (2 + 3C_\alpha\xi)e^{4\xi} + (2 - 3C_\alpha\xi)}{-4\xi(1 + C_\alpha)e^{2\xi} + (1 + 2C_\alpha\xi)e^{4\xi} + (2C_\alpha\xi - 1)},
\end{aligned} \tag{11}$$

and $C_\alpha := \dfrac{\mu}{\alpha\delta}$ is a dimensionless parameter governing the Beaver-Joseph-Saffman condition (3). To simplify the presentation, we develop in the following the analysis for the case $\alpha = +\infty$, i.e. $C_\alpha = 0$, corresponding to replacing the Beaver-Joseph-

Saffman condition by the no slip condition $\mathbf{u} \cdot \boldsymbol{\tau} = 0$. This approximation is valid for a wide range of not too large rock permeabilities. The discussion of the general case is postponed to Sect. 4.

2.2 Reduced Transmission Conditions

An asymptotic expansion of H_i^{ex}, $i = 1, 2$, with respect to small ξ provides the reduced transmission conditions

$$\mu|k| \begin{pmatrix} H_1^{red}(|k|\delta) & 0 \\ 0 & H_2^{red}(|k|\delta) \end{pmatrix} \begin{pmatrix} \hat{u}_2 + \hat{u}_2 \\ \hat{u}_1 - \hat{u}_2 \end{pmatrix} = \begin{pmatrix} \hat{p}_1 - \hat{p}_2 \\ \hat{p}_1 + \hat{p}_2 \end{pmatrix}, \tag{12}$$

with the approximation H_i^{red} of H_i^{ex} given by

$$H_1^{red}(\xi) = \xi, \qquad H_2^{red}(\xi) = \frac{3}{\xi^3}\Big(1 + \frac{4}{5}\xi^2\Big),$$

at order $O(\xi^5)$ and $O(\xi)$. Note that these orders of approximation are the highest ones providing a well-posed reduced model, i.e. such that $|k|\ H_i^{red}(|k|\delta) > 0$ for all $k > 0$. Setting for $i = 1, 2$

$$\gamma_i^{\mathbf{n}}\mathbf{u}_i = \mathbf{u}_i \cdot \mathbf{n}\Big(\delta_i, \cdot\Big), \qquad \gamma_i p_i = p_1\Big(\delta_i, \cdot\Big),$$

provides the following reduced model with elimination of the fracture unknowns:

$$\begin{aligned}
&\operatorname{div}(\mathbf{u}_i) = f_i && \text{on } \Omega_i,\ i = 1, 2,\\
&\mathbf{u}_i = -\mathbb{K}_i \nabla p_i && \text{on } \Omega_i,\ i = 1, 2,\\
&-\mu\partial_{yy}\frac{(\gamma_1^{\mathbf{n}}\mathbf{u}_1 + \gamma_2^{\mathbf{n}}\mathbf{u}_2)}{2} = \frac{\gamma_1 p_1 - \gamma_2 p_2}{2\delta} && \text{on } \mathbb{R},\\
&\mu\Big(1 - \frac{4}{5}\delta^2\,\partial_{yy}\Big)\frac{(\gamma_1^{\mathbf{n}}\mathbf{u}_1 - \gamma_2^{\mathbf{n}}\mathbf{u}_2)}{2\delta} = -\frac{\delta^2}{3}\partial_{yy}\frac{(\gamma_1 p_1 + \gamma_2 p_2)}{2} && \text{on } \mathbb{R}.
\end{aligned} \tag{13}$$

2.3 Reconstruction Along the Fracture

As in [2, 3], the reconstruction along the fracture starts with averaging both the Stokes unknowns and equations along the fracture width, setting

$$\hat{P} := \frac{1}{2\delta}\int_{-\delta}^{\delta} \hat{p}(x,k)dx, \quad \hat{U} := \frac{1}{2\delta}\int_{-\delta}^{\delta} \hat{u}(x,k)dx, \quad \hat{V} := \frac{1}{2\delta}\int_{-\delta}^{\delta} \hat{v}(x,k)dx.$$

From the divergence free condition (6), we obtain by integration along the fracture width the reduced material conservation equation

$$ik2\delta\ \hat{V} = \hat{u}_1 - \hat{u}_2. \tag{14}$$

By integration of the momentum equation (4), and taking into account the pressure jump condition (7), we get that

$$\mu|k|^2 2\delta \hat{U} = (\hat{p}_1 - \hat{p}_2). \tag{15}$$

By integration of the momentum equation (5), we get the relation

$$-\mu(\partial_x \hat{v}(\delta, k) - \partial_x \hat{v}(-\delta, k)) + \mu|k|^2 2\delta\ \hat{V} + ik2\delta\ \hat{P} = 0. \tag{16}$$

Then, the classical approach developed in [2, 3] amounts to make profile assumptions along the width for U, V and P in order to derive both the coupling conditions and the approximation of the wall friction term $-\mu(\partial_x \hat{v}(\delta, k) - \partial_x \hat{v}(-\delta, k))$.

In our approach the coupling conditions were already derived by Fourier analysis and asymptotic expansions. The approximation of the friction term is obtained in the same way from the Fourier expression of $\partial_x \hat{v}(x, k)$ which can be shown to lead to

$$F^{ex}(\xi) = \delta \frac{(\partial_x \hat{v}(-\delta, k) - \partial_x \hat{v}(\delta, k))}{\hat{V}} = -2 \frac{\xi^2 \left(4\xi e^{2\xi} + e^{4\xi} - 1\right)}{4\xi e^{2\xi} - e^{4\xi} + 1}.$$

By asymptotic expansion for small $\xi = |k|\delta$, we obtain the following approximation F^{red} of F^{ex} at order $O(\xi^4)$:

$$F^{red}(\xi) = 6 + \frac{4}{5}\xi^2,$$

which leads to

$$\frac{6\mu}{\delta}\hat{V} + \widetilde{\mu}|k|^2 2\delta\ \hat{V} + ik2\delta\ \hat{P} = 0, \tag{17}$$

with the modified tangential viscosity $\widetilde{\mu} = \left(1 + \frac{2}{5}\right)\mu$.

Equations (14)–(17) are the reconstructed equations along the fracture. These equations can be combined with (13) in order to obtain the following coupled formulation of the reduced model:

$$\begin{aligned}
&\mathrm{div}(\mathbf{u}_i) = f_i && \text{on } \Omega_i,\ i = 1, 2,\\
&\mathbf{u}_i = -\mathbb{K}_i \nabla p_i && \text{on } \Omega_i,\ i = 1, 2,\\
&2\delta\ \partial_y V = \gamma_1^{\mathbf{n}} \mathbf{u}_1 - \gamma_2^{\mathbf{n}} \mathbf{u}_2, && \text{on } \mathbb{R},\\
&-2\mu\delta\ \partial_{yy} U = \gamma_1 p_1 - \gamma_2 p_2 && \text{on } \mathbb{R},
\end{aligned}$$

$$
\begin{aligned}
&6\frac{\mu}{\delta}V - 2\widetilde{\mu}\delta\, \partial_{yy}V + 2\delta\, \partial_y P = 0 && \text{on } \mathbb{R},\\
&U = \frac{\gamma_1^{\mathbf{n}}\mathbf{u}_1 + \gamma_2^{\mathbf{n}}\mathbf{u}_2}{2} && \text{on } \mathbb{R},\\
&\frac{\mu}{\delta}\left(\gamma_1^{\mathbf{n}}\mathbf{u}_1 - \gamma_2^{\mathbf{n}}\mathbf{u}_2\right) = \gamma_1 p_1 + \gamma_2 p_2 - 2P && \text{on } \mathbb{R}. \qquad (18)
\end{aligned}
$$

Compared with the classical approach developed in [2, 3] our methodology leads to a correction term which amounts to replace the tangential viscosity μ by $\widetilde{\mu}$ in the fifth equation of (18). This correction plays an essential role to obtain the error estimates shown in the next section.

3 Error Estimates

We use the same setting as in [1] for the Darcy subproblems assuming for simplicity that $\mathbb{K}_1 = \mathbb{K}_2 = I$ and considering homogeneous Dirichlet conditions on $\partial\Omega_i \setminus \overline{\Gamma}$. For each subdomain $i = 1, 2$, we denote by $\hat{s}_i \geq 0$ the Fourier transform of the Steklov Poincaré operator with $\hat{s}_i = |k| \coth(|k|(L_i - \delta))$, and we denote by $\widehat{R(f_i)}$ the Fourier transform of $\gamma_i^n \nabla(\Delta^{-1} f_i)$ with Δ^{-1} defined on Ω_i with homogeneous Dirichlet boundary conditions on $\partial\Omega_i$. In this section, the superscripts red and ex are used for the reduced and exact model solutions. We assume in the following that δ is such that $\delta \leq L = \min(\frac{L_1}{2}, \frac{L_2}{2})$.

3.1 *Error Estimates on the Traces $\gamma_i p_i$ and $\gamma_i^{\mathbf{n}}\mathbf{u}_i$*

For the exact and reduced solutions we have, with $\bullet =$ red, ex,

$$
\hat{u}_1^\bullet = -\hat{s}_1 \hat{p}_1^\bullet - \widehat{R(f_1)}, \quad \hat{u}_2^\bullet = \hat{s}_2 \hat{p}_2^\bullet - \widehat{R(f_2)}.
$$

We want to provide an error estimate for the errors on the traces

$$
\hat{e}_{p_i} = \hat{p}_i^{ex} - \hat{p}_i^{red}, \qquad \hat{e}_{u_i} = \hat{u}_i^{ex} - \hat{u}_i^{red},
$$

for $i = 1, 2$ which are linked by the relations $\hat{e}_{u_i} = (-1)^i \hat{s}_i \hat{e}_{p_i}$.

From the exact and reduced transmission conditions (10) and (12), setting

$$
E_i = H_i^{ex} - H_i^{red},
$$

and

$$D(k)=\left(\frac{1}{\mu|k|\hat{s}_1}+H_1^{red}\right)\left(\frac{1}{\mu|k|\hat{s}_2}+H_2^{red}\right)+\left(\frac{1}{\mu|k|\hat{s}_2}+H_1^{red}\right)\left(\frac{1}{\mu|k|\hat{s}_1}+H_2^{red}\right),$$

we obtain that

$$\hat{e}_{u_1}=\frac{-(\frac{1}{\mu|k|\hat{s}_2}+H_2^{red})E_1(\hat{u}_1^{ex}+\hat{u}_2^{ex})-(\frac{1}{\mu|k|\hat{s}_2}+H_1^{red})E_2(\hat{u}_1^{ex}-\hat{u}_2^{ex})}{D(k)},$$
$$\hat{e}_{u_2}=\frac{-(\frac{1}{\mu|k|\hat{s}_1}+H_2^{red})E_1(\hat{u}_1^{ex}+\hat{u}_2^{ex})+(\frac{1}{\mu|k|\hat{s}_1}+H_1^{red})E_2(\hat{u}_1^{ex}-\hat{u}_2^{ex})}{D(k)}.$$

It remains to estimate $|\hat{e}_{u_i}|$. We can establish the following bounds

$$\frac{|E_2(\xi)|}{\xi}\le C_2,\quad \frac{|E_1(\xi)|}{\xi^5}\le C_1,\quad \forall\xi\ge 0,$$

and

$$\left|\frac{1}{H_2^{ex}(\xi)}\right|\le C_3\xi^3,\quad \frac{1}{H_2^{red}(\xi)}\le C_3\xi^3,\quad k\le\hat{s}_i(k)\le k+\frac{1}{L},\quad \forall\xi,k\ge 0,$$

with $C_1=\frac{1}{45}$, $C_2=\frac{81}{175}$, $C_3=\frac{1}{3}$. We deduce the estimates

$$|\hat{e}_{u_i}|=\left[\mu|k|(|k|+\frac{1}{L})C_1|k|\delta|\hat{u}_1^{ex}+\hat{u}_2^{ex}|+C_2C_3|\hat{u}_1^{ex}-\hat{u}_2^{ex}|\right]|k|^4\delta^4,\tag{19}$$

and

$$|\hat{e}_{u_i}|=\left[\mu|k|(|k|+\frac{1}{L})C_1|\hat{u}_1^{ex}+\hat{u}_2^{ex}|+\frac{1}{\mu|k|}C_2(C_3)^2|k|^2\delta^2|\hat{p}_1^{ex}+\hat{p}_2^{ex}|\right]|k|^5\delta^5.\tag{20}$$

Estimates on $\hat{e}_{p_i}$ are readily deduced from the relations $\hat{e}_{u_i}=(-1)^i\hat{s}_1\hat{e}_{p_i}$. An improved estimate can also be derived on $\hat{e}_{p_1}-\hat{e}_{p_2}$ using the additional bound $|\frac{1}{\hat{s}_1}-\frac{1}{\hat{s}_2}|\le\frac{1}{|k|(L|k|+1)}$:

$$|\hat{e}_{p_1}-\hat{e}_{p_2}|\le\mu\left[2(|k|+\frac{1}{L})C_1|\hat{u}_1^{ex}+\hat{u}_2^{ex}|+\frac{1}{2L}C_2C_3|\hat{u}_1^{ex}-\hat{u}_2^{ex}|\right]|k|^5\delta^5.\tag{21}$$

3.2 Error Estimates on the Fracture Mean Values U, V and P

Let us proceed with the error estimates on the fracture mean values $\hat{V}$, $\hat{U}$ and $\hat{P}$. For the error $\hat{e}_V=\hat{V}^{ex}-\hat{V}^{red}$, we have from (14) the bound

$$|\hat{e}_V| \leq \frac{1}{|k|2\delta}|\hat{e}_{u_1} - \hat{e}_{u_2}|,$$

then, it suffices to apply (19) or (20) providing respectively an $O(\delta^3)$ or an $O(\delta^4)$ error estimate.

Similarly, for the error $\hat{e}_U = \hat{U}^{ex} - \hat{U}^{red}$, we have from (15) the bound

$$|\hat{e}_U| \leq \frac{1}{\mu 2\delta |k|^2}|\hat{e}_{p_1} - \hat{e}_{p_2}|.$$

Then, it suffices to apply (21) providing an $O(\delta^4)$ error estimate.

To estimate the error on the mean pressure, it can be shown that there exists $C_4 = \frac{22}{175}$ such that

$$\frac{|F^{ex}(\xi) - F^{red}(\xi)|}{\xi^4} \leq C_4, \quad \forall \xi \geq 0.$$

Then, we deduce from (16) and the definition of F^{ex} the following error estimate for $\hat{e}_P = \hat{P}^{ex} - \hat{P}^{red}$:

$$|\hat{e}_P| \leq \mu\Big[\Big((1 + \frac{2}{15C_3})|k| + \frac{1}{C_3}|k|^{-1}\delta^{-2}\Big)|\hat{e}_V| + \frac{C_4}{2}|k|^3\delta^2|\hat{V}^{ex}|\Big],$$

of order $O(\delta^2)$.

4 Extension to the General Beaver Joseph Saffman Condition

In the general case, the functions H_i^{ex} and F^{ex} depend on two dimensionless parameters, namely $|k|\delta$ and $C_\alpha = \frac{\mu}{\alpha\delta}$. The extension distinguishes two cases, first $\alpha > 0$ (including the previous case $\alpha = +\infty$ i.e. $C_\alpha = 0$) and second $\alpha = 0$. In the first case, the asymptotic expansions of H_i^{ex} and F^{ex} are done for small values of $|k|\delta$ at given $C_\alpha < +\infty$. This choice permits to recover the proper wall friction term in the V momentum equation (22). We obtain the same model as in (18) with modified coefficients for the fifth equation:

$$\frac{6\frac{\mu}{\delta}}{1 + 3C_\alpha}V - 2\widetilde{\mu}\delta\,\partial_{yy}V + 2\delta\,\partial_y P = 0. \tag{22}$$

The tangential viscosity $\widetilde{\mu} = \Big(1 + \frac{2}{5(3C_\alpha+1)^2}\Big)\mu$ is again corrected compared with the classical model reduction approach for which $\widetilde{\mu} = \mu$. The error estimates are the same as in Sects. 3.1 and 3.2 with constants C_i, $i \in \{1, 2, 3, 4\}$ depending on C_α.

In the second case, for $\alpha = 0$ corresponding to $C_\alpha = +\infty$, the expansions of H_i^{ex} are done w.r.t. small values of $|k|\delta$ and $F^{ex} = F^{red} = 0$. We obtain the following reduced model:

$$\begin{aligned}
&\operatorname{div}(\mathbf{u}_i) = f_i && \text{on } \Omega_i,\ i = 1, 2,\\
&\mathbf{u}_i = -\mathbb{K}_i \nabla p_i && \text{on } \Omega_i,\ i = 1, 2,\\
&2\delta\, \partial_y V = \gamma_1^{\mathbf{n}}\mathbf{u}_1 - \gamma_2^{\mathbf{n}}\mathbf{u}_2 && \text{on } \mathbb{R},\\
&-2\mu\delta\, \partial_{yy} U = \gamma_1 p_1 - \gamma_2 p_2 && \text{on } \mathbb{R},\\
&-\mu\, \partial_{yy} V + \partial_y P = 0 && \text{on } \mathbb{R},\\
&U = \frac{\gamma_1^{\mathbf{n}}\mathbf{u}_1 + \gamma_2^{\mathbf{n}}\mathbf{u}_2}{2} && \text{on } \mathbb{R},\\
&\frac{\mu}{\delta}\Big(1 - \frac{\delta^2}{6}\partial_{yy}\Big)\Big(\gamma_1^{n}\mathbf{u}_1 - \gamma_2^{\mathbf{n}}\mathbf{u}_2\Big) = \gamma_1 p_1 + \gamma_2 p_2 - 2P && \text{on } \mathbb{R},
\end{aligned} \tag{23}$$

which differs in the last equation from the model obtained by the classical model reduction approach [3] providing the equation $\frac{\mu}{\delta}\Big(\gamma_1^{\mathbf{n}}\mathbf{u}_1 - \gamma_2^{\mathbf{n}}\mathbf{u}_2\Big) = \gamma_1 p_1 + \gamma_2 p_2 - 2P$. The error estimates for the case $\alpha = 0$ differ from the ones of Sects. 3.1 and 3.2. Setting $C_1 = \frac{2}{15}$ and $C_2 = \frac{2}{945}$, we obtain

$$|\hat{e}_{u_i}| \le \epsilon|k|(|k| + \frac{1}{L})|k|^5\Big(C_1|\hat{u}_1^{ex} + \hat{u}_2^{ex}| + C_2|\hat{u}_1^{ex} - \hat{u}_2^{ex}|\Big)\delta^5, \quad i = 1, 2,$$

and

$$\begin{aligned}
|\hat{e}_V| &\le \frac{1}{|k|}\frac{|\hat{e}_{u_1} - \hat{e}_{u_2}|}{2\delta} \le \epsilon(|k| + \frac{1}{L})|k|^5\Big(C_1|\hat{u}_1^{ex} + \hat{u}_2^{ex}| + C_2|\hat{u}_1^{ex} - \hat{u}_2^{ex}|\Big)\delta^4,\\
|\hat{e}_P| &\le \epsilon|k||\hat{e}_V| \le \epsilon^2(|k| + \frac{1}{L})|k|^6\Big(C_1|\hat{u}_1^{ex} + \hat{u}_2^{ex}| + C_2|\hat{u}_1^{ex} - \hat{u}_2^{ex}|\Big)\delta^4,\\
|\hat{e}_U| &\le \frac{1}{\epsilon 2\delta|k|^3}(|\hat{e}_{u_1}| + |\hat{e}_{u_2}|) \le (|k| + \frac{1}{L})|k|^3\Big(C_1|\hat{u}_1^{ex} + \hat{u}_2^{ex}| + C_2|\hat{u}_1^{ex} - \hat{u}_2^{ex}|\Big)\delta^4.
\end{aligned}$$

5 Conclusions

This work extends the dimensional reduction methodology based on Fourier analysis developed in [1] to the case of a Darcy-Stokes matrix fracture coupled model. This analysis leads to correction terms which cannot be a priori obtained by the classical technique based on averaging along the fracture width combined with profile assumptions on the velocities and pressure in the fracture [2, 3]. More precisely, the new mixed-dimensional model exhibits a correction of the tangential viscosity along the fracture in the case $\alpha > 0$ and a second order correction term in the second closure equation in the case $\alpha = 0$. These terms play an essential role in the error estimates between the equip and mixed-dimensional models derived by the Fourier analysis. Numerical tests are ongoing in order to assess numerically these results.

References

1. Gander, M., J., Hennicker, J., Masson, R., Modeling and analysis of the coupling in discrete fracture matrix models. SIAM J. Numer. Anal. **59**(1), 195–218 (2021)
2. Rybak, I., Metzger, S.: A dimensionally reduced Stokes-Darcy model for fluid flow in fractured porous media. Appl. Math. Comput. **384** (2020)
3. Lesinigo, M., D'Angelo, C., Quarteroni, A.: A multiscale Darcy-Brinkman model for fluid flow in fractured porous media. Numer. Math. **117**(4), 717–752 (2011)

Finite Volumes for Simulation of Large Molecules

Martin Heida

Abstract We study a finite volume scheme for simulating the evolution of large molecules within their reduced state space. The finite volume scheme under consideration is the SQRA scheme developed by Lie, Weber and Fackeldey. We study convergence of a more general family of FV schemes in up to 3 dimensions and provide a convergence result for the SQRA-scheme in arbitrary space dimensions.

Keywords Finite Volume · SQRA · Voronoi

1 Smoluchovski Equation in High Dimension

The evolution of a large molecule over time can be modelled using the Smoluchovski equation where the state of the molecule is described by its position in the state space. While the true state space consists of the positions and velocities of all atoms of the molecule, for large molecules we can often identify several critical degrees of freedom that dominate the behavior and the state of the molecule, which can be used to reduce the dimension of the state space. Considering e.g. a critical bond within the molecule, which can vary its angles $(\theta, \phi) \in [0, \pi) \times [0, 2\pi)$. If the molecule has 3 such bonds, this leads to a polygonal subset $\mathbf{Q}$ of a $d = 6$ dimensional state space $\mathbb{X}$. The variable $u(t, \cdot) : \mathbf{Q} \to \mathbb{R}$ will henceforth be indicating the probability distribution to find the molecule in the state $x \in \mathbf{Q}$ at time t and $u_0 = u(0, \cdot)$ is the initial distribution (or initial state $u_0 = \delta_{x_0}$, in case this is known precisely). The evolution of u over time is described by the Smoluchovski equation with mobility κ and chemical potential V

M. Heida (✉)
Weierstrass Institute for Applied Analysis and Stochastics, Berlin 10117, Germany
e-mail: heida@wias-berlin.de

E. Franck et al. (eds.), *Finite Volumes for Complex Applications X—Volume 1, Elliptic and Parabolic Problems*, Springer Proceedings in Mathematics & Statistics 432,
https://doi.org/10.1007/978-3-031-40864-9_25

$$\dot{u} = \nabla \cdot (\kappa \nabla u) + \nabla \cdot (\kappa u \nabla V) \quad \text{on } [0, T] \times \mathbf{Q}$$

Without going into details but referring to [5] we claim that the major point for the understanding of long-term evolution of the molecule is the understanding of the right hand side linear operator, i.e. its eigenvalues and eigenvectors.

From the numerical point of view, this results in the necessity to discretize the following elliptic equation:

$$-\nabla \cdot (\kappa \nabla u) - \nabla \cdot (\kappa u \nabla V) = f \qquad \text{on } \mathbf{Q} \tag{1}$$

and to study the convergence behavior of the discretization. For simplicity we assume in the following that κ, $V \in C^2(\overline{\mathbf{Q}})$, $f \in L^2(\mathbf{Q})$ are real-valued functions.

The assumption $V \in C^2(\overline{\mathbf{Q}})$ implies strict positivity of $\pi := \exp(-V)$. Using a transformation $U = u/\pi$ we find that (1) is equivalent with

$$-\nabla \cdot (\pi \kappa \nabla U) = f. \tag{2}$$

The particular challenges we address are, first, the choice of discretization approach for π, as addressed in [3], and second the issues that arise from high dimensionality of the problem, i.e. the curse of dimensionality, and the issue arising from $V(x) \to +\infty$ as $x \to \partial\mathbf{Q}$ at least for some models, addressed in [4].

1.1 Discretization

Discretizing (2) on an admissible mesh in the sense of Definition 10.1 in Chapter 3 of [1] or in [2] we write $\mathcal{T} = (\mathcal{V}, \mathcal{E}, \mathcal{P})$ for the mesh consisting of convex polytope control volumes $\mathcal{V} := \{\Omega_i, i = 1, \ldots, N\}$ with mass m_i,$(d-1)$-dimensional flat interfaces $\mathcal{E}_{\mathbf{Q}} = \left\{\sigma_{i,j}\right\}$ with measure $m_{i,j}$ and points $\mathcal{P}_{\mathbf{Q}} = \{x_i, i = 1, \ldots, N\}$ which we sometimes call the cell centers. Two cells Ω_i, Ω_j are neighbors if $\sigma_{i,j} := \partial\Omega_i \cap \partial\Omega_j$ has positive measure and we write $i \sim j$. If $i \sim j$, the distance of the cell centers is $h_{i,j} := \left|x_i - x_j\right|$.

In order to formulate discrete Dirichlet conditions, we follow [2] and enrich the mesh with finitely many points $\mathcal{P}_{\partial\mathbf{Q}} = (y_k)_k \subset \partial\mathbf{Q}$ and virtual interfaces $\mathcal{E}_{\partial\mathbf{Q}} = \left\{\sigma_{i,k} \text{ flat} : \exists i \text{ with } \sigma_{i,k} \subset \partial\mathbf{Q} \cap \partial\Omega_i\right\}$ i.e., for every flat segment $\sigma_{i,k} \subset \partial\mathbf{Q} \cap \partial\Omega_i$ we chose $y_k \in \sigma_{i,k}$ such that $(y_k - x_i) \perp \sigma_{i,k}$ and denote $m_{i,k} := \left|\sigma_{i,k}\right|$ with $h_{i,k} := |y_k - x_i|$. We further generalize the notation $i \sim j$ if $\sigma_{i,j} \subset \partial\Omega_i$ or $\sigma_{i,j} \subset \partial\Omega_j$. Then, when summing up over the interfaces in the calculations below, we do not have to distinguish between inner interface of type $\partial\Omega_i \cap \partial\Omega_j$ and outer interfaces of type $\partial\mathbf{Q} \cap \partial\Omega_i$.

We finally denote $\mathcal{P} = \mathcal{P}_{\mathbf{Q}} \cup \mathcal{P}_{\partial\mathbf{Q}}$ and $\mathcal{E} = \mathcal{E}_{\mathbf{Q}} \cup \mathcal{E}_{\partial\mathbf{Q}}$ and write $\sum_{j:\, j\sim i}$ for the sum over all interfaces belonging to Ω_i and $\sum_{j\sim i}$ for the sum over all interfaces $\mathcal{E}$.

Given a family of admissible meshes $\mathcal{T}_h = (\mathcal{V}_h, \mathcal{E}_h, \mathcal{P}_h)$ we denote for $\Omega_i \in \mathcal{V}_h$ the diameter $h_i = \operatorname{diam}\Omega_i$. The family of meshes is called *quasi uniform* if for every

$x_i, x_j \in \mathcal{P}_h$, $i \sim j$, it holds $h_{i,j} < h$ and if there exists $R, r > 0$ independent from $\mathcal{T}_h$ such that the following holds: For every $\Omega_i \in \mathcal{V}_h$ there exists $x \in \Omega_i$ such that $\mathbb{B}_{rh_i}(x) \subset \Omega_i \subset \mathbb{B}_{Rh_i}(x)$.

We make the following proposal for a discretization of (2)

$$\forall x_i \in \mathcal{P}_{\mathbf{Q}} \qquad - \sum_{j:\, j\sim i} \frac{m_{i,j}}{h_{i,j}} S_{i,j} \left(U_{\mathcal{T},j} - U_{\mathcal{T},i} \right) = m_i f_{\mathcal{T},i}, \tag{3}$$

where $f_{\mathcal{T},i} = ⨍_{\Omega_i} f$ is the average of f over Ω_i and $S_{i,j} = S_{\alpha,\beta}\left(\pi_i, \pi_j\right)$ is a Stolarsky mean of π_i and π_j [6], $\pi_i = \mathrm{e}^{-V_i}$, $V_i = V(x_i)$ resp. $V_i = V(y_i)$ and

$$S_{\alpha,\beta}(x, y) = \left(\frac{\beta\,(x^\alpha - y^\alpha)}{\alpha\left(x^\beta - y^\beta\right)} \right)^{\frac{1}{\alpha-\beta}}, \qquad \alpha \neq 0,\ \beta \neq 0,\ \alpha \neq \beta,\ x \neq y \tag{4}$$

Stolarsky means can be extended to the critical points $\alpha = 0,\ \beta = 0,\ \alpha = \beta,\ x = y$ in a continuous way and generalize the logarithmic mean and other means. Interestingly, for a choice $\alpha = 0$ $\beta = -1$ one obtains the Scharfetter–Gummel scheme with $S_{0,-1}(x, y) = xy(x - y)^{-1} \log \frac{x}{y}$. While we do not want to go into detail on this aspect, we mention that $\alpha = 1$, $\beta = -1$ yields $S_{1,-1}(x, y) = \sqrt{xy}$, which is the SQRA scheme and refer for more information on motivation and background to [3].

From a discrete solution $U_{\mathcal{T}}$ one can obtain a discrete $u_{\mathcal{T}}$ reversing the above transformation $U = u/\pi$. One obtains that $u_{\mathcal{T},i} := U_{\mathcal{T},i}\pi_i$ solves the discrete Smoluchovski

$$\forall x_i \in \mathcal{P}_{\mathbf{Q}} \qquad - \sum_{j:\, j\sim i} \frac{m_{i,j}}{h_{i,j}} S_{i,j} \left(\frac{u_{\mathcal{T},j}}{\pi_j} - \frac{u_{\mathcal{T},i}}{\pi_i} \right) = m_i f_{\mathcal{T},i}, \tag{5}$$

In what follows we will provide convergence results for the above discretizations in low dimensions, i.e. $d \leq 3$ and in high dimensions for (3) only.

1.2 Results and Challenges

Our results are centered around two different questions that arise from the convergence analysis of (3) and (5) in high dimensions: The first results in Sect. 2 deal with the convergence of (3) and (5) in low dimensions up to $d = 3$. We will see that all schemes converge with the same rate for U but that there is a different convergence behavior in u: The classical Scharfetter-Gummel scheme has a better convergence behavior than the other schemes for high gradients of V. From the analytical point of view, it is interesting that for any choice of the Stolarsky mean, the rate of convergence is not worse than the consistency of the mesh for the ordinary Laplace operator, i.e. $\kappa = \pi = 1$.

The results of Sect. 3 are centered around the convergence of a general finite volume scheme of type (3) in high dimensions when the resolution of the underlying grid is not homogeneous: In particular, we assume that the expected solution is almost constant in some regions, where the resolution is chosen rough, while the resolution is fine in regions of strong oscillations of the solution or the coefficients κ and π. We will see that this can lead to good results by simultaneously bypassing the curse of dimensionality to some extend. Furthermore, we deal with the case that the elliptic parameter degenerates locally close to the boundary. This scenario is relevant in chemistry as the potential V might tend to $+\infty$ in some regions of the state space.

The mathematical challenge in the second case is that one cannot rely on the "classical" pointwise evaluation of the limit function, but one has to compare the discrete solution with a locally averaged continuous solution. In particular, Taylor arguments have to be carried out in an averaged sense and one needs to be very careful that averaged lower order terms really cancel each other out. Furthermore, also the proof of the Poincaré inequality has to rely on dimensionless averaging arguments.

2 Convergence Results Based on Consistency, [3]

In this section, we assume $\kappa = 1$ for simplicity of notation, but mention that the results in [3] hold more general. We then denote

$$L^2(\mathcal{P}) := \left\{U : \mathcal{P}_{\mathbf{Q}} \to \mathbb{R}\right\} \qquad H_{\mathcal{T}} := \left\{U : \mathcal{P} \to \mathbb{R} \mid U|_{\mathcal{P}_{\partial\mathbf{Q}}} \equiv 0\right\}$$

with the embedding $H_{\mathcal{T}} \hookrightarrow L^2(\mathcal{P})$ and for $\tilde{v} \in L^2(\mathcal{P})$, $v \in H_{:\mathcal{T}}$ we introduce

$$\|v\|^2_{H_{\mathcal{T}}} := \sum_{i\sim j} \frac{m_{i,j}}{h_{i,j}} \left(v_j - v_i\right)^2 , \quad \|\tilde{v}\|^2_{L^2(\mathcal{P})} := \sum_{\Omega_i} m_i \tilde{v}_i^2 . \tag{6}$$

Definition 1 (*Inf-sup stability*) Let $\mathcal{T}_h = (\mathcal{V}_h, \mathcal{E}_h, \mathcal{P}_h)$ be a quasi uniform family of admissible meshes. A family of bilinear forms a_h on $H_{\mathcal{T}_h}$ is called *uniformly inf-sup stable* with respect to two norms $\|\cdot\|_{h,1}$, $\|\cdot\|_{h,2}$ if there exists $\gamma > 0$ (independent from h) such that

$$\forall u \in H_{\mathcal{T}_h} : \quad \gamma \|u\|_{h,1} \leq \sup_{v \in H_{\mathcal{T}_h}} \frac{a_h(u, v)}{\|v\|_{h,2}} .$$

We write $(\mathcal{R}_h u)_i := (\mathcal{R}_{\mathcal{T}_h} u)_i := u(x_i)$ on Ω_i. For a continuous and coercive bilinear form $a : H^1_0(\mathbf{Q}) \times H^1_0(\mathbf{Q}) \to \mathbb{R}$, the associated linear operator $A : H^2(\mathbf{Q}) \to L^2(\mathbf{Q})$ is defined by

$$\forall u \in H^2(\mathbf{Q}) \cap H^1_0(\mathbf{Q}),\ v \in H^1_0(\mathbf{Q}) : \quad a(u, v) = \int_{\mathbf{Q}} v\, Au . \tag{7}$$

Definition 2 (*Consistency*) Let $a : H_0^1(\mathbf{Q}) \times H_0^1(\mathbf{Q}) \to \mathbb{R}$ be bilinear and continuous with linear operator A such that (7) holds and let $\mathcal{T}_h = (\mathcal{V}_h, \mathcal{E}_h, \mathcal{P}_h)$ be a family of admissible meshes with $a_h : H_{\mathcal{T}_h} \times H_{\mathcal{T}_h} \to \mathbb{R}$ continuous bilinear forms. The *variational consistency error* of a_h in $u \in H^2(\mathbf{Q}) \cap H_0^1(\mathbf{Q})$ is the linear form $E_h(u;\cdot) : H_{\mathcal{T}_h} \to \mathbb{R}$ where

$$\forall v \in H_{\mathcal{T}_h} : \quad E_h(u; v) := \sum_i v_i \int_{\Omega_i} Au - a_h(\mathcal{R}_h u, v) . \tag{8}$$

We say *consistency* holds for $\|\cdot\|_{h,2}$ on $H_{\mathcal{T}_h}$ and $u \in H^2(\mathbf{Q}) \cap H_0^1(\mathbf{Q})$ if

$$\|E_h(u;\cdot)\|_{h,2,*} := \sup_{v \in H_{\mathcal{T}_h}\setminus\{0\}} \frac{|E_h(u; v)|}{\|v\|_{h,2}} \to 0 \quad \text{as} \quad h \to 0 .$$

A special role is played by

$$a_{\mathrm{D}}(u, v) = \int_{\mathbf{Q}} \nabla u \cdot \nabla v , \qquad a_{h,\mathrm{D}}(u, v) = \sum_{i \sim j} \frac{m_{i,j}}{h_{i,j}} \left(u_j - u_i\right)\left(v_j - v_i\right) ,$$

with the corresponding consistency error $E_{h,\mathrm{D}}$. This is the underlying concept of the following definition:

Definition 3 (φ-*consistency*) Let $\mathcal{T}_h = (\mathcal{V}_h, \mathcal{E}_h, \mathcal{P}_h)$ be a quasi uniform family of admissible meshes. We say that $\mathcal{T}_h$ is φ-*consistent* for a continuous monotone increasing φ with $\varphi(0) = 0$ if for every $u \in H^2(\mathbf{Q}) \cap H_0^1(\mathbf{Q})$ there exists $C \geq 0$ such that for every $h > 0$

$$\left\| E_{h,\mathrm{D}}(u;\cdot) \right\|_{H^*_{\mathcal{T}_h}} \leq C \, \|u\|_{H^2} \, \varphi(h) .$$

Our main results are formulated in terms of φ-consistency as follows:

Theorem 1 ([3], Theorem 1.4) *Let $d \leq 3$ and $\mathcal{T}_h = (\mathcal{V}_h, \mathcal{E}_h, \mathcal{P}_h)$ be a quasi uniform family of admissible meshes and let the above assumptions on κ, V and f hold. Moreover, let $\mathcal{T}_h$ be φ-consistent (Definition 3). If $U \in H^2(\mathbf{Q}) \cap H_0^1(\mathbf{Q})$ is the solution of (2) and $U_{\mathcal{T}_h}$ the solution of (3) with discrete homogeneous Dirichlet boundary conditions then*

$$\left\| U_{\mathcal{T}_h} - \mathcal{R}_{\mathcal{T}_h} U \right\|^2_{H_{\mathcal{T}_h}} \leq C_1 \, \|\pi\|^2_\infty \, \varphi(h)^2 + C_2 h^k ,$$

where $k = 2$ in general and $k = 4$ if the grid is cubic or $d = 1$. Here, C_1 and C_2 depend only on d and $\mathbf{Q}$, r and R.

Theorem 2 ([3], Theorem 1.5) *Let $d \leq 3$ and $\mathcal{T}_h = (\mathcal{V}_h, \mathcal{E}_h, \mathcal{P}_h)$ be a quasi uniform family of admissible meshes and let the above assumptions on κ, V and f hold. Moreover, let $\mathcal{T}_h$ be φ-consistent (Definition 3). If $u \in H^2(\mathbf{Q}) \cap H_0^1(\mathbf{Q})$ is the solution* *of (1) and $u_{\mathcal{T}_h}$ the solution of (5) with discrete homogeneous Dirichlet boundary conditions then*

$$\left\| u_{\mathcal{T}_h} - \mathcal{R}_{\mathcal{T}_h} u \right\|_{H_{\mathcal{T}_h}}^2 \leq C_1 \left(\|u\|_{H^2}^2 + \|u\|_\infty^2 \|V\|_{H^2}^2 \right) \varphi(h)^2 + C_2 h^k \,,$$

where $k = 2$ in general and $k = 4$ if $\alpha + \beta = -1$ and where C_1 depends on $\mathbf{Q}$, d, r and R and C_2 additionally depends on $\|V\|_{C^2}$ and $\|u\|_{H^2}$.

On cubic grids, the above estimates further simplify.

Theorem 3 ([3], Theorem 1.7) *Let $d \leq 3$ and $\mathcal{T}_h = (\mathcal{V}_h, \mathcal{E}_h, \mathcal{P}_h)$ be a sequence of cubic grids $h\mathbb{Z}^d$ and let the above assumptions on κ, V and f hold. If $u \in H^2(\mathbf{Q}) \cap H_0^1(\mathbf{Q})$ is the solution of (1) and $u_{\mathcal{T}_h}$ the solution of (5) with discrete homogeneous Dirichlet boundary conditions then*

$$\left\| u_{\mathcal{T}_h} - \mathcal{R}_{\mathcal{T}_h} u \right\|_{H_{\mathcal{T}_h}}^2 \leq C h^k \,,$$

where $k = 2$ in general and $k = 4$ if $\alpha + \beta = -1$ and where C depends on on $\mathbf{Q}$, d, $\|V\|_{C^2}$ and $\|u\|_{H^2}$.

3 Finite Volume in High Dimension, [4]

We will now focus on (2) with $\kappa = 1$. When we speak of periodic boundary conditions below, we assume that $\mathbf{Q}$ is a cube. We further assume $\pi \in C^2(\mathbf{Q})$. **Since molecules could face non self-penetrating conditions, the event** $V(x) \to +\infty$ **as** $x \to x_0 \in \mathbf{Q}$, **i.e.** $\pi(x_0) = 0$ **is a plausible scenario.** However, we will be very general on our assumptions on π. For every $\Omega_i \in \mathcal{T}$ we take a value π_i and for every $\sigma \in \mathcal{E}_\mathbf{Q}$ we take a value π_σ. It may or may not hold for $\sigma = \sigma_{ij}$ that $\pi_{ij} = \pi_{\sigma_{ij}} = S_{\alpha,\beta}(\pi_i, \pi_j)$ but we always assume that the discretization is such that $\pi_i, \pi_j, \pi_{ij} > 0$ for every $i \sim k$. Finally, for every cell Ω_i we assume that there exist positive constants $R_i > r_i$ such that

$$\mathbb{B}_{r_i}(x_i) \subset \Omega_i \subset \mathbb{B}_{R_i}(x_i) \,.$$

Since we are in high dimension and want to break the curse of dimensionality by using a high resolution (i.e. small R_i) only in a region as small as possible, we will replace the typically used upper bound for R_i by a distribution of R_i.

In what follows, we write $H_0^2(\mathbf{Q}) := H^2(\mathbf{Q}) \cap H_0^1(\mathbf{Q})$ as well as $H_{per}^2(\mathbf{Q})$ for periodic $H^2(\mathbf{Q})$ functions with mean value 0 and

$$H_{(0)}^2(\mathbf{Q}) := \left\{ U \in H^2(\mathbf{Q}) \mid \int_\mathbf{Q} U = 0,\ \partial_\nu U = 0 \text{ on } \partial\mathbf{Q} \right\} \,.$$

These spaces clearly correspond to homogeneous Dirichlet boundary conditions (BC), periodic or homogeneous Neumann boundary conditions.

In what follows, we write $\mathcal{E}_i = \{\sigma_{ij} : i \sim j\}$ and for $\sigma = \sigma_{ij} \in \mathcal{E}_i$ we write $\partial_{i,\sigma_{ij}} = \frac{1}{h_{ij}}(U_j - U_i)$. If $\sigma \subset \partial\mathbf{Q} \cap \partial\Omega_i$ exists, we write $\mathcal{E}_{i,\partial}$ for the set of all such piecewise flat subsets σ and include $\mathcal{E}_{i,\partial}$ into $\mathcal{E}_i$ and write $\partial_{i,\sigma}$ accordingly.

We then define discrete spaces incorporating discrete Dirichlet, Neumann (Neu) and periodic boundary conditions (DBC) as follows:

- Dirichlet: $H_{\mathcal{T},0} := \{U : \mathcal{P} \to \mathbb{R} \mid \forall\sigma \in \mathcal{E}_\partial\ U_\sigma = 0\}$
- Neu: $H_{\mathcal{T},(0)} := \left\{U : \mathcal{P} \to \mathbb{R} \mid \forall i,\ \sigma \in \mathcal{E}_{i,\partial} : \partial_{K,\sigma} U = 0\,,\ \sum_K m_k U_K = 0\right\}$
- Periodic: we periodize the discretization, consider discrete functions on the full space and require identical values on "periodically shifted" cells. The corresponding space will be called $H_{\mathcal{T},per}$.

In the following, we always match discrete with the corresponding continuous BC. When there is no need to distinguish between the cases, we simply write $H^2_{\mathrm{BC}}(\mathbf{Q})$ and $H_{\mathcal{T},\mathrm{BC}}$ and use the index BC accordingly throughout this work. We study the discrete equation (9) i.e.,

$$\forall i : \quad \sum_{\sigma\in\mathcal{E}_i} m_\sigma \pi_\sigma \partial_{i,\sigma} U_{\mathcal{T}} = m_i f_i\,, \tag{9}$$

in either one of the spaces $H_{\mathcal{T},0}$, $H_{\mathcal{T},(0)}$ or $H_{\mathcal{T},per}$ and with the additional condition $\int_{\mathbf{Q}} \pi U = 0$ in case of Neumann or periodic boundary conditions (BC) i.e. $\sum_i m_i U_{\mathcal{T},i} = 0$.

Defining $L^2(\mathcal{T}) := \left\{v \,|\, \mathcal{P}_{\mathbf{Q}} \to \mathbb{R}\right\}$ and

$$\|v\|^2_{L^2(\mathcal{T})} := \sum_{i\in\mathcal{V}} m_i v_i^2\,, \qquad \|v\|^2_{H_{\mathcal{T},\pi}} := \sum_{\sigma\in\mathcal{E}} m_\sigma h_\sigma \pi_\sigma \left|\partial_\sigma v\right|^2\,, \tag{10}$$

as well as the pair of operators

$$\tilde{\mathcal{R}}_{\mathcal{T}} : L^2(\mathbf{Q}) \to L^2(\mathcal{T})\,, \left(\tilde{\mathcal{R}}_{\mathcal{T}} U\right)_i := \fint_{\mathbb{B}_{r_i}(x_i)} U\,, \tag{11}$$

$$\mathcal{R}^*_{\mathcal{T}} : L^2(\mathcal{T}) \to L^2(\mathbf{Q})\,, \left(\mathcal{R}^*_{\mathcal{T}} U\right)(x) := U_i \text{ if } x \in \Omega_i \tag{12}$$

We extend $\tilde{\mathcal{R}}_{\mathcal{T}}$ to account for discrete Dirichlet BC by $\left(\mathcal{R}_{\mathcal{T},0} U\right)_i := \left(\tilde{\mathcal{R}}_{\mathcal{T}} U\right)_i$ and

$$\forall\sigma \in \mathcal{E}_\partial : \quad \left(\mathcal{R}_{\mathcal{T},0} U\right)_\sigma := 0\,, \tag{13}$$

and for Neumann BC by $\mathcal{R}_{\mathcal{T},(0)} U := \tilde{\mathcal{R}}_{\mathcal{T}} U - \left(\sum_i m_i \left(\tilde{\mathcal{R}}_{\mathcal{T}_h} U\right)_i\right)$ and

$$\forall\sigma\in\mathcal{E}_\partial : \quad \left(\mathcal{R}_{\mathcal{T},(0)}U\right)_\sigma := (\mathcal{R}_\mathcal{T}U)_K , \quad K\in\mathcal{V}_\sigma . \tag{14}$$

For periodic BC, we set $\mathcal{R}_{\mathcal{T},per}U := \tilde{\mathcal{R}}_\mathcal{T}U - \left(\sum_K m_K\left(\tilde{\mathcal{R}}_{\mathcal{T}_h}U\right)_K\right)$ and find the general relation $\mathcal{R}_{\mathcal{T},\mathrm{BC}} : H^2_{\mathrm{BC}}(\mathbf{Q}) \to H_{\mathcal{T},\mathrm{BC}}$.

Theorem 4 ([4] Theorem 2.5) *Given a polygonal bounded domain $\mathbf{Q}\subset\mathbb{R}^d$ and $U\in H^2(\mathbf{Q})$ a solution to (2) with $f\in L^2(\mathbf{Q})$ satisfying the boundary conditions* BC *then for every admissible mesh $\mathcal{T}$ it holds: there exists a unique solution $U_\mathcal{T}$ to (9) for $f_\mathcal{T}$ given by (9) satisfying the discrete boundary conditions* BC. *Furthermore*

$$\left\|U_\mathcal{T}-\mathcal{R}_{\mathcal{T},\mathrm{BC}}U\right\|_{H_{\mathcal{T},\pi}} \le \left(I_{1,\mathcal{T}}(U)+I_{2,\mathcal{T}}(U)\right) , \tag{15}$$

$$I_{1,\mathcal{T}}(U) = \left(\sum_{\sigma\in\mathcal{E}} h_\sigma m_\sigma \pi_\sigma^{-1}\left(\fint_\sigma |\pi-\pi_\sigma|\,|\nabla U|\right)^2\right)^{\frac{1}{2}} ,$$

$$I_{2,\mathcal{T}}(U) = \left(\sum_{\sigma\in\mathcal{E}} m_\sigma\pi_\sigma h_\sigma\left(\fint_\sigma \nabla U\cdot\nu_{\sigma,K}-\partial_{\sigma,K}\mathcal{R}_\mathcal{T}U\right)^2\right)^{\frac{1}{2}} .$$

Furthermore, there exists a constant $C>0$ depending only on d such that for every $U\in H^2(\mathbf{Q})\cap H^1_0(\mathbf{Q})$ the following holds:

$$\left|I_{1,\mathcal{T}}(U)\right|^2 \le C\left(\sum_i \frac{R_i^3}{r_i^3}R_i^2\left\|\sqrt{\pi}\nabla U\right\|^2_{H^1(\Omega_i)}\|\nabla\pi\|^2_{L^\infty(\Omega_i)}\sum_{\sigma\in\mathcal{E}_i}\fint_\sigma\frac{1}{\pi\kappa_\sigma}\right) , \tag{16}$$

$$\left|I_{1,\mathcal{T}}(U)\right|^2 \le C\left(\sum_i \frac{R_i^3}{r_i^3}R_i^2\left\|\nabla U\right\|^2_{H^1(\Omega_i)}\|\nabla\pi\|^2_{L^\infty(\Omega_i)}\sum_{\sigma\in\mathcal{E}_i}\frac{1}{\pi_\sigma}\right) , \tag{17}$$

$$\left|I_{2,\mathcal{T}}(U)\right|^2 \le C\left(\sum_i R_i^2\left(\frac{R_i}{r_i}\right)^{d+1}\left\|\nabla^2 U\right\|^2_{L^2(\Omega_i)}\sum_{\sigma\in\mathcal{E}_i}\pi_\sigma\right) . \tag{18}$$

Theorem 4 provides only an estimate on the $H_{\mathcal{T},\pi}$-norm while we seek convergence also in $L^2(\mathcal{T})$. For this it is convenient to derive a Poincaré inequality. As the above discussion suggests, we will seek for such an inequality with respect to the weighted norms. In what follows, we assume that $\mathbf{Q}$ has the following structure, even though there are more general possible structures:

Definition 4 Let $\mathbf{Q}$ be simply connected, let $\omega\subset\mathbf{Q}$ be open convex and let $\pi:\overline{\mathbf{Q}}\to\mathbb{R}$ be a piecewise constant function. Let $\omega(\pi,\pi_0):=\{x\in\omega|\,\pi(x)\ge\pi_0\}$. Given $\pi_0\ge\pi_1>0$ we say that π is pseudo monotone on ω w.r.t π_0,π_1 and an open ball $\mathbb{B}\subset\omega(\pi,\pi_0)$ if for every $x\in\omega\backslash\omega(\pi,\pi_0)$ and every $y\in\mathbb{B}$ there exists $z\in\partial\omega(\pi,\pi_0)$

such that $t \mapsto \pi(x + t(z - x))$ is monotone increasing on $[0, 1]$ and if π restricted to the closed convex hull of $\omega(\pi, \pi_0)$ is bigger or equal to π_1.

Definition 5 Using $\mathcal{E}_{\mathcal{T},x,y} := \{\sigma \in \mathcal{E} \mid [x, y] \cap \sigma \neq \emptyset\}$ we define values $\pi_{\mathcal{T}}(x) := \left(\mathcal{R}_{\mathcal{T}}^* \pi_{\mathcal{T}}\right)(x)$ and the following for $x \in \omega_i$ and corresponding $\mathbb{B}_{ij} \subset \omega_i$:

$$a_{\pi,\mathcal{T}}(x) := \min \left\{ \left(\mathcal{R}_{\mathcal{T}}^* \pi_{\mathcal{T}}\right)(x)\,,\ \inf_{y \in \mathbb{B}_{ij}} \inf_{\sigma \in \mathcal{E}_{\mathcal{T},x,y}} \pi_\sigma \right\},$$

$$\tilde{\pi}_{\mathcal{T}}(x) := \begin{cases} \left(\mathcal{R}_{\mathcal{T}}^* \pi_{\mathcal{T}}\right)(x) & \text{if } \left(\mathcal{R}_{\mathcal{T}}^* \pi_{\mathcal{T}}\right)(x) \geq \pi_0 \text{ and } a_{\pi,\mathcal{T}}(x) \geq \pi_1 \\ a_{\pi,\mathcal{T}}(x) & \text{else} \end{cases}.$$

Next, we introduce the notation $\tilde{\pi}_{\mathcal{T},K} := m_K^{-1} \int_K \tilde{\pi}_{\mathcal{T}}$. Based on this we write for $U \in L^2(\mathcal{T})$:

$$\overline{\pi}^{\mathcal{V}} := \int_{\mathbf{Q}} \tilde{\pi}_{\mathcal{T}}(x)\,, \qquad \overline{U}^{\tilde{\pi}} := \frac{1}{\overline{\pi}^{\mathcal{V}}} \int_{\mathbf{Q}} \tilde{\pi}_{\mathcal{T}} \mathcal{R}_{\mathcal{T}}^* U\,.$$

Theorem 5 ([4] Theorem 2.14) *Under the above assumptions on* $\mathbf{Q}$ *and* π *and* $\mathcal{T}$ *exists a constant* C *depending only on* d, $\tilde{\mathbf{Q}}$, $C(\mathcal{T}, \pi_0)$, π_0 *and* $\|\pi\|_\infty$ *such that*

$$\sum_K \tilde{\pi}_{\mathcal{T},K} m_K \left(U_K - \overline{U}^{\tilde{\pi}}\right)^2 \leq C \, \|U\|^2_{H_{\mathcal{T},\pi}}\,. \tag{19}$$

References

1. Eymard, R., Gallouët, T., Herbin, R.: Finite volume methods. Handb. Numer. Anal. **7**, 713–1018 (2000)
2. Gallouët, T., Herbin, R., Vignal, M.H.: Error estimates on the approximate finite volume solution of convection diffusion equations with general boundary conditions. SIAM J. Numer. Anal. **37**(6), 1935–1972 (2000)
3. Heida, M., Kantner, M., Stephan, A.: Consistency and convergence for a family of finite volume discretizations of the fokker–planck operator. ESAIM M2AN **55**, 3017–3042 (2021)
4. Heida, M., Sikorski, A., Weber, M.: Consistency and order 1 convergence of cell-centered finite volume discretizations of degenerate elliptic problems in any space dimension. WIAS Preprint **2913** (2022)
5. Lie, H.C., Fackeldey, K., Weber, M.: A square root approximation of transition rates for a markov state model. SIAM J. Matrix Anal. Appl. **34**, 738–756 (2013)
6. Stolarsky, K.B.: Generalizations of the logarithmic mean. Math. Mag. **48**(2), 87–92 (1975)

PDE Models of Virus Replication Coupling 2D Manifold and 3D Volume Effects Evaluated at Realistic Reconstructed Cell Geometries

Markus M. Knodel, Arne Nägel, Eva Herrmann, and Gabriel Wittum

Abstract Virus pandemics and endemics cause enormous pain and costs. Major processes of the intracellular Hepatitis C virus (HCV) viral RNA (vRNA) replication cycle are restricted to the 2D Endoplasmatic Reticulum (ER) manifold, while others take place in the 3D cytosol volume. Modeling the interplay of the major components of the vRNA replication cycle with partial differential equations (PDEs), we establish a system of surface PDEs (sufPDEs) for effects restricted to manifolds coupled to PDEs describing volume effects. Using the diffusion coefficient of viral proteins which we estimated based on experimental data, we discretize the population-dynamics inspired nonlinear diffusion-reaction equation PDE/sufPDE system with the aid of a vertex-centered Finite Volume scheme and evaluate it at unstructured grids representing data based realistic reconstructed cell geometries. We describe the numerical techniques applied and demonstrate the numerical robustness of our simulations. Our framework might contribute to efficient development of antiviral drugs and potent vaccines.

Keywords PDEs · Vertex centered finite volumes · Virus replication

M. M. Knodel (✉)
Simulation in Technology, TechSim, Ölbronn-Dürrn, Germany
e-mail: markus.knodel@techsim.org

A. Nägel
Universität Frankfurt, MSQC, Frankfurt, Germany

E. Herrmann
Universität Frankfurt, IBMM, Frankfurt, Germany

G. Wittum
KAUST, CEMSE, MaS, Thuwal, Saudi Arabia

E. Franck et al. (eds.), *Finite Volumes for Complex Applications X—Volume 1, Elliptic and Parabolic Problems*, Springer Proceedings in Mathematics & Statistics 432,
https://doi.org/10.1007/978-3-031-40864-9_26

1 Introduction

Viruses are a major challenge to animal and human health, global prosperity, economy, political and social systems, as the recent Covid19 pandemics has unveiled. Infection with the Hepatitis C virus (HCV) [2] belongs to current global pandemic virus diseases.

Spatial dependence is a crucial factor in the process all viruses use in order to replicate. In case of HCV, the virus genome - viral RNA (vRNA) - replication takes place within specific compartments called membranous web (MW) [2]. The MWs are derived from altered regions of the the Endoplasmatic Reticulum (ER). The ER is an interconnected intracellular membrane network, its surface is a connected sum of g tori, $g \geq 1$, enclosing the ER lumen. The intracellular vRNA replication cycle is based upon complex vRNA, non structural virus proteins (NSPs) and host interactions between the 3D cell volume space (called cytosol) and the curved 2D ER surface/manifold embedded in the 3D volume. While the ER manifold is embedded inside the volume space of the cell, the volume enclosed by the ER (lumen) is exempt from the replication cycle processes.

In the recent years, we started to develop a framework to allow for fully 3D spatio-temporal resolved virus replication models at an intracellular level for the case of HCV. Our framework aims to mirror in vitro/in vivo experiments by means of fully spatio-temporal resolved diffusion-reaction partial differential equation (PDE) models. From the very beginning, all simulations are performed at geometries which we reconstructed based on experimental data. As major processes of the vRNA cycle are restricted to the curved 2D ER manifold embedded in 3D, our first models focused upon the evaluation of surface PDEs (sufPDEs) and described the interplay of vRNA, NSPs, and a generic host factor. Next we introduced nonlinear population dynamics inspired diffusion and reaction coefficients and different aggregate states for vRNA and NSPs which allowed qualitative realistic modeling, cf. [5] and references given therein. Further, we estimated the diffusion coefficient of the so-called NS5A HCV protein by means of adjusting sufPDE simulation values to experimental time series data, cf. [4].

This study merges effects restricted to the 2D ER manifold with those taking place in the full 3D cytosol volume. We couple sufPDEs with PDEs where manifold attachment and detachment is described by fluxes and reactions ensuring mass conservation. As far as known, we use quantitative reliable parameters motivated by experimentally based values to approach to quantitative reliable simulations.

2 Grids, PDE Models, Discretization and Solvers

2.1 Data Based Unstructured Surface Mesh and Volume Grid

The geometry of our simulations is based on the same cell compartment data as in our former studies [5]: We used fluorescence data z-stacks of stained ER and MW surfaces of liver cells and reconstructed the surfaces of ER and MWs. Applying a GPU implementation of an inertia moment based anisotropic filter and segmenting the data, we created a triangular surface mesh (by means of the marching cube algorithm), post processed with ProMesh [7]. The crucial difference of this study to all our former studies is that up to now, all computations were performed upon the (unstructured) triangular surface grids exclusively. Here, we enclose the surface grids into a rectangular hexahedron which is filled with a full tetrahedral volume (unstructured) mesh with the aid of tetgen [8]. Figure 1 displays screenshots of the surface grid and volume mesh (at base level) opened in parts by a clip plane. Table 1 lists the subdomains of the 3D computational domain $\Omega = \Gamma \cup C \cup W$. Γ is the external boundary of $\Omega \subset \mathbb{R}^3$.

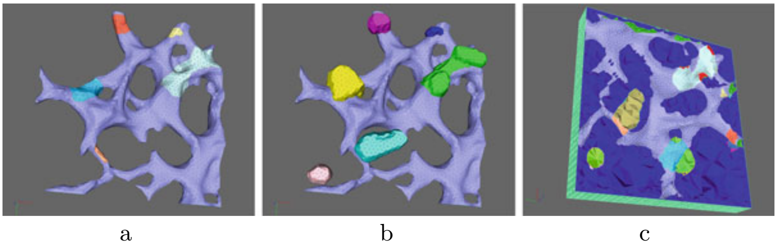


Fig. 1 Triangular surface and tetrahedral volume mesh. Different colors refer to different subdomains. **a** ER surface grid of $\mathcal{M}$ (ribosome subdomains $\mathcal{R}_i$ appear as intersection of ER and - in **a** not displayed—MW surfaces), **b** ER and MW surface grids, **c** Volume mesh of Ω opened with a clip plane. Perspective of **a**, **b** differs to **c**

Table 1 Subdomains of computational domain Ω

Subdomain	Property
2D manifold $\mathcal{E} \cup \mathcal{R} = \mathcal{M} \subset \Gamma$, embedded in 3D	
$\mathcal{E} \subset \mathcal{M}$	Reconstructed ER surfaces except for
$\cup_{i=1}^{7} \mathcal{R}_i = \mathcal{R} \subset \mathcal{M}$	7 ribosomic zones: intersection ER/MW surfaces
3D volume $\mathcal{C} \cup \mathcal{W} \subset \Omega$	
$\mathcal{C} \subset \Omega$	Cytosol (enclosed by the box $\mathcal{B}$ enclosing in fact Ω)
$\cup_{i=1}^{7} \mathcal{W}_i = \mathcal{W} \subset \Omega$	7 MW zones: volume enclosed by MW surfaces

While $\mathcal{M}$ is embedded inside $\mathcal{C} \cup \mathcal{W}$, the volume enclosed by $\mathcal{M}$ (lumen) is excluded from the computational domain and not meshed - Ω is not stellated, but $\mathcal{M}$ is entire part of the "external" boundary of Ω, i.e. $\Gamma = \mathcal{M} \cup \mathcal{B}$.

2.2 Coupled sufPDE/PDE Model of Virus Replication

Those components whose movement is restricted to the ER surface are modeled with the aid of sufPDEs on the 3D embedded 2D curved manifold domain $\mathcal{M}$. Fully 3D "volume" PDEs describe the other components defined in the full volume domain Ω, which we also call "vPDEs" to account for their "volume" property. Technically, the sufPDEs are defined on faces which as well act as faces of the boundary Γ of Ω, where the vPDEs are defined. We model the exchange between surfaces and the volume by means of coupling the vPDEs with the sufPDEs via Neumann boundary conditions of the vPDEs which are mirrored by reactions of the sufPDEs. Exchange may appear where Γ geometrically coincides with $\mathcal{M}$, as $\mathcal{M} \subset \Gamma$. To ensure mass conservation, the values of the local reactions of the sufPDEs on $\mathcal{M}$ mandatory have to match the values of the local flux conditions of the vPDEs on Γ. We describe the interplay of the components listed in Table 2.

All vPDE and sufPDE systems have to be understood as local. For simplicity of notation, we omit the notation of spatial and temporal variables everywhere in the equation system notations. With the convention that a term with a subdomain

Table 2 Concentrations considered in the PDE model

Concentration	Region	Biophysical meaning
Surface concentrations defined at the 3D embedded curved 2D manifold $\mathcal{M}$		
R_R^S	$\mathcal{R}$	Ribosomal bound RNA
P_R^S	$\mathcal{R}$	Viral polyprotein translated at ribosomes
W_C^S	$\mathcal{R}$	Web (NSP) protein cleaved from the polyprotein
N_E^S	$\mathcal{M}$	NS5a NSP cleaved from the polyprotein
R_E^S	$\mathcal{M}$	polymerized free RNA attached to the ER
Volume concentrations defined in the 3D volume Ω		
W_W^V	$\mathcal{W}$	Web (NSP) protein detached from ribosomes to form MWs
N_W^V	$\mathcal{W}$	NS5a NSP detached from ribosomes incorporated into MW
C_W^V	$\mathcal{W}$	Replication complex as combination of detached R_R^S and W_C^S
R_P^V	Ω	Polymerized free RNA moving in the full volume
H^V	Ω	Host factor

subscript indicates that the corresponding term contributes only in this specific subdomain, the sufPDE equation system to be evaluated only on the manifold $\mathcal{M}$ reads:

$$\partial_t R^S_R = \left[\mathrm{div}_{\mathrm{T}}(D_R \,\nabla_{\mathrm{T}}\, R^S_R) - r_1 R^S_R \frac{R^S_R}{R^S_R + p_1} \frac{W^S_C}{W^S_C + p_2} + r_7 R^S_E (r_0 - R^S_R)\right]_{\mathcal{R}} \tag{1}$$

$$\partial_t P^S_R = \left[\mathrm{div}_{\mathrm{T}}(D_P \,\nabla_{\mathrm{T}}\, P^S_R) + r_2 R^S_R \frac{H^V}{H^V + p_3} - r_3 P^S_R\right]_{\mathcal{R}} \tag{2}$$

$$\partial_t W^S_C = \left[\mathrm{div}_{\mathrm{T}}(D_N \,\nabla_{\mathrm{T}}\, W^S_C) + r_3 P^S_R - r_4 W^S_C - v_1 r_1 R^S_R \frac{R^S_R}{R^S_R + p_1} \frac{W^S_C}{W^S_C + p_2}\right]_{\mathcal{R}} \tag{3}$$

$$\partial_t N^S_E = \mathrm{div}_{\mathrm{T}}(D_N \,\nabla_{\mathrm{T}}\, N^S_E) + \left[r_3 P^S_R - r_5 N^S_E W^S_C\right]_{\mathcal{R}} \tag{4}$$

$$\partial_t R^S_E = \mathrm{div}_{\mathrm{T}}\left[D^S_R \left(1 + k_1 \frac{N^S_E}{N^S_E + p_4}\right) \nabla_{\mathrm{T}}\, R^S_E\right] + r_6 R^V_P - \left[r_7 R^S_E (r_0 - R^S_R)\right]_{\mathcal{R}} \tag{5}$$

whereas the vPDE equation system to be evaluated in the full domain Ω reads

$$\partial_t W^V_W = \left[\mathrm{div}(D_N \nabla W^V_W)\right]_{\mathcal{W}} \tag{6}$$

$$\partial_t N^V_W = \left[\mathrm{div}\left(D_N \frac{W^V_W}{W^V_W + p_5} \nabla N^V_W\right)\right]_{\mathcal{W}} \tag{7}$$

$$\partial_t C^V_W = \left[\mathrm{div}\left(D_C \frac{W^V_W}{W^V_W + p_6} \nabla C^V_W\right)\right]_{\mathcal{W}} \tag{8}$$

$$\partial_t R^V_P = \mathrm{div}\left[D^V_R \left(1 + k_2 \frac{N^V_W}{N^V_W + p_7}\right) \nabla R^V_P\right] + \left[r_6 C^V_W \frac{H^V}{H^V + p_8}\right]_{\mathcal{W}} \tag{9}$$

$$\partial_t H^V = \mathrm{div}\left[D_H \left(1 + k_3 \frac{W^V_W}{W^V_W + p_9}\right) \nabla H^V\right] - \left[v_2 r_6 C^V_W \frac{H^V}{H^V + p_8}\right]_{\mathcal{W}} \tag{10}$$

with the Neumann boundary conditions which connect sufPDEs and vPDEs

$$\begin{aligned}
n \cdot \left[D_N \nabla W^V_W\right] &= +r_4 W^S_C & \forall \mathbf{x} \in \mathcal{R} \\
n \cdot \left[D_N \tfrac{W^V_W}{W^V_W + p_5} \nabla N^V_W\right] &= +r_5 N^S_E W^S_C & \forall \mathbf{x} \in \mathcal{R} \\
n \cdot \left[D_C \tfrac{W^V_W}{W^V_W + p_6} \nabla C^V_W\right] &= +r_1 R^S_R \tfrac{R^S_R}{R^S_R + p_1} \tfrac{W^S_C}{W^S_C + p_2} & \forall \mathbf{x} \in \mathcal{R} \\
n \cdot \left[D^V_R \left(1 + k_2 \tfrac{N^V_W}{N^V_W + p_7}\right) \nabla R^V_P\right] &= -r_6 R^V_P & \forall \mathbf{x} \in \mathcal{M}
\end{aligned} \tag{11}$$

where all other boundary conditions are no flux, i.e. Neumann zero conditions. The initial conditions are such that all concentrations are zero everywhere for all components, except for R^S_R, which is a nonzero constant at one specific $\mathcal{R}_i$,

$$R^S_R(t = 0, \mathbf{x}) = \begin{cases} r_0, & \forall \mathbf{x} \in \mathcal{R}_2, \\ 0, & \text{else.} \end{cases} \tag{12}$$

2.3 Technical Framework: Discretization, Solvers and DoF Numbers

Temporal discretization: The complete PDE/sufPDE system is discretized by means of an implicit Euler scheme of first order. The time step size is chosen adaptive and is regulated with the aid of the corresponding number of Newton steps of the respective former time step. So far, all reaction and diffusion terms are incorporated by means of their implicit form, but we plan to switch the incorporation of corresponding signed reaction terms into the explicit form, which, in part, might allow for bigger time step sizes.

Spatial discretization: We perform the spatial discretization of the sufPDEs and vPDEs with the aid of a vertex-centered finite volume (vcFV) scheme, called also "box method" [1, 3], as such a scheme ensures the mass conservation of the transported components on the discrete level. Sketched in a nutshell, a dual grid repartitions the computational domain of the PDE of type $\nabla D \nabla u + ru = 0$, defined in $\Omega \subset \mathbb{R}^d$, by means of non-overlapping control volumes (boxes), $\bar{\Omega} \simeq \bigcup_{i=1}^{n} B_i$, where $B_i \cap B_j = \emptyset$ if $i \neq j$. Each box encloses exactly one vertex of the finite element grid. Choosing $u \in C(\Omega)$ as in the FE case, but $v \in L^2(\Omega)$, the weak form of the PDE leads to $\sum_{i=1}^{n}(\int_{B_i} rud^d x - \int_{\partial B_i} \mathbf{n} \cdot D\nabla u d\sigma) = 0$. The numerical scheme ensures the balance law, namely the fluxes into and from each box are balanced. In case of the manifolds, the method was adopted, i.e. the boxes form lower-dimensional parts of the manifold. At the boundary of the boxes, the normals are constructed tangential to the manifold, as well the differential operators; e.g., the Laplace-Beltrami operator "replaces" the Laplace operator.

Nonlinear solver: A Newton solver is applied to solve the highly nonlinear equation system (1)–(10) leading to a huge systems of linear equations (SLEs).

Linear solver: At each iteration of the Newton solver, the SLEs are solved with a BiCGStab solver which is preconditioned by means of a Geometric Multigrid Solver (GMG).

GMG solver: The GMG applies a V-Cycle with 3 pre- and 3 postsmoothing steps using Symmetric Gauss-Seidel (SGS). As base solver, we use an ILU preconditioned BiCGStab. The GMG solver does not apply any kind of coarsening of the grid. Moreover, the GMG solver applies a global refinement strategy of the already (due to its experimental data based origin) quite fine coarse grid.

DoF numbers: Fig. 4 reports the number of the degrees of freedom (DoFs).

Technical framework: The numerical computations were performed with the UG4 framework [6] at the HLRS Stuttgart Apollo Hawk supercomputer.

3 Simulations

Simulation screenshots: The simulations are running robust and reproduce all major effects of the vRNA cycle, namely also with respect to the interplay of surface and volume effects. The vRNA translates viral protein, which induces MW growth. Inside the MW, the vRNA gets replicated, while the host factor gets consumed, and some newly produced vRNA attach to other ribosomes $\mathcal{R}$, such that the cycle is closed. Figure 2 displays screenshots of different components.

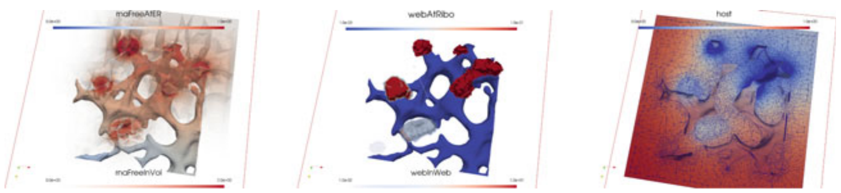
Mass conservation As the vPDE system incorporates several fluxes realized with Neumann boundary conditions which mirror reactions (detachment and attachment) of components from/to the manifold (modelling exchange between manifold and volume), we checked exhaustively the numerical robustness and mass conservation of these processes and found very robust mass conservation properties. Figure 3 displays the mass conservation properties for two scenarios:

I Exclusive production of P^S on $\mathcal{R}$, only nonzero reaction: $r_2 \neq 0$. Comparing with the sums of P^S, W_C^S, W_W^V in case when P^S reacts to W_C^S on $\mathcal{R}$, which itself detaches to W_W^V within $\mathcal{W}$, i.e. $r_2, r_3, r_4 \neq 0$; $r_i = 0$ for $i \neq 2, 3, 4$.

II All reactions nonzero except for $r_7 = 0$, $r_i \neq 0 \forall i \neq 7$; R_E^S at $\mathcal{M}$ does not "switch" into R_R^S at $\mathcal{R}$. Comparing with the case when in addition, R_P^V in Ω does not attach at $\mathcal{M}$ to form R_E^S, i.e. $r_6 = r_7 = 0$, and $r_i \neq 0 \forall i \neq 6, 7$.

Data shown for grid level 2 and $\Delta_t = 40$s (constant time step possible at beginning). The vcFV scheme induced integrals (due to Neumann conditions for the vPDEs) are the same as those for the reaction terms of the sufPDEs. The sum of the compared data agrees very well in both cases, as the relative differences demonstrate. (Note: Mass conservation not dependent on grid level.)

Numerical grid convergence is demonstrated in Fig. 4 by comparing absolute values for selected unknowns c integrated over their domain $\mathcal{D}, \mathcal{I}_L^{\mathcal{D}}(c) = \int_{\mathcal{D}} c d\mathfrak{m}_{\mathcal{D}}$ ($d\mathfrak{m}_{\Omega} = dx^3, d\mathfrak{m}_{\mathcal{M}} = d\sigma$), as the relative differences $R_L^{\mathcal{D}}(c) = |(\mathcal{I}_L^{\mathcal{D}}(c) - \mathcal{I}_{L+1}^{\mathcal{D}}(c))/(\mathcal{I}_L^{\mathcal{D}}(c) + \mathcal{I}_{L+1}^{\mathcal{D}}(c))|$ decrease for increasing grid refinement level L.



(a) (b) (c)

Fig. 2 Simulation screenshots at $t \simeq 1$h (grid level 1). **a** R_E^S concentration on $\mathcal{M}$ merged with R_P^V concentration in complete Ω displayed with the aid of an opacity mapping functionality. **b** W_C^S concentration on $\mathcal{M}$ merged with W_W^V concentration defined in $\mathcal{W}$ (opacity mapping). **c** H^V concentration in Ω, domain disclosed by means of a cut plane (no opacity mapping; edges of grid faces visible)

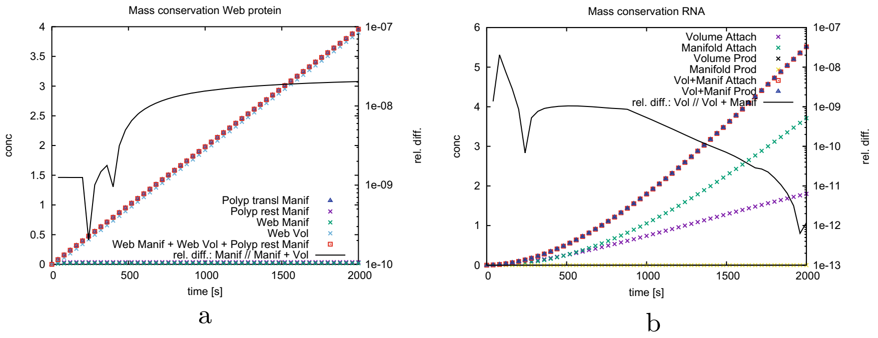


Fig. 3 Mass conservation: absolute values and relative differences. **a** Case I (P_R^S, $r_i = 0 \forall i \neq 2$ versus P_R^S, W_C^S, W_W^V, $r_i = 0 \forall i \neq 2, 3, 4$); **b** Case II ($R_P^V$, $r_i \neq 0 \forall i \neq 6, 7$, versus R_P^V, R_E^S, $r_i \neq 0 \forall i \neq 7$). Results computed at grid level 2 and constant $\Delta_t = 40s$

Level	DoFs	Vols
0	96,370	41,446
1	667,280	331,568
2	4,890,410	2,652,544
3	37,266,790	21,220,352
4	290,579,070	169,762,816

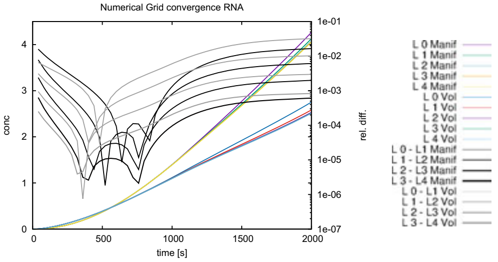


Fig. 4 Left: DoF number at different grid levels and number of elements (tetrahedra). Right: Absolute values $\mathcal{I}_L^{\Omega}(R_P^V)$ (volume) and $\mathcal{I}_L^{\mathcal{M}}(R_E^S)$ (manifold) for $L \in \{0, 1, 2, 3, 4\}$ displayed with "real" colors; relative differences $R_L^{\Omega}(R_P^V)$ (black lines) and $R_L^{\mathcal{M}}(R_E^S)$ (gray lines) with $L \in \{0, 1, 2, 3\}$ demonstrate profound numerical grid convergence

Some challenges for new 2d-3d coupling robustness: To avoid the need for an extremely high number of Newton iterations per time step, the implementation of an adaptive time step size governed by the number of Newton steps was crucial. To simulate several biophysical hours using realistic parameters, the thus established time step size varies drastically, up to five orders of magnitude.

For efficient parallel performance, we apply a hierarchical grid redistribution for spatial grid level refinement higher than level one [6].

The correct assignment of the subdomains after tetrahedralization for all nodes, faces and volumes was solved with the aid of a ProMesh script.

4 Discussion and Conclusions

We have presented a PDE model of the intracellular vRNA cycle which couples volume and surface effects and we have described the numerical techniques we applied. We have demonstrated mass conservation for the coupled surfPDE/vPDE system as well as numerical grid convergence. The results shown in this paper demonstrate the numerical robustness of our simulations, and are compatible with experimental observations. Our study is a building block to establish an advanced quantitative spatially-resolved understanding of virus replication dynamics to unveil the relation of form and function. In the long run, our framework might help facilitate the development of direct antiviral agents and potent vaccines.

Acknowledgements The authors acknowledge the HLRS Stuttgart for the supplied computing time on the HLRS Apollo Hawk supercomputer. MMK thanks Florian Frank (Erlangen University) for stimulating discussions about the numerical methods.

References

1. Bank, R.E., Rose, D.: Some error estimates for the box method. SIAM J. Nu. Anal. **24**, 777–787 (1987)
2. Chatel-Chaix, L., Bartenschlager, R.: Dengue virus and hepatitis c virus-induced replication and assembly compartments: The enemy inside - caught in the web. J. Virol. **88**(11), 5907–11 (2014)
3. Hackbusch, W.: On first and second order box schemes. Computing **41**, 277–296 (1989)
4. Knodel, M.M., Nägel, A., Reiter, S., Rupp, M., Vogel, A., Targett-Adams, P., McLauchlan, J., Herrmann, E., Wittum, G.: Quantitative analysis of hepatitis c ns5a viral protein dynamics on the er surface. Viruses **10**(1), 28 (2018)
5. Knodel, M.M., Reiter, S., Targett-Adams, P., Grillo, A., Herrmann, E., Wittum, G.: Advanced hepatitis c virus replication pde models within a realistic intracellular geometric environment. Int. J. Environ. Res. Public Health **16**(3), 513 (2019)
6. Reiter, S., Vogel, A., Heppner, I., Rupp, M., Wittum, G.: A massively parallel geometric multigrid solver on hierarchically distributed grids. Comp. Vis. Sci. **16**(4), 151–164 (2013)
7. Reiter, S., Wittum, G.: (2017). http://promesh3d.com/
8. Si, H.: Tetgen. A quality tetrahedral mesh generator and 3d delaunay triangulator. WIAS Technical Report 13 (2013). http://www.tetgen.org

Structure-Preserving Schemes for Drift-Diffusion Systems on General Meshes: DDFV Versus HFV

Stella Krell and Julien Moatti

Abstract We made a comparison between a Discrete Duality Finite Volume (DDFV) scheme and a Hybrid Finite Volume (HFV) scheme for a drift-diffusion model with mixed boundary conditions on general meshes. Both schemes are based on a nonlinear discretisation of the convection-diffusion fluxes, which ensures the positivity of the discrete densities. We investigate the behaviours of the schemes on numerical test cases.

Keywords DDFV · HFV · Positivity preserving methods · Discrete entropy/dissipation relation · Long-time behaviour

1 Motivation

We are interested in the numerical discretization of drift-diffusion model. Let Ω be a polygonal connected open bounded subset of $\mathbb{R}^2$, whose boundary $\Gamma = \partial\Omega$ is divided into two parts $\Gamma = \Gamma^D \cup \Gamma^N$ with $\mathrm{m}(\Gamma^D) > 0$. The problem writes:

$$\left\{\begin{array}{cl} \partial_t N - div(\nabla N - N\nabla\phi) = 0 & \text{in } \mathbb{R}_+ \times \Omega, \\ \partial_t P - div(\nabla P + P\nabla\phi) = 0 & \text{in } \mathbb{R}_+ \times \Omega, \\ -\lambda^2 div(\nabla\phi) = C + P - N & \text{in } \mathbb{R}_+ \times \Omega, \\ N = N^D,\ P = P^D \text{ and } \phi = \phi^D & \text{on } \mathbb{R}_+ \times \Gamma^D, \\ (\nabla N - N\nabla\phi)\cdot n = (\nabla P + P\nabla\phi)\cdot n = \nabla\phi\cdot n = 0 & \text{on } \mathbb{R}_+ \times \Gamma^N, \\ N(0,\cdot) = N^{in} \text{ and } P(0,\cdot) = P^{in} & \text{in } \Omega, \end{array}\right. \tag{1}$$

S. Krell (✉)
Université Côte d'Azur, CNRS, Inria, LJAD, Nice, France
e-mail: stella.krell@univ-cotedazur.fr

J. Moatti
Laboratoire Paul Painlevé, Inria, Univ. Lille, CNRS, UMR 8524, F-59000 Lille, France
e-mail: julien.moatti@inria.fr

E. Franck et al. (eds.), *Finite Volumes for Complex Applications X—Volume 1, Elliptic and Parabolic Problems*, Springer Proceedings in Mathematics & Statistics 432,
https://doi.org/10.1007/978-3-031-40864-9_27

where n denotes the unit normal vector to $\partial\Omega$ pointing outward Ω. Regarding the data, (i) the parameter $\lambda > 0$ is the rescaled Debye length of the system, which accounts for the nondimensionalisation (relevant values of this parameter can be very small, inducing some stiff behaviours), (ii) the initial conditions N^{in} and P^{in} belong to $L^{\infty}(\Omega)$ and are positive, (iii) the doping profile C is in $L^{\infty}(\Omega)$, and characterises the semiconductor device used. In the following, we also assume that the boundary conditions are the trace of some H^1 function on Ω, such that the following relation holds:

$$\log(N^D) - \phi^D = \alpha_N \text{ and } \log(P^D) + \phi^D = \alpha_P \text{ on } \Gamma^D, \tag{2}$$

where α_N and α_P are two real constants. It follows that N^D and P^D are positive.

The solution to (1) enjoys some natural physical properties: the densities N and P are positive for all time, and the solution converges exponentionaly fast towards some thermal equilibrium (N^e, P^e, ϕ^e) -which is a stationary solution to (1)—where $N^e = e^{\alpha_N+\phi^e}$, $P^e = e^{\alpha_P-\phi^e}$ and ϕ^e is the solution to the Poisson-Boltzmann equation

$$\begin{cases} -\lambda^2 div(\nabla\phi^e) = C + \exp(\alpha_P - \phi^e) - \exp(\alpha_N + \phi^e) & \text{in } \Omega, \\ \phi^e = \phi^D \text{ on } \Gamma^D \qquad \text{and} \qquad \nabla\phi^e \cdot n = 0 & \text{on } \Gamma^N. \end{cases} \tag{3}$$

Relation (2) is a compatibility condition in order to ensure the existence of the thermal equilibrium (3). When designing numerical schemes for (1), it is crucial to ensure that the scheme preserves these properties at the discrete level. This structure preserving feature is ensured by classical TPFA schemes on admissible orthogonal meshes (see [1]). Unfortunately, these schemes cannot be used on general meshes. Following the ideas introduced in [3], a nonlinear positivity preserving DDFV scheme for Fokker-Planck equations has been introduced in [2]. In the spirit of these works, a nonlinear structure preserving HFV scheme was introduced and partially analysed in [5]. The aim of this paper is to introduce a nonlinear structure preserving DDFV scheme for (1) based on the scheme of [2] and to compare it numerically with the HFV scheme of [5].

2 Descriptions of the Schemes

The schemes used here are based on the same nonlinear strategy, introduced in [3], consisting in the reformulation of the convection-diffusion fluxes:

$$\nabla N - N\nabla\phi = N\nabla\left(\log(N) - \phi\right) \text{ and } \nabla P + P\nabla\phi = P\nabla\left(\log(P) + \phi\right).$$

At the discrete level, both schemes relie on discrete gradients operators to approximate the continuous gradients. The major issue lies in the discretisation of the prefactors P and N, which will be handled by local reconstruction operators.

This discretisation strategy is a way of ensuring (at the theoretical level) the positivity of the discrete densities. We refer to [5, Theorem 1] (HFV scheme for drift-diffusion system) and [2, Theorem 2.1] (DDFV scheme for a single advection-diffusion equation) for proofs of this statement. We also refer the reader to these proofs for more insight about the reconstruction operators. Both schemes are based on a backward Euler discretisation in time. To fix ideas, we will use a constant time step $\Delta t > 0$. For more precise descriptions and statements about the schemes and the meshes, we refer to [2] (DDFV) and [5] (HFV).

Remark 1 (*Generalisation to anisotropic models*) In this paper, we consider isotropic convection-diffusion equations for the charges carriers for the sake of brevity. One could add anisotropic diffusion tensors and consider the framework described in [5].

Both schemes rely on a spatial discretisation (or mesh) of the domain Ω. The (primal interior) mesh $\mathfrak{M}$ is a partition of Ω in polygonal control volumes (or cells). We let $\partial\mathfrak{M}$ be the set of boundary edges, seen either as degenerate control volumes (DDFV framework) or as edges (HFV framework). The primal mesh $\overline{\mathfrak{M}}$ is defined as the reunion of $\mathfrak{M}$ and $\partial\mathfrak{M}$. Given a cell $K \in \overline{\mathfrak{M}}$, we fix a point $x_K \in K$, called the center of K. For all neighboring primal cells K and L, we assume that $\partial K \cap \partial L$ is a segment, corresponding to an internal edge of the mesh $\mathfrak{M}$, denoted by $\sigma = K|L$ and we let $\mathcal{E}_{int}$ be the set of such edges. We denotes by $\mathcal{E} = \mathcal{E}_{int} \cup \partial\mathfrak{M}$ the set of all (internal and exterior) edges of the mesh, and define $\mathcal{E}_K$ the set of edges of the cell $K \in \mathfrak{M}$. For any $K \in \mathfrak{M}$ and $\sigma \in \mathcal{E}_K$, we define $\mathbf{n}_{\sigma K}$ as the unit normal to σ outward K. Given any measurable $X \subset \mathbb{R}^2$, we denote by m_X the measure of the object X.

2.1 The DDFV Scheme

In order to define the DDFV scheme, we need to introduce two other meshes: the dual mesh denoted $\overline{\mathfrak{M}^*}$ and the diamond mesh denoted $\mathfrak{D}$ (see [2] for more details). The dual mesh $\overline{\mathfrak{M}^*}$ is also composed of interior dual mesh $\mathfrak{M}^*$ (corresponding of cells around vertex in Ω) and of boundary dual mesh $\partial\mathfrak{M}^*$ (corresponding of cells around vertex on $\partial\Omega$). 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. 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 and x_{K^*}. We define the set $\mathcal{E}^*_{int}$ of internal edges of the dual mesh similarly as $\mathcal{E}_{int}$. We denote by $\mathbf{n}_{\sigma^* K^*}$ the unit normal to σ^* outward K^*. For each couple $(\sigma, \sigma^*) \in \mathcal{E} \times \mathcal{E}^*_{int}$ such that $\sigma = [x_{K^*}, x_{L^*}]$ and $\sigma^* = K^*|L^*$, 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 (Fig. 1). Finally, the DDFV mesh is made of $\mathcal{T} = (\overline{\mathfrak{M}}, \overline{\mathfrak{M}^*})$

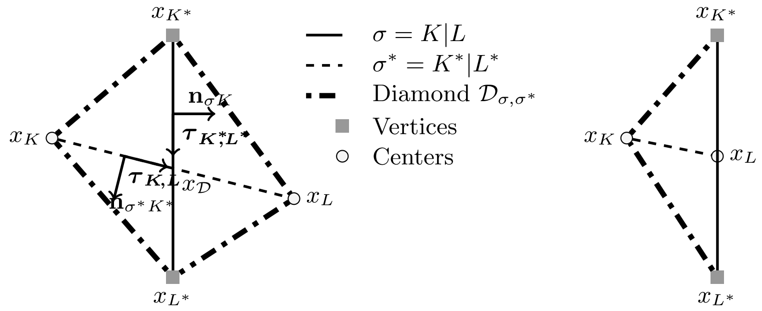


Fig. 1 Definition of the diamonds $\mathcal{D}_{\sigma,\sigma^*}$ and related notations

and $\mathfrak{D}$. We now introduce the space of scalar fields which are associated to each cell $\mathbb{R}^{\mathcal{T}}$, and space of vector fields constant on the diamonds $\left(\mathbb{R}^2\right)^{\mathfrak{D}}$:

$$u_{\mathcal{T}} \in \mathbb{R}^{\mathcal{T}} \iff u_{\mathcal{T}} = \left((u_K)_{K\in\overline{\mathfrak{M}}}, (u_{K^*})_{K^*\in\overline{\mathfrak{M}^*}}\right) \text{ and } \xi_{\mathfrak{D}} \in \left(\mathbb{R}^2\right)^{\mathfrak{D}} \iff \xi_{\mathfrak{D}} = \left(\xi_{\mathcal{D}}\right)_{\mathcal{D}\in\mathfrak{D}}.$$

To enforce Dirichlet boundary conditions, we introduce the set of Dirichlet boundary primal and dual cells: $\partial\mathfrak{M}_D = \{K \in \partial\mathfrak{M} : K \subset \Gamma_D\}$ and $\partial\mathfrak{M}^*_D = \{K^* \in \partial\mathfrak{M}^* : x_{K^*} \in \overline{\Gamma}_D\}$, and, for a given $v \in C(\Gamma^D)$, we define

$$\mathrm{E}_v^{\Gamma_D} = \{u_{\mathcal{T}} \in \mathbb{R}^{\mathcal{T}} \mid \forall K \in \partial\mathfrak{M}_D,\ u_K = v(x_K) \text{ and } \forall K^* \in \partial\mathfrak{M}^*_D,\ u_{K^*} = v(x_{K^*})\}.$$

We also define discrete bilinear forms on $\mathbb{R}^{\mathcal{T}}$ and $\left(\mathbb{R}^2\right)^{\mathfrak{D}}$ by

$$\begin{aligned} [\![v_{\mathcal{T}}, u_{\mathcal{T}}]\!]_{\mathcal{T}} &= \frac{1}{2}\sum_{K\in\mathfrak{M}} \mathrm{m}_K u_K v_K + \frac{1}{2}\sum_{K^*\in\overline{\mathfrak{M}^*}} \mathrm{m}_{K^*} u_{K^*} v_{K^*}, \quad \forall (u_{\mathcal{T}}, v_{\mathcal{T}}) \in \left(\mathbb{R}^{\mathcal{T}}\right)^2, \\ \left(\xi_{\mathfrak{D}}, \varphi_{\mathfrak{D}}\right)_{\mathfrak{D}} &= \sum_{\mathcal{D}\in\mathfrak{D}} \mathrm{m}_{\mathcal{D}}\, \xi_{\mathcal{D}} \cdot \varphi_{\mathcal{D}}, \quad \forall (\xi_{\mathfrak{D}}, \varphi_{\mathfrak{D}}) \in \left(\left(\mathbb{R}^2\right)^{\mathfrak{D}}\right)^2. \end{aligned}$$

The DDFV method is based on the definition of a discrete gradient operator $\nabla^{\mathfrak{D}} : \mathbb{R}^{\mathcal{T}} \to \left(\mathbb{R}^2\right)^{\mathfrak{D}}$, defined by $\nabla^{\mathfrak{D}} u_{\mathcal{T}} = \left(\nabla^{\mathcal{D}} u_{\mathcal{T}}\right)_{\mathcal{D}\in\mathfrak{D}}$, where

$$\nabla^{\mathcal{D}} u_{\mathcal{T}} = \frac{1}{2m_{\mathcal{D}}} \left(\mathrm{m}_\sigma (u_L - u_K)\mathbf{n}_{\sigma K} + \mathrm{m}_{\sigma^*}(u_{L^*} - u_{K^*})\mathbf{n}_{\sigma^* K^*}\right) \quad \forall \mathcal{D} \in \mathfrak{D}. \tag{4}$$

Finally, we introduce a reconstruction operator on diamonds $r^{\mathfrak{D}}$. It is a mapping from $\mathbb{R}^{\mathcal{T}}$ to $\mathbb{R}^{\mathfrak{D}}$ defined for all $u_{\mathcal{T}} \in \mathbb{R}^{\mathcal{T}}$ by $r^{\mathfrak{D}} u_{\mathcal{T}} = \left(r^{\mathcal{D}} u_{\mathcal{T}}\right)_{\mathcal{D}\in\mathfrak{D}}$, where for $\mathcal{D} \in \mathfrak{D}$, whose vertices are x_K, x_L, x_{K^*}, x_{L^*}, $r^{\mathcal{D}} u_{\mathcal{T}} = \frac{1}{4}(u_K + u_L + u_{K^*} + u_{L^*})$. One can now introduce a DDFV discretisation of $(u, w, v) \mapsto \int_\Omega u \nabla w \cdot \nabla v$, defined by

$$T_\mathfrak{D} : (u_\mathcal{T}, w_\mathcal{T}, v_\mathcal{T}) \mapsto \sum_{\mathcal{D}\in\mathfrak{D}} \mathrm{m}_\mathcal{D} r^\mathcal{D} u_\mathcal{T}\ \nabla^\mathcal{D} w_\mathcal{T} \cdot \nabla^\mathcal{D} v_\mathcal{T}.$$

Now, we first discretise the data by taking the mean values of N^{in}, P^{in} and C on the primal and dual cells, which define $N_\mathcal{T}^0$, $P_\mathcal{T}^0$ and $C_\mathcal{T}$. Then, for all $n \geq 0$, we look for $(N_\mathcal{T}^{n+1}, P_\mathcal{T}^{n+1}, \phi_\mathcal{T}^{n+1}) \in \mathrm{E}_{N^D}^{\Gamma_D} \times \mathrm{E}_{P^D}^{\Gamma_D} \times \mathrm{E}_{\phi^D}^{\Gamma_D}$ solution to:

$$[\![\frac{N_\mathcal{T}^{n+1} - N_\mathcal{T}^n}{\Delta t}, v_\mathcal{T}]\!]_\mathcal{T} + T_\mathfrak{D}(N_\mathcal{T}^{n+1}, \log(N_\mathcal{T}^{n+1}) - \phi_\mathcal{T}^{n+1}, v_\mathcal{T}) = 0 \quad \forall v_\mathcal{T} \in \mathrm{E}_0^{\Gamma_D}, \tag{5a}$$

$$[\![\frac{P_\mathcal{T}^{n+1} - P_\mathcal{T}^n}{\Delta t}, v_\mathcal{T}]\!]_\mathcal{T} + T_\mathfrak{D}(P_\mathcal{T}^{n+1}, \log(P_\mathcal{T}^{n+1}) + \phi_\mathcal{T}^{n+1}, v_\mathcal{T}) = 0 \quad \forall v_\mathcal{T} \in \mathrm{E}_0^{\Gamma_D}, \tag{5b}$$

$$\lambda^2 \left(\nabla^\mathcal{D} \phi_\mathcal{T}^{n+1}, \nabla^\mathcal{D} v_\mathcal{T}\right)_\mathfrak{D} = [\![C_\mathcal{T} + P_\mathcal{T}^{n+1} - N_\mathcal{T}^{n+1}, v_\mathcal{T}]\!]_\mathcal{T} \quad \forall v_\mathcal{T} \in \mathrm{E}_0^{\Gamma_D}. \tag{5c}$$

In (5a) and (5b), we use the notation $\log(u_\mathcal{T}) = \left((\log(u_K))_{K\in\overline{\mathfrak{M}}}, (\log(u_{K^*}))_{K^*\in\overline{\mathfrak{M}^*}}\right)$.

2.2 *The HFV Scheme*

In order to define the HFV scheme, we need to introduce a pyramidal submesh. To do so, one has to assume that each cell $K \in \mathfrak{M}$ is star-shaped with respect to its center x_K (we recall that x_K is not necessarily the barycentre of K). We then define $P_{K,\sigma}$ as the pyramid (triangle) of base σ and apex x_K. Given any $\sigma \in \mathcal{E}$, we denote by $\overline{x}_\sigma$ the barycentre of σ, and by $d_{K,\sigma}$ the euclidean distance between σ and x_K. Finally, we define the hybrid discretisation (or mesh) as $\mathcal{D} = (\mathfrak{M}, \mathcal{E})$.

We now introduce the space of discrete (scalar) hybrid unknowns $\underline{V}_\mathcal{D}$:

$$\underline{u}_\mathcal{D} \in \underline{V}_\mathcal{D} \Longleftrightarrow \underline{u}_\mathcal{D} = \left((u_K)_{K\in\mathfrak{M}}, (u_\sigma)_{\sigma\in\mathcal{E}}\right),$$

where the $u_K \in \mathbb{R}$ are the cell unknowns and the $u_\sigma \in \mathbb{R}$ are the edges unknowns (approximation of the trace of the solutions on the edges). To enforce Dirichlet boundary conditions, for a given $v \in C(\Gamma^D)$, we define

$$\underline{V}_{\mathcal{D},v}^{\Gamma_D} = \{\underline{u}_\mathcal{D} \in \underline{V}_\mathcal{D} \mid \forall \sigma \in \partial\mathfrak{M}_D,\ u_\sigma = v(\overline{x}_\sigma)\}.$$

As for the DDFV framework, we define a bilinear form on $\underline{V}_\mathcal{D}$, discrete counterpart of the inner product on $L^2(\Omega)$ as

$$[\![\underline{u}_\mathcal{D}, \underline{v}_\mathcal{D}]\!]_\mathfrak{M} = \sum_{K\in\mathfrak{M}} \mathrm{m}_K u_K v_K, \quad \forall(\underline{u}_\mathcal{D}, \underline{v}_\mathcal{D}) \in \underline{V}_\mathcal{D}^2.$$

The HFV method is based on the definition of a discrete gradient operator $\nabla_\mathcal{D} : \underline{V}_\mathcal{D} \to (\mathbb{R}^2)^\Omega$ which maps discrete hybrid unknowns onto piecewise constant functions on the pyramidal submesh. More precisely, given $\underline{v}_{\mathfrak{D}} \in \underline{V}_{\mathfrak{D}}$, $K \in \mathfrak{M}$ and $\sigma \in \mathcal{E}_K$,

$$\nabla_{\mathfrak{D}} \underline{v}_{\mathfrak{D}|P_{K,\sigma}} = G_K \underline{v}_{\mathfrak{D}} + S_{K,\sigma} \underline{v}_{\mathfrak{D}},$$

where, for some $\eta > 0$, the consistent and stabilisation parts of the gradient are given by

$$G_K \underline{v}_{\mathfrak{D}} = \frac{1}{\mathrm{m}_K} \sum_{\sigma' \in \mathcal{E}_K} m_{\sigma'} v_{\sigma'} n_{K,\sigma'} \text{ and } S_{K,\sigma} \underline{v}_{\mathfrak{D}} = \frac{\eta}{d_{K,\sigma}} \big(v_\sigma - v_K - G_K \underline{v}_K \cdot (\overline{x}_\sigma - x_K) \big) n_{K,\sigma}.$$

One can now define the discrete counterpart of $(u, v) \mapsto \int_\Omega \nabla u \cdot \nabla v$ as

$$a_{\mathfrak{D}} : (\underline{u}_{\mathfrak{D}}, \underline{v}_{\mathfrak{D}}) \mapsto \int_\Omega \nabla_{\mathfrak{D}} \underline{u}_{\mathfrak{D}} \cdot \nabla_{\mathfrak{D}} \underline{v}_{\mathfrak{D}}.$$

We introduce as previously local reconstruction operators on cells $r^K : \underline{V}_{\mathfrak{D}} \to \mathbb{R}$, such that for any $u_{\mathfrak{D}} \in \underline{V}_{\mathfrak{D}}$, $r^K(\underline{u}_{\mathfrak{D}}) = \dfrac{1}{|\mathcal{E}_K|} \displaystyle\sum_{\sigma \in \mathcal{E}_K} \dfrac{u_K + u_\sigma}{2}$, where $|\mathcal{E}_K|$ is the cardinal of the finite set $\mathcal{E}_K$. One can now introduce a HFV discretisation of $(u, w, v) \mapsto \int_\Omega u \nabla w \cdot \nabla v$, defined by

$$T_{\mathfrak{D}} : (\underline{u}_{\mathfrak{D}}, \underline{w}_{\mathfrak{D}}, \underline{v}_{\mathfrak{D}}) \mapsto \sum_{K \in \mathfrak{M}} r^K(\underline{u}_{\mathfrak{D}}) \int_K \nabla_{\mathfrak{D}} \underline{w}_{\mathfrak{D}} \cdot \nabla_{\mathfrak{D}} \underline{v}_{\mathfrak{D}}.$$

We now discretise the data by taking the mean values of N^{in}, P^{in} and C on the cells and edges, which define $\underline{P}^0_{\mathfrak{D}}$, $\underline{N}^0_{\mathfrak{D}}$ and $\underline{C}_{\mathfrak{D}}$. Then, for all $n \geq 0$, we look for $(\underline{N}^{n+1}_{\mathfrak{D}}, \underline{P}^{n+1}_{\mathfrak{D}}, \underline{\phi}^{n+1}_{\mathfrak{D}}) \in \underline{V}^{\Gamma_D}_{\mathfrak{D},N^D} \times \underline{V}^{\Gamma_D}_{\mathfrak{D},P^D} \times \underline{V}^{\Gamma_D}_{\mathfrak{D},\phi^D}$ solution to:

$$[\![\frac{\underline{N}^{n+1}_{\mathfrak{D}} - \underline{N}^n_{\mathfrak{D}}}{\Delta t}, \underline{v}_{\mathfrak{D}}]\!]_{\mathfrak{M}} + T_{\mathfrak{D}}(\underline{N}^{n+1}_{\mathfrak{D}}, \log(\underline{N}^{n+1}_{\mathfrak{D}}) - \underline{\phi}^{n+1}_{\mathfrak{D}}, \underline{v}_{\mathfrak{D}}) = 0 \; \forall \underline{v}_{\mathfrak{D}} \in \underline{V}^{\Gamma_D}_{\mathfrak{D},0}, \tag{6a}$$

$$[\![\frac{\underline{P}^{n+1}_{\mathfrak{D}} - \underline{P}^n_{\mathfrak{D}}}{\Delta t}, \underline{v}_{\mathfrak{D}}]\!]_{\mathfrak{M}} + T_{\mathfrak{D}}(\underline{P}^{n+1}_{\mathfrak{D}}, \log(\underline{P}^{n+1}_{\mathfrak{D}}) + \underline{\phi}^{n+1}_{\mathfrak{D}}, \underline{v}_{\mathfrak{D}}) = 0 \; \forall \underline{v}_{\mathfrak{D}} \in \underline{V}^{\Gamma_D}_{\mathfrak{D},0}, \tag{6b}$$

$$\lambda^2 a_{\mathfrak{D}} \left(\underline{\phi}^{n+1}_{\mathfrak{D}}, \underline{v}_{\mathfrak{D}} \right) = [\![\underline{C}_{\mathfrak{D}} + \underline{P}^{n+1}_{\mathfrak{D}} - \underline{N}^{n+1}_{\mathfrak{D}}, \underline{v}_{\mathfrak{D}}]\!]_{\mathfrak{M}} \; \forall \underline{v}_{\mathfrak{D}} \in \underline{V}^{\Gamma_D}_{\mathfrak{D},0}. \tag{6c}$$

As previously, we use the notation $\log(\underline{u}_{\mathfrak{D}}) = \big((\log(u_K))_{K \in \mathfrak{M}}, (\log(u_\sigma))_{\sigma \in \mathcal{E}} \big)$.

2.3 Some Structural Differences Between Schemes

As highlighted by the unified presentation above, both schemes are very similar and rely on the same features. Note that both local reconstruction operators $r^{\mathcal{D}}$ and r^K take into account all the local unknowns of the geometric entity considered (diamond or cells), this property is the key point of the analysis of this kind of schemes, see [2, 5]. However, the schemes exhibit differences, some of which are listed below:

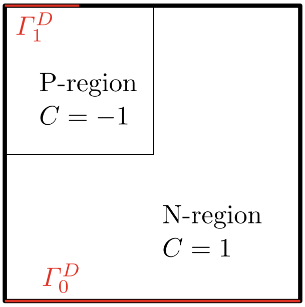
- the discrete HFV gradient $\nabla_{\mathfrak{D}}$ includes a stabilisation term for the sake of coercivity and the stabilisation parameter η has to be chosen a priori, whereas the DDFV one is simpler and do not need any choice of parameter;
- the DDFV unknowns are all "volumic", in the sense that there are associated to geometric entities with non-zero two-dimensional measures, whereas the faces unknowns of the HFV method have no mass and have no influence on the discrete time derivative terms $[\![\underline{N}_{\mathfrak{D}}^{n+1} - \underline{N}_{\mathfrak{D}}^{n}, \underline{v}_{\mathfrak{D}}]\!]_{\mathfrak{M}}$ and $[\![\underline{P}_{\mathfrak{D}}^{n+1} - \underline{P}_{\mathfrak{D}}^{n}, \underline{v}_{\mathfrak{D}}]\!]_{\mathfrak{M}}$;
- the cells unknowns of the HFV scheme can be eliminated before solving linear systems, using a static condensation procedure (see [5, Sect. 5.1.2.]), this procedure cannot be performed for the DDFV method;
- the HFV scheme can be used in 3D without any modification (the edges become faces), whereas using a DDFV method in 3D requires more sophisticated changes (see [4]).

3 Numerical Experiments

The two numerical schemes described here are nonlinear, hence their algebraic realisations boil down to the resolution of nonlinear systems of equations. To do so, we use Newton method, with an adaptative time stepping strategies: if the Newton method does not converge, we try to compute the solution for a smaller time step $0.5 \times \Delta t$. If the method converges, we use a bigger time step $1.4 \times \Delta t$. The initial time step is denoted by Δt_{ini}, and we also impose a maximal time step Δt_{max}. For the HFV scheme, at each system resolution, a static condensation is used to eliminate the cell unknowns (see [5, Sect. 5.1.2.]), and we use $\eta = 1.5$. Note that we use N, P and ϕ as discrete unknowns in the schemes.

The test case used below follow the framework used in [5] to describe a 2D PN-junction, whose geometry is described in Fig. 2. The domain Ω is the unit square $]0, 1[^2$. For the boundary conditions, we split $\Gamma^D = \Gamma_0^D \cup \Gamma_1^D$ with $\Gamma_0^D =$

Fig. 2 PN diode geometry



$[0, 1] \times \{0\}$ and $\Gamma_1^D = [0, 0.25] \times \{1\}$. For $i \in \{0, 1\}$, we let $N^D = N_i^D$, $P^D = P_i^D$ and $\phi^D = \frac{\log(N_i^D) - \log(P_i^D)}{2}$ on Γ_i^D. To be consistent with the compatibility condition (2) we assume that there exists a constant α_0 such that $\log(N^D \times P^D) = \alpha_0$. Therefore for given N^D and α_0 we set $P^D = \frac{e^{\alpha_0}}{N^D}$ on Γ^D.

Thus, one has $\alpha_N = \alpha_P = \frac{\alpha_0}{2}$. The doping profile C is piecewise constant, equal to -1 in the P-region and 1 in the N-region (see Fig. 2). Last, we use the following smooth initial conditions: $N_0(x, y) = N_1^D + (N_0^D - N_1^D)(1 - \sqrt{y})$ and $P_0(x, y) = P_1^D + (P_0^D - P_1^D)(1 - \sqrt{y})$.

3.1 Positivity

The test uses the following values: $\lambda = 0.05$, $N_0^D = 0.1$, $N_1^D = 1$ and $\alpha_0 = -4$. We perform a test on a distorted quadrangle mesh (mesh_quad_6 of the FVCA 8 Benchmark), with $\Delta t_{ini} = 1.4\,10^{-3}$ and $\Delta t_{max} = 0.1$. We show in Fig. 3 the evolution of the minimal values of P and N, along with the time step and the number of Newton's iterations needed to compute the solutions at a given time for each time step. The minimal values are taken on every unknowns (primal and dual cells for the DDFV scheme, cells and faces for the HFV one). One can see that both schemes compute, as expected by the theoretical results, positive densities. The minimal values computed are of the same order for both schemes. Moreover, both computations proceed without the need of a time step reduction. Regarding the cost, it appears that the HFV scheme needs more Newton iterations than the DDFV one (90 vs. 63). For

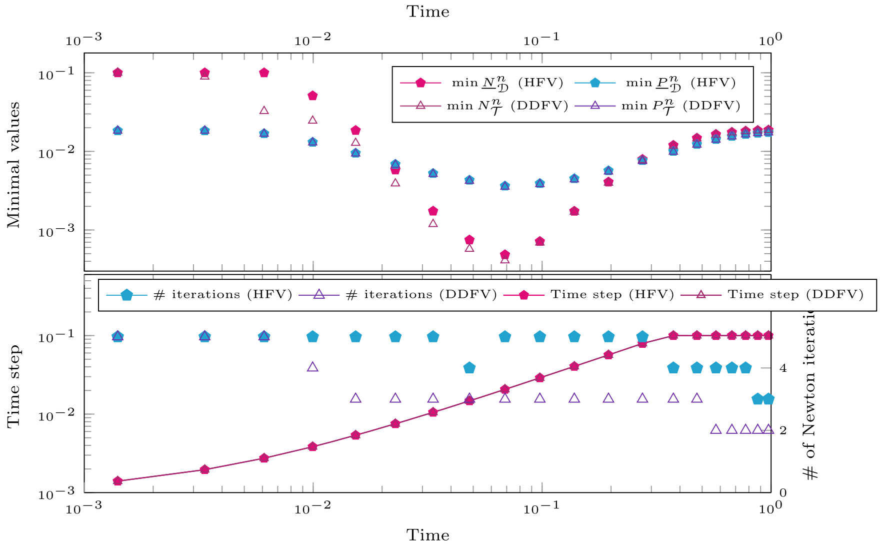


Fig. 3 **Test-case**. Evolution of the discrete minimal values, time step and cost

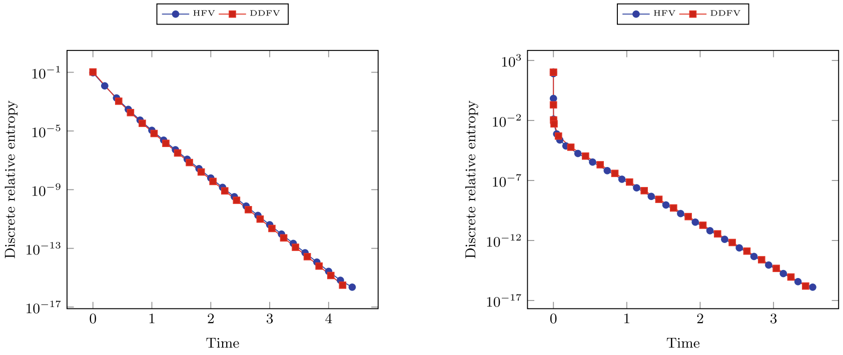
both schemes, the number of iteration decay as the time increases, since the solutions converge exponentially fast towards the equilibrium.

3.2 Long-Time Behaviour

At the continuous level, one usually quantify the distance between the solution (N, P, ϕ) and the equilibrium (N^e, P^e, ϕ^e) by looking at the relative entropy, defined as $\mathbb{E}(t) = \int_\Omega N^e H\left(\frac{N}{N^e}\right) + \int_\Omega P^e H\left(\frac{P}{P^e}\right) + \frac{\lambda^2}{2}\|\nabla(\phi - \phi^e)\|^2_{L^2(\Omega)}$, with $H : s \mapsto s\log(s) - s + 1$. One can check that (N, P, ϕ) coincides with the equilibrium if and only if the relative entropy cancels. In the following, we are interested in the evolution of the discrete counterparts of this quantities, defined as

$$\mathbb{E}^n_{\mathfrak{D}} = [\![\underline{N}^e_{\mathfrak{D}} H\left(\frac{\underline{N}^n_{\mathfrak{D}}}{\underline{N}^e_{\mathfrak{D}}}\right), \underline{1}_{\mathfrak{D}}]\!]_{\mathfrak{M}} + [\![\underline{P}^e_{\mathfrak{D}} H\left(\frac{\underline{P}^n_{\mathfrak{D}}}{\underline{P}^e_{\mathfrak{D}}}\right), \underline{1}_{\mathfrak{D}}]\!]_{\mathfrak{M}} + \frac{\lambda^2}{2} a_{\mathfrak{D}}\left(\underline{\phi}^n_{\mathfrak{D}} - \underline{\phi}^e_{\mathfrak{D}}, \underline{\phi}^n_{\mathfrak{D}} - \underline{\phi}^e_{\mathfrak{D}}\right)$$

for the HFV scheme (where $\underline{1}_{\mathfrak{D}}$ is the discrete elements whose coordinates are 1, and the product, quotient and functions are applied coordinate-wise) and similar definition for the DDFV scheme. Note that the HFV entropy does not take into account the edge unknowns of the discrete densities. To compute the discrete equilibrium, we use a nonlinear scheme for (3) and get $\underline{\phi}^e_{\mathfrak{D}}$, then we defined the associated densities following the continuous relations $N^e = e^{\alpha_N + \phi^e}$ and $P^e = e^{\alpha_P - \phi^e}$. We consider a test case with physical data $N_0^D = e$, $N_1^D = 1$ and $\alpha_0 = 0$. We also use two different values of the Debye length λ, respectively 1 and 0.01. We perform simulations on a triangular mesh, with a $\Delta t_{ini} = \Delta t_{max} = 0.1$. On Fig. 4, we show the evolutions of the discrete relative entropies along time, for the two values of λ and both schemes.



(a) $\lambda = 1$

(b) $\lambda = 0.01$

Fig. 4 Long-time behaviour. Evolution of the discrete relative entropies

As expected, the convergence towards the equilibrium is exponentially fast, as in the continuous framework. Moreover, it is remarkable to notice that the decay rates are almost the same for both schemes. Moreover, with the small Debye length (Fig. 4b), both schemes are able to capture the behaviour with a very fast evolution far from the equilibrium, then slower once close to it.

References

1. Bessemoulin-Chatard, M., Chainais-Hillairet, C.: J. Numer. Math. **25**(3), 147–168 (2017)
2. Cancès, C., Chainais-Hillairet, C., Krell, S.: Comput. Methods Appl. Math. **18**, 407–432 (2018)
3. Cancès, C., Guichard, C.: Found. Comput. Math. **17**(6), 1525–1584 (2017)
4. Coudière, Y., Hubert, F.: SIAM J. Sci. Comput. **33**(4), 1739–1764 (2011)
5. Moatti, J.: ESAIM, Math. Model. Numer. Anal., in press (2023)

Reduced Basis Approach for Convection-Diffusion Equations with Non-linear Boundary Reaction Conditions

S. Matera, C. Merdon, and D. Runge

Abstract This paper aims at an efficient strategy to solve drift-diffusion problems with non-linear boundary conditions as they appear, e.g., in heterogeneous catalysis. Since the non-linearity only involves the degrees of freedom along (a part of) the boundary, a reduced basis ansatz is suggested that computes discrete Green's-like functions for the present drift-diffusion operator such that the global non-linear problem reduces to a smaller non-linear problem for a boundary method. The computed basis functions are completely independent of the non-linearities. Thus, they can be reused for problems with the same differential operator and geometry. Corresponding scenarios might be inverse problems in heterogeneous catalysis but also modeling the effect of different catalysts in the same reaction chamber. The strategy is explained for a mass-conservative finite volume method and demonstrated on a simple numerical example for catalytic CO oxidation.

Keywords Reduced basis · Non-linear boundary conditions · Finite volume methods

S. Matera
Fritz Haber Institute of the Max Planck Society, Faradayweg 4-6, 14195 Berlin, Germany
e-mail: matera@fhi-berlin.mpg.de

C. Merdon · D. Runge (✉)
Weierstrass Institute for Applied Analysis and Stochastics, Mohrenstraße 39, 10117 Berlin, Germany
e-mail: daniel.runge@wias-berlin.de

C. Merdon
e-mail: christian.merdon@wias-berlin.de

E. Franck et al. (eds.), *Finite Volumes for Complex Applications X—Volume 1, Elliptic and Parabolic Problems*, Springer Proceedings in Mathematics & Statistics 432,
https://doi.org/10.1007/978-3-031-40864-9_28

1 Introduction

Reduced basis approaches have gained popularity for the solution of partial differential equations (PDE), especially to address parameter dependencies [1, 11]. The idea is to employ only a few simulations with a general purpose discretization and high resolution to determine a (small) set of problem-specific basis functions (offline phase). This then allows for a fast solution of the PDE using the previously computed basis functions (online phase).

We present a new reduced basis approach for linear convection-diffusion equations with highly non-linear flux boundary conditions. Such models are often good approximations in problems which involve surface chemistry, e.g., heterogeneous catalysis or electrochemistry [7, 8]. Our approach decomposes the problem into a set of linear problems to obtain the reduced basis and a non-linear problem which only depends on the degrees of freedom on the boundary. Being essentially a discrete representation of a type of Green's function of the linear operator, the reduced basis is independent of the non-linearity and can be reused for very different scenarios, e.g. inverse problems for parametrizing the non-linear boundary condition, which is a common task in surface chemistry applications. The connection to a type of Green's functions renders the approach closely related to boundary integral methods like [12], but without explicit knowledge of the former. In principle, the problem can be discretized by any method, e.g., finite volumes or finite element methods. Here, we demonstrate the approach with Voronoi finite volume methods, which allow for advantageous structural properties like conservation of mass or the non-negativity of Y [5]. An algebraic formulation reveals that the reduced basis solution agrees with the solution of the fully coupled problem discretized with the same underlying method.

The rest of this paper is structured as follows. Section 2 introduces the model problem and lays out its finite volume discretization. Section 3 concerns the design and implementation of the set of reduced basis functions. Section 4 demonstrates the approach in one model application. Finally, Section 5 gives an outlook to further aspects and target applications.

2 Model Problem and FV Discretisation

For a given domain Ω, a (divergence-free) velocity field $\boldsymbol{v}$, a positive diffusion coefficient $D \in \mathbb{R}$, and a right-hand side $f \in L^2(\Omega)$, the model problem seeks some function $Y \in H^1(\Omega)$ such that

$$\operatorname{div}(Y\boldsymbol{v} - D\nabla Y) = f. \tag{1}$$

On the boundary $\Gamma := \partial\Omega = \Gamma_{\text{in}} \cup \Gamma_{\text{out}} \cup \Gamma_{\text{nl}} \cup \Gamma_{\text{rest}}$, various boundary conditions may apply. Here, we assume some inlet condition

$$Y = Y_{\text{in}} \quad \text{along} \quad \Gamma_{\text{in}},$$

an outflow condition

$$\boldsymbol{n} \cdot (Y\boldsymbol{v} - D\nabla Y) = Y\boldsymbol{v} \cdot \boldsymbol{n} \quad \text{along} \quad \Gamma_{\text{out}},$$

and a non-linear boundary condition

$$\boldsymbol{n} \cdot (Y\boldsymbol{v} - D\nabla Y) = R(Y|_{\Gamma_{\text{nl}}}) \quad \text{along} \quad \Gamma_{\text{nl}} \tag{2}$$

for some given functional R. On the remaining part of the boundary Γ_{rest} we assume homogeneous Neumann boundary conditions for simplicity. But this approach can easily be extended to linear Robin boundary conditions.

Here, we consider a finite volume discretization of the model problem based on some boundary-conforming Delaunay triangulation $\mathcal{T}$, which computes a piecewise constant approximation $Y_h \in P_0(\mathcal{K})$ to Y with respect to the set of open, convex Voronoi cells $\mathcal{K}$. Each cell $K \in \mathcal{K}$ has some associated collocation point $\boldsymbol{x}_K \in \overline{K}$. The subset of cells at the boundary Γ_{nl} are denoted by $\mathcal{K}_{\text{nl}}$ and their associated collocation points are located on the boundary Γ_{nl}. Details on the implementation can be found, e.g., in the documentation of the Julia package `VoronoiFVM.jl` [3] which was also used as a basis for the implementation of the reduced bases.

Note, that we employ an exponential fitting flux discretization that ensures desirable structural properties like non-negativity and a maximum principle for Y under certain conditions, e.g., if $R = 0$ and $\boldsymbol{v}$ is divergence-free [5]. A discussion and numerical investigation of convergence rates of the exponential fitting scheme can be found in e.g. [6, p. 536].

3 Reduced Basis Approach

This section describes the main idea to speed up the computation of the solution of the model problem with the help of a reduced basis related to the boundary degrees of freedom of the non-linear boundary $\mathcal{K}_{\text{nl}}$. For this, observe that the discrete solution Y_h can be decomposed into

$$Y_h = Y_0 + Y_{\text{nl}} := Y_0 + \sum_{K \in \mathcal{K}_{\text{nl}}} \alpha_K Y_K$$

where Y_0 solves the discretized linear sub-problem for $R = 0$ and each Y_K solves the discretized fully linear problem

$$\begin{aligned} \operatorname{div}(Y_K\boldsymbol{v} - D\nabla Y_K) &= 0 \\ \boldsymbol{n} \cdot (Y_K\boldsymbol{v} - D\nabla Y_K) &= \chi_K \quad \text{along} \quad \Gamma_{\text{nl}} \end{aligned} \tag{3}$$

and homogeneous boundary conditions on the remaining boundary $\Gamma \setminus \Gamma_{\mathrm{nl}}$. Here, χ_K is the characteristic function of $\partial K \cap \Gamma_{\mathrm{nl}}$ for $K \in \mathcal{K}_{\mathrm{nl}}$. Note, that Y_K can be interpreted as a type of discrete Green's function of the drift-diffusion operator for its corresponding part $\partial K \subset \Gamma_{\mathrm{nl}}$ of the boundary.

To further investigate this on the algebraic level, let x_{nl}, x_0 and x_K denote the coefficient vectors of the Y_{nl}, Y_0 and Y_K parts, respectively. They are given by a solution of the linear systems of equations

$$Ax_0 = b_0 \tag{4}$$

$$Ax_K = b_K, \tag{5}$$

where A is the finite-volume discretization of the drift-diffusion operator, b_0 encodes all linear (Y-independent) boundary and right-hand side data, and b_K encodes the boundary condition (3). Note here, that the matrix A is the same in all computations and it is therefore straightforward to solve for x_0 and all x_K efficiently and in parallel. Also, for a constant inlet concentration and $f \equiv 0$, it holds $Y_0 \equiv Y_{\mathrm{in}}$ which allows to avoid the computation of Y_0 in that case.

To determine the coefficients x_{nl} for $Y_{\mathrm{nl}} = \sum_{K \in \mathcal{K}_{\mathrm{nl}}} \alpha_K Y_K$, it is required to solve the non-linear system

$$A(x_0 + x_{\mathrm{nl}}) = b_0 + b_{\mathrm{nl}}\left(x_0 + x_{\mathrm{nl}}\right) \quad \Leftrightarrow \quad Ax_{\mathrm{nl}} = b_{\mathrm{nl}}\left(x_0 + x_{\mathrm{nl}}\right) \tag{6}$$

where $b_{\mathrm{nl}}(x)$ encodes the finite volume discretisation of the (non-linear) catalytic boundary data. Inserting the decomposition into the reduced boundary basis $x_{\mathrm{nl}} = \sum_{K \in \mathcal{K}_{\mathrm{nl}}} \alpha_K x_K$ and using (5), this is equivalent to seeking α_K such that

$$\sum_{K \in \mathcal{K}_{\mathrm{nl}}} \alpha_K b_K = b_{\mathrm{nl}} \left(x_0 + \sum_{K \in \mathcal{K}_{\mathrm{nl}}} \alpha_K x_K \right).$$

In a finite volume method that approximates the boundary integrals in the assembly of b_K and $b_{\mathrm{nl}}(x)$ by a quadrature rule evaluating in the collocation point, the non-linear system to determine the coefficients α_K can be rewritten into

$$\alpha_L = R \left(Y_0(\boldsymbol{x}_L) + \sum_{K \in \mathcal{K}_{\mathrm{nl}}} \alpha_K Y_K\left(\boldsymbol{x}_L\right) \right) \quad \text{for all } L \in \mathcal{K}_{\mathrm{nl}}. \tag{7}$$

Here, $\boldsymbol{x}_L \in \Gamma_{\mathrm{nl}}$ denotes the collocation point of the cell $L \in \mathcal{K}_{\mathrm{nl}}$. Note, that this is a much smaller system to solve than the global system (6).

Remark 1 As usual in a reduced basis setting, computations can be split into an offline and online phase. The offline phase computes the coefficients of the linear part x_0 and the reduced basis functions Y_K resp. their coefficients x_K. The online phase solves the reduced system (7) for a given function R.

Remark 2 The amount of work in the online phase can be further reduced by combining basis functions to larger ones or applying other compression techniques. In some applications it might be sufficient to make use of only one basis function that combines all Y_K into a single basis function. Then (7) turns into a single equation.

Remark 3 For solving (7) only boundary values of the reduced basis functions are needed. There is no need to store the whole vector x_K. In case volume information is needed, e.g. for plotting or evaluating quantities of interest, the full solution can be obtained from a linear solve, where the right-hand side vector $\sum_{K \in \mathcal{K}_{\text{nl}}} \alpha_K b_K$ is used.

4 Model Application and Numerical Example

This section studies a simple, but realistic model application that is based on the catalytic CO oxidation according to the reaction equation

$$2\text{CO} + \text{O}_2 \rightarrow 2\text{CO}_2$$

in a two-dimensional channel domain $\Omega := (0, 5) \times (0, 1)$ at a small catalytic boundary $\Gamma_{\text{nl}} := 0 \times (2, 3)$. The involved species mass fractions $\boldsymbol{Y} := (Y_{\text{CO}}, Y_{\text{O}_2}, Y_{\text{CO}_2})$ with inlet mass fractions $\boldsymbol{Y}_{\text{in}} = (0.2, 0.8, 0)$ at Γ_{in} are advected by a Hagen–Poiseuille flow $\boldsymbol{v}(x, y) := v_{\text{in}}(y(y-1), 0)^T$ from the inlet $\Gamma_{\text{in}} := \{0\} \times (0, 1)$ to the outlet $\Gamma_{\text{out}} := \{5\} \times (0, 1)$. For the rest of the boundary Γ_{inert}, inert wall boundary conditions are prescribed. For simplicity, *mass action kinetics* at the non-linear boundary are assumed, which result in the reaction function

$$R(Y_{\text{CO}}, Y_{\text{O}_2}, Y_{\text{CO}_2}) = k\,(Y_{\text{CO}})^2 (Y_{O_2})^1,$$

where k is a *reaction rate constant*. Note, that all three species are involved and their dynamics are coupled through *stoichiometric coefficients* according to the reaction above. Altogether, we seek mass fractions $\boldsymbol{Y}$ that satisfy

$$\begin{aligned}
\operatorname{div}(\boldsymbol{Y}\boldsymbol{v} - D\nabla \boldsymbol{Y}) &= 0 && \text{in } \Omega \\
\boldsymbol{Y} &= \boldsymbol{Y}_{\text{in}} && \text{along } \Gamma_{\text{in}} \\
\boldsymbol{n} \cdot (\boldsymbol{Y}\boldsymbol{v} - D\nabla \boldsymbol{Y}) &= 0 && \text{along } \Gamma_{\text{inert}} \\
\boldsymbol{n} \cdot (\boldsymbol{Y}\boldsymbol{v} - D\nabla \boldsymbol{Y}) &= \boldsymbol{Y}\boldsymbol{v} \cdot \boldsymbol{n} && \text{along } \Gamma_{\text{out}} \\
\boldsymbol{n} \cdot (\boldsymbol{Y}\boldsymbol{v} - D\nabla \boldsymbol{Y}) &= R(\boldsymbol{Y})\,(-2, -1, 1)^T && \text{along } \Gamma_{\text{nl}}.
\end{aligned} \tag{8}$$

For the sake of simplicity, we assume that D is a positive scalar, i.e., that the diffusion coefficients for all species coincide and that there is no cross-diffusion.

The global solution is computed using `VoronoiFVM.jl` which employs a damped Newton method where the sparse linear systems are solved using a direct

Table 1 Numbers of degrees of freedom for the global system and the reduced basis method as functions of the refinement level

Refinement level	Global degrees of freedom (N)	Reduced basis degrees of freedom
0	108	2
1	363	3
2	1,323	5
3	5,043	9
4	19,683	17
5	77,763	33
6	309,123	65
7	1,232,643	129

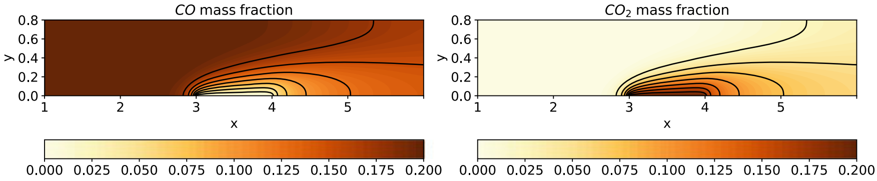


Fig. 1 Computed mass fractions of CO and CO_2 for $D = 10^{-2}$, $k = 10^{10}$ and $v_{\text{in}} = 1$

solver. For the offline phase of the reduced basis method, we employ the same linear solver. For the online phase, we exploit the linear dependence of the three non-linear boundary conditions in (8) which reduces the number of degrees of freedom per cell to one instead of three and automatically ensures the stoichiometry. The non-linear system (7) is solved using the implemented Newton solver with default line search and residual norm tolerance `ftol` $= 10^{-11}$ from the `NLsolve.jl` package [10]. This tolerance is selected as low as possible while still yielding convergence across all tested values of the reaction rate constant k. We test the approach for uniformly refined meshes with $n_0 = 6, n_{\text{level}} = 2 \cdot (n_{\text{level}-1}) - 1$ nodes in each direction. Table 1 lists the number N of degrees of freedom for the global problem and the reduced basis up to level 7.

Figure 1 shows a characteristic development of the mass fractions of CO and CO_2 along the catalytic surface. The results obtained by the reduced basis approach and the global solution agree within a tolerance close to `ftol`. While CO is consumed and CO_2 is produced, a boundary layer forms whose thickness is determined by the ratio of D and v_{in}. Note that the concentrations appear very uniform along the catalytic surface. This suggests that the compression described in Remark 2 could be applied here to further reduce the degrees of freedom for the online phase.

Figure 2 (left) compares the runtimes of a global solve via `VoronoiFVM.jl` [3] and the reduced basis scheme as functions of the number of uniform refinements for the parameters $D = 10^{-2}$, $k = 10^{10}$, $v_{\text{in}} = 1$. It features the offline phase for

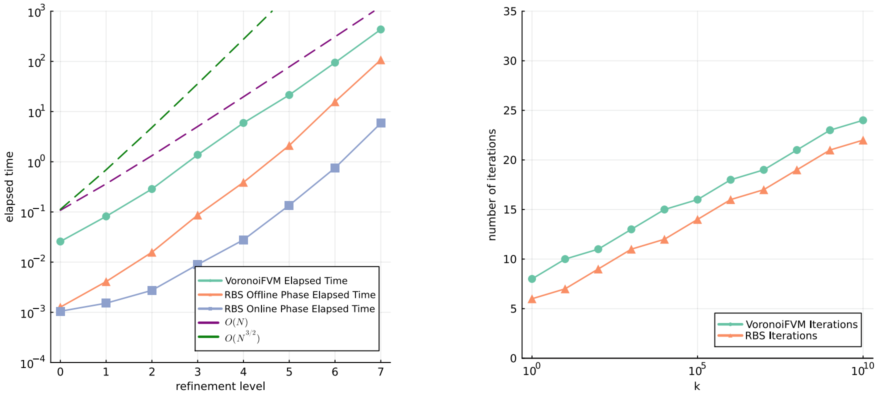


Fig. 2 Runtimes as a function of the refinement level (left) for $k = 10^{10}$ and number of Newton iterations as a function of k (right), both for $D = 10^{-2}$ and $v_{\text{in}} = 1$

the reduced basis setup and the online phase that solves the global problem via the reduced basis. For this and all other tested parameter settings, the online phase of the reduced basis method outperforms the global solver by about two orders of magnitude. In addition, the offline phase also comes at a significantly lower runtime. This is partly due to the fact that the assembly of the drift-diffusion operator A in (4) is executed in every Newton step by `VoronoiFVM.jl` whereas our offline phase requires this to be executed only once.

The number of Newton iterations for a fixed k in all experiments was largely independent of the refinement level and therefore is not shown. However, Fig. 2 (right) shows the number of Newton iterations on a fixed mesh (the finest one) versus k. Here, we see that the number of iterations increases with k, but that the reduced basis solver always requires a comparable number of Newton iterations as the global solver.

5 Outlook

Here, we only discussed a simple model problem for catalytic CO oxidation and demonstrated the approach for the case where the reduced basis covers the full resolution of the global problem, simply by singling out the boundary ansatz functions. As demonstrated, this already reduces the computational cost dramatically, especially when many problems of the same type have to be solved on the same geometry where only the non-linearity $R(Y|_{\Gamma_{\text{nl}}})$ varies. The proposed methodology is thus particularly suited for inverse problems or uncertainty quantification with a parameter-dependent non-linearity. Since the reduced basis is completely independent of the non-linearity, this might even include qualitatively different models for the boundary reaction, e.g. for different catalyst materials.

However, there are a number of aspects which require or allow for an extension. Particularly, this concerns the velocity field $\boldsymbol{v}$ for which no analytical expression is available in many practical applications. Therefore $\boldsymbol{v}$ has to be obtained from a CFD simulation. To ensure mass conservation within the model transport problem and also for the reduced basis, the discrete velocity field needs to be divergence-free according to the methodology derived in [6]. Here we plan to investigate novel, less costly divergence-free coupling strategies developed in the context of electrochemistry [4, 9]. A way to improve efficiency is to exploit that Y and thereby $R(Y|_{\Gamma_{\mathrm{nl}}})$ often is smooth along the boundary. This motivates to reduce the basis at the boundary, e.g., by combining the basis functions of neighbouring cells or considering wavelet basis approaches. In fact, a single basis approach has already been successfully applied [8].

A typical class of applications is flow problems coupled with surface chemical reaction. Indeed, the current study is part of a joint effort to combine transport simulations with detailed microkinetic models of heterogeneous catalysis. These hybrid models shall be employed to interpret modern in situ surface characterization experiments [2]. The corresponding complex employed instrumentation typically requires rather large and non-standard reaction chambers whereas the catalyst samples are rather small. We therefore expect the proposed methodology to be particularly effective, also because the resulting non-linear problems often are very challenging.

Acknowledgements The authors gratefully acknowledge the funding by the German Science Foundation (DFG) within the project "ME 4819/2-1", the CRC 1114 "Scaling Cascades in Complex Systems" (Project No. 235221301) and under Germany's Excellence Strategy–EXC 2008–390540038–UniSysCat.

References

1. Benner, P., Ohlberger, M., Cohen, A., Willcox, K.: Model Reduction and Approximation. SIAM, Philadelphia, PA (2017)
2. Frenken, J., Groot, I.: Operando research in heterogeneous catalysis. Springer (2017)
3. Fuhrmann, J., contributors: VoronoiFVM.jl: Finite volume solver for coupled nonlinear partial differential equations (2019–2021). https://doi.org/10.5281/zenodo.3529808
4. Fuhrmann, J., Guhlke, C., Linke, A., Merdon, C., Müller, R.: Induced charge electroosmotic flow with finite ion size and solvation effects. Electrochim. Acta **317**, 778–785 (2019)
5. Fuhrmann, J., Langmach, H.: Stability and existence of solutions of time-implicit finite volume schemes for viscous nonlinear conservation laws. Appl. Numer. Math. **37**(1), 201–230 (2001)
6. Fuhrmann, J., Linke, A., Langmach, H.: A numerical method for mass conservative coupling between fluid flow and solute transport. Appl. Numer. Math. **61**(4), 530–553 (2011)
7. Fuhrmann, J., Zhao, H., Langmach, H., Seidel, Y.E., Jusys, Z., Behm, R.J.: The role of reactive reaction intermediates in two-step heterogeneous electrocatalytic reactions: a model study. Fuel Cells **11**(4), 501–510 (2011)
8. Matera, S., Blomberg, S., Hoffmann, M.J., Zetterberg, J., Gustafson, J., Lundgren, E., Reuter, K.: Evidence for the active phase of heterogeneous catalysts through in situ reaction product imaging and multiscale modeling. ACS Catal. **5**, 4514–4518 (2015)

9. Merdon, C., Fuhrmann, J., Linke, A., Streckenbach, T., Neumann, F., Khodayari, M., Baltruschat, H.: Inverse modeling of thin layer flow cells for detection of solubility, transport and reaction coefficients from experimental data. Electrochim. Acta **211**, 1–10 (2016)
10. Mogensen, P.K., Carlsson, K., Villemot, S., Lyon, S., Gomez, M., Rackauckas, C., Holy, T., Widmann, D., Kelman, T., Karrasch, D., Levitt, A., Riseth, A.N., Lucibello, C., Kwon, C., Barton, D., TagBot, J., Baran, M., Lubin, M., Choudhury, S., Byrne, S., Christ, S., Arakaki, T., Bojesen, T.A., Benneti, Macedo, M.R.G.: JuliaNLSolvers/NLsolve.jl: v4.5.1 (2020). https://doi.org/10.5281/ZENODO.2682214
11. Quarteroni, A., Manzoni, A., Negri, F.: Reduced Basis Methods for Partial Differential Equations: An Introduction. Springer (2015)
12. Träuble, M., Kirchner, C.N., Wittstock, G., Simos, T.E., Maroulis, G.: Nonlinear boundary conditions in simulations of electrochemical experiments using the boundary element method. In: AIP Conference Proceedings, vol. 963, pp. 500–503. AIP (2007)

A Skeletal High-Order Structure Preserving Scheme for Advection-Diffusion Equations

Julien Moatti

Abstract We introduce a nonlinear structure preserving high-order scheme for anisotropic advection-diffusion equations. This scheme, based on Hybrid High-Order methods, can handle general meshes. It also has an entropy structure, and preserves the positivity of the solution. We present some numerical simulations showing that the scheme converges at the expected order, while preserving positivity and long-time behaviour.

Keywords Anisotropic advection-diffusion equations · General meshes · High-order schemes · Structure preserving methods

1 Motivations and Context

We are interested in the discretisation of a linear advection-diffusion equation on general meshes with a high-order scheme. Let Ω be an open, bounded, connected polytopal subset of $\mathbb{R}^d$, $d \in \{2, 3\}$. We consider the following problem with homogeneous Neumann boundary conditions: find $u : \mathbb{R}_+ \times \Omega \to \mathbb{R}$ solution to

$$\begin{cases} \partial_t u - \operatorname{div}(\Lambda(\nabla u + u\nabla\phi)) = 0 & \text{in } \mathbb{R}_+ \times \Omega, \\ \Lambda(\nabla u + u\nabla\phi) \cdot n = 0 & \text{on } \mathbb{R}_+ \times \partial\Omega, \\ u(0, \cdot) = u^{in} & \text{in } \Omega, \end{cases} \tag{1}$$

where n is the unit normal vector to $\partial\Omega$ pointing outwards from Ω. We assume that the data satisfy: (i) $\Lambda \in L^\infty(\Omega; \mathbb{R}^{d\times d})$ is a uniformly elliptic diffusion tensor: there exists $\lambda_\flat > 0$ such that, for a.e. x in Ω, $\Lambda(x)\xi \cdot \xi \geq \lambda_\flat|\xi|^2$ for all $\xi \in \mathbb{R}^d$; (ii) $\phi \in C^1(\overline{\Omega})$ is a regular potential; (iii) $u^{in} \in L^1(\Omega)$ is a non-negative initial datum, such that $\int_\Omega u^{in} \log\left(u^{in}\right) < \infty$. The solutions to (1) enjoy some specific and well-known properties. First the mass is preserved along time, i.e. for almost every $t > 0$,

J. Moatti (✉)
Inria, Laboratoire Paul Painlevé, Univ. Lille, CNRS, UMR 8524, F-59000 Lille, France
e-mail: julien.moatti@inria.fr

E. Franck et al. (eds.), *Finite Volumes for Complex Applications X—Volume 1, Elliptic and Parabolic Problems*, Springer Proceedings in Mathematics & Statistics 432,
https://doi.org/10.1007/978-3-031-40864-9_29

$\int_\Omega u(t) = \int_\Omega u^{in} = M$ where $M > 0$ is the initial mass. Second, the solution is positive for $t > 0$. Last, the solution has a specific long-time behaviour: it converges exponentially fast when $t \to \infty$ towards the thermal equilibrium u^∞, solution to the stationary problem associated to (1), defined as $u^\infty = \frac{M}{\int_\Omega \mathrm{e}^{-\phi}} \mathrm{e}^{-\phi}$.

In order to get a reliable numerical approximation of such problems, one has to preserve these structural properties at the discrete level. It is well-known that two-point finite volume methods are structure preserving (see [2] for the long-time behaviour), but these methods can only be used for isotropic tensors on meshes satisfying some orthogonality conditions. On the other hand, finite volume methods (using auxiliary unknowns) for anisotropic problems on general meshes were introduced in the past twenty years, but none of these linear methods preserve the positivity of the solutions (see [7]). A possible alternative was proposed in [1], with the introduction and analysis of a nonlinear positivity preserving Vertex Approximate Gradient VAG scheme. Following these ideas, a nonlinear Hybrid Finite Volume (HFV) scheme was designed in [3].

All the schemes discussed above are at most of order two in space (in L^2 norm). The aim of this paper is to introduce a high-order scheme preserving the three structural properties discussed above. Since the HFV method coincides with the low-order version of the Hybrid High-Order (HHO) scheme introduced in [5], we propose an HHO generalisation of the scheme introduced in [3]. Numerical results indicate that this scheme offers a better efficiency in terms of computational cost than low order schemes.

2 Discrete Setting and Scheme

2.1 Mesh

We define a discretisation of Ω as a pair $\mathcal{D} = (\mathcal{M}, \mathcal{E})$, where:

- the mesh $\mathcal{M}$ is a partition of Ω into *cells*, i.e., a finite family of nonempty disjoint open polytopal subsets K of Ω such that $\overline{\Omega} = \bigcup_{K\in\mathcal{M}} \overline{K}$,
- the set of faces $\mathcal{E}$ is a partition of the mesh skeleton $\bigcup_{K\in\mathcal{M}} \partial K$ into *faces* σ which are subsets contained in hyperplanes of $\overline{\Omega}$. We denote by $\mathcal{E}_K$ the set of faces of the cell K, and we define $n_{K,\sigma} \in \mathbb{R}^d$ as the unit normal vector to σ pointing outwards from K.

The diameter of a subset $X \subset \overline{\Omega}$ is denoted by $h_X = \sup\{|x - y| \mid (x, y) \in X^2\}$. We define the mesh size of $\mathcal{D}$ as $h_\mathcal{D} = \sup\{h_K \mid K \in \mathcal{M}\}$. We refer to [6, Sect. 1.1] for more detailed statements about the mesh and its regularity.

2.2 Polynomials, Discrete Unknowns and Discrete Operators

In the following, k is a fixed non-negative integer. First, we introduce polynomial spaces on a subset $X \subset \overline{\Omega}$: $\mathbb{P}^k(X)$ and $\mathbb{P}^k(X)^d$ denote respectively the spaces of polynomial functions $X \to \mathbb{R}$ and polynomial vector fields $X \to \mathbb{R}^d$ of degree at most k. Given $Y \subset \overline{X}$, we also define the L^2-projector $\Pi_Y^k : C^0(\overline{X}) \to \mathbb{P}^k(Y)$ by the relation $\forall w \in \mathbb{P}^k(Y)$, $\int_Y \Pi_Y^k(v)w = \int_Y vw$.

We now introduce the set of discrete unknowns corresponding to the mixed-order HHO method [4, 6], with face unknowns of degree k and (enriched) cells unknowns of degree $k+1$:

$$\underline{V}_{\mathcal{D}}^{k,k+1} = \left\{ \underline{v}_{\mathcal{D}} = \big((v_K)_{K\in\mathcal{M}}, (v_\sigma)_{\sigma\in\mathcal{E}}\big) \,\middle|\, \begin{array}{ll} \forall K \in \mathcal{M}, & v_K \in \mathbb{P}^{k+1}(K) \\ \forall \sigma \in \mathcal{E}, & v_\sigma \in \mathbb{P}^k(\sigma) \end{array} \right\}.$$

Given a cell $K \in \mathcal{M}$, we let $\underline{V}_K^{k,k+1} = \mathbb{P}^{k+1}(K) \times \prod_{\sigma\in\mathcal{E}_K} \mathbb{P}^k(\sigma)$ be the restriction of $\underline{V}_{\mathcal{D}}^{k,k+1}$ to K, and for any generic discrete unknown $\underline{v}_{\mathcal{D}} \in \underline{V}_{\mathcal{D}}^{k,k+1}$ we denote by $\underline{v}_K = \big(v_K, (v_\sigma)_{\sigma\in\mathcal{E}_K}\big) \in \underline{V}_K^{k,k+1}$ its local restriction to the cell K. Given any $\underline{v}_{\mathcal{D}} \in \underline{V}_{\mathcal{D}}^{k,k+1}$, we associate two piecewise polynomial functions $v_{\mathcal{M}} : \Omega \to \mathbb{R}$ and $v_{\mathcal{E}} : \bigcup_{K\in\mathcal{M}} \partial K \to \mathbb{R}$ such that

$$v_{\mathcal{M}|K} = v_K \text{ for all } K \in \mathcal{M} \text{ and } v_{\mathcal{E}|\sigma} = v_\sigma \text{ for all } \sigma \in \mathcal{E}.$$

We also introduce $\underline{1}_{\mathcal{D}} \in \underline{V}_{\mathcal{D}}^{k,k+1}$ the discrete element such that $1_K = 1$ for any cell $K \in \mathcal{M}$ and $1_\sigma = 1$ for any face $\sigma \in \mathcal{E}$.

Now, given a cell $K \in \mathcal{M}$, we define a local *discrete gradient operator* $G_K^k : \underline{V}_K^{k,k+1} \to \mathbb{P}^k(K)^d$ such that, for any $\underline{v}_K \in \underline{V}_K^{k,k+1}$, $G_K^k(\underline{v}_K)$ satisfies

$$\int_K G_K^k(\underline{v}_K)\cdot\tau = \int_K \nabla v_K \cdot \tau + \sum_{\sigma\in\mathcal{E}_K} \int_\sigma (v_\sigma - v_K)\tau \cdot n_{K,\sigma} \quad \forall \tau \in \mathbb{P}^k(K)^d. \tag{2}$$

For any face $\sigma \in \mathcal{E}_K$, we also define the *jump operator* $J_{K,\sigma} : \underline{V}_K^{k,k+1} \to \mathbb{P}^k(\sigma)$ by

$$J_{K,\sigma}(\underline{v}_K) = \Pi_\sigma^k(v_K) - v_\sigma. \tag{3}$$

2.3 Scheme

Following the ideas from [1, 3] our scheme relies on a nonlinear reformulation of Problem (1). To do so, we introduce the logarithm potential $\ell = \log(u)$ and the quasi-Fermi potential $w = \ell + \phi$. At least formally, one has the following relation:

$$\nabla u + u\nabla\phi = u\nabla\left(\log(u) + \phi\right) = \mathrm{e}^\ell\,\nabla w. \tag{4}$$

The scheme relies on this formulation. We will *discretise the potentials as polynomials*, i.e. approximate ℓ and w as discrete unknowns in $\underline{V}_{\mathcal{D}}^{k,k+1}$. Then, mimicking the relation $u = \mathrm{e}^{\ell}$, we will reconstruct the density thus ensuring its positivity. Therefore, a solution $\left(\underline{\ell}_{\mathcal{D}}^{n}\right)_{n\geq 1}$ to the scheme (9) corresponds to an approximation of the logarithms of the solution u (density).

More specifically, for a given discretisation $\underline{\ell}_{\mathcal{D}} \in \underline{V}_{\mathcal{D}}^{k,k+1}$ of the potential ℓ, one associates a discrete density $\underset{\sim}{u}_{\mathcal{D}} = (u_{\mathcal{M}}, u_{\mathcal{E}})$ defined as a pair of piecewise smooth functions where $u_{\mathcal{M}} : \Omega \to \mathbb{R}$ corresponds to the cells unknowns and $u_{\mathcal{E}} : \bigcup_{K\in\mathcal{M}} \partial K \to \mathbb{R}$ corresponds to the face unknowns, defined as

$$u_{\mathcal{M}} = \exp(\ell_{\mathcal{M}}) \text{ and } u_{\mathcal{E}} = \exp(\ell_{\mathcal{E}}). \tag{5}$$

Note that a discrete density $\underset{\sim}{u}_{\mathcal{D}}$ is not a collection of polynomials (which is highlighted by the use of the wave under u), but it enjoys positivity, both on cells and faces, since it is defined as the exponential of real functions.

Our scheme is based on local contributions on cells, split into a consistent term and a stabilisation term. Given $K \in \mathcal{M}$ and $\eta_l > 0$, the classical discrete counterpart of $(w, v) \mapsto \int_K \Lambda \nabla w \cdot \nabla v$ is the bilinear form (see [4, Sect. 3.2.1])

$$a_K : (\underline{w}_K, \underline{v}_K) \mapsto \int_K \Lambda G_K^k(\underline{w}_K) \cdot G_K^k(\underline{v}_K) + \eta_l \sum_{\sigma\in\mathcal{E}_K} \frac{\Lambda_{K\sigma}}{h_\sigma} \int_\sigma J_{K,\sigma}(\underline{w}_K) J_{K,\sigma}(\underline{v}_K),$$

where $\Lambda_{K\sigma} = \|\Lambda_{|K} n_{K\sigma} \cdot n_{K\sigma}\|_{L^\infty(\sigma)}$. Similarly, given $\eta_{nl} > 0$, we define a local discretisation of $(\ell, w, v) \mapsto \int_K \mathrm{e}^{\ell} \Lambda \nabla w \cdot \nabla v$ as a sum of nonlinear consistent (6a) and stabilisation (6b) contributions:

$$\mathcal{C}_K(\underline{\ell}_K, \underline{w}_K, \underline{v}_K) = \int_K \mathrm{e}^{\ell_K} \Lambda G_K^k(\underline{w}_K) \cdot G_K^k(\underline{v}_K), \tag{6a}$$

$$\mathcal{S}_K(\underline{\ell}_K, \underline{w}_K, \underline{v}_K) = \eta_{nl} \sum_{\sigma\in\mathcal{E}_K} \frac{\Lambda_{K\sigma}}{h_\sigma} \int_\sigma \frac{\mathrm{e}^{\Pi_\sigma^k(\ell_K)} + \mathrm{e}^{\ell_\sigma}}{2} J_{K,\sigma}(\underline{w}_K) J_{K,\sigma}(\underline{v}_K). \tag{6b}$$

We can now define a local application $\mathcal{T}_K : \underline{V}_K^{k,k+1} \times \underline{V}_K^{k,k+1} \times \underline{V}_K^{k,k+1} \to \mathbb{R}$ by

$$\mathcal{T}_K(\underline{\ell}_K, \underline{w}_K, \underline{v}_K) = \mathcal{C}_K(\underline{\ell}_K, \underline{w}_K, \underline{v}_K) + \mathcal{S}_K(\underline{\ell}_K, \underline{w}_K, \underline{v}_K) + \varepsilon h_K^{k+2} a_K(\underline{w}_K, \underline{v}_K), \tag{7}$$

where ε is a non-negative parameter. At the global level, we define $\mathcal{T}_{\mathcal{D}} : \underline{V}_{\mathcal{D}}^{k,k+1} \times \underline{V}_{\mathcal{D}}^{k,k+1} \times \underline{V}_{\mathcal{D}}^{k,k+1} \to \mathbb{R}$ by summing the local contributions:

$$\mathcal{T}_{\mathcal{D}}(\underline{\ell}_{\mathcal{D}}, \underline{w}_{\mathcal{D}}, \underline{v}_{\mathcal{D}}) = \sum_{K\in\mathcal{M}} \mathcal{T}_K(\underline{\ell}_K, \underline{w}_K, \underline{v}_K). \tag{8}$$

We let $\underline{\phi}_{\mathcal{D}} \in \underline{V}_{\mathcal{D}}^{k,k+1}$ be the interpolate of ϕ: for any $K \in \mathcal{M}$, $\phi_K = \Pi_K^{k+1}(\phi)$ and for all $\sigma \in \mathcal{E}$, $\phi_\sigma = \Pi_\sigma^k(\phi)$. Now, using a backward Euler discretisation in time with time step $\Delta t > 0$, we introduce the following scheme for (1): find $\left(\underline{\ell}_{\mathcal{D}}^{n}\right)_{n\geq 1} \in \left(\underline{V}_{\mathcal{D}}^{k,k+1}\right)^{\mathbb{N}^*}$ such that

$$\begin{cases} \displaystyle\int_{\Omega} \frac{u_{\mathcal{M}}^{n+1} - u_{\mathcal{M}}^{n}}{\Delta t} v_{\mathcal{M}} = -\mathcal{T}_{\mathcal{D}}(\underline{\ell}_{\mathcal{D}}^{n+1}, \underline{\ell}_{\mathcal{D}}^{n+1} + \underline{\phi}_{\mathcal{D}}, \underline{v}_{\mathcal{D}}) & \forall \underline{v}_{\mathcal{D}} \in \underline{V}_{\mathcal{D}}^{k,k+1}, \quad \text{(9a)} \\ u_K^0 = u^{in}{}_{|K} & \forall K \in \mathcal{M}. \quad \text{(9b)} \end{cases}$$

Given a solution $\left(\underline{\ell}_{\mathcal{D}}^{n}\right)_{n\geq 1}$ to the scheme (9), as discussed above, we associate a sequence of positive discrete densities $\left(\underset{\sim}{u}_{\mathcal{D}}^{n}\right)_{n\geq 1}$.

Remark 1 (*Parameter* ε) Note that $\mathcal{T}_{\mathcal{D}}$ is to be understood as a discretisation of $(\ell, w, v) \mapsto \int_{\Omega} (\mathrm{e}^{\ell} + \epsilon)\Lambda \nabla w \cdot \nabla v$, with $\epsilon \sim \varepsilon h_{\mathcal{D}}^{k+2}$ a small parameter. The ϵ perturbation is used in order to show the existence result of Proposition 2 and can be seen as a kind of stabilisation. The scaling factor h_K^{k+2} in (7) is used to get the expected order of convergence. In practice, numerical results for $\varepsilon = 1$ and $\varepsilon = 0$ are almost the same. The influence of this term will be investigated in future works.

We define the discrete thermal equilibrium as $\underset{\sim}{u}_{\mathcal{D}}^{\infty} = (\rho\, \mathrm{e}^{-\phi_{\mathcal{M}}}, \rho\, \mathrm{e}^{-\phi_{\mathcal{E}}})$, with $\rho = M / \int_{\Omega} \mathrm{e}^{-\phi_{\mathcal{M}}}$. One can show that $\underset{\sim}{u}_{\mathcal{D}}^{\infty}$ (and the associated logarithm potential $\underline{\ell}_{\mathcal{D}}^{\infty} \in \underline{V}_{\mathcal{D}}^{k,k+1}$) is the only stationary solution to (9) with mass M.

3 Main Features of the Scheme

In this section, we present some results regarding the analysis of the scheme (9). Given $\underline{\ell}_{\mathcal{D}} \in \underline{V}_{\mathcal{D}}^{k,k+1}$ a discrete logarithm, we associate a discrete quasi-Fermi potential defined as $\underline{w}_{\mathcal{D}} = \underline{\ell}_{\mathcal{D}} + \underline{\phi}_{\mathcal{D}} - \log(\rho)\underline{1}_{\mathcal{D}}$. By definition of ρ, one has $w_{\mathcal{M}} = \log\left(\frac{u_{\mathcal{M}}}{u_{\mathcal{M}}^{\infty}}\right)$. Note that, for any $(\underline{\ell}_{\mathcal{D}}, \underline{v}_{\mathcal{D}}) \in \underline{V}_{\mathcal{D}}^{k,k+1} \times \underline{V}_{\mathcal{D}}^{k,k+1}$, we have $\mathcal{T}_{\mathcal{D}}(\underline{\ell}_{\mathcal{D}}, \underline{\ell}_{\mathcal{D}} + \underline{\phi}_{\mathcal{D}}, \underline{v}_{\mathcal{D}}) = \mathcal{T}_{\mathcal{D}}(\underline{\ell}_{\mathcal{D}}, \underline{w}_{\mathcal{D}}, \underline{v}_{\mathcal{D}})$. We now state our fundamental a priori results.

Proposition 1 (Fundamental a priori relations) *Let* $\left(\underline{\ell}_{\mathcal{D}}^{n}\right)_{n\geq 1}$ *be a solution to the scheme (9), and* $\left(\underset{\sim}{u}_{\mathcal{D}}^{n}\right)_{n\geq 1}$ *be the associated reconstructed discrete density. Then, the following a priori results hold:*

(i) the mass is preserved along time: $\forall n \in \mathbb{N}^*$, $\displaystyle\int_{\Omega} u_{\mathcal{M}}^{n} = \int_{\Omega} u^{in} = M$,

(ii) a discrete entropy/dissipation relation holds: $\forall n \in \mathbb{N}$, $\displaystyle\frac{\mathbb{E}^{n+1} - \mathbb{E}^{n}}{\Delta t} \leq -\mathbb{D}^{n+1}$, where the discrete entropy and dissipation are defined by $\mathbb{E}^n = \int_{\Omega} u_{\mathcal{M}}^{\infty} \Phi_1\left(\frac{u_{\mathcal{M}}^{n}}{u_{\mathcal{M}}^{\infty}}\right)$ and $\mathbb{D}^n = \mathcal{T}_{\mathcal{D}}(\underline{\ell}_{\mathcal{D}}^{n}, \underline{w}_{\mathcal{D}}^{n}, \underline{w}_{\mathcal{D}}^{n}) \geq 0$ with $\Phi_1 : s \mapsto s\log(s) - s + 1$ (and $\Phi_1(0) = 1$).

Proof Using $\underline{1}_{\mathcal{D}}$ as a test function in (9a), alongside with (9b), we get the mass conservation identity (i). To get (ii), we test (9a) with $\underline{w}_{\mathcal{D}}^{n+1}$, and we use the convexity of Φ_1 alongside with the expression of $w_{\mathcal{M}}^{n+1}$.

Note that the previous results hold for any $\varepsilon \geq 0$. Following the ideas of [1, 3], the entropy/dissipation relation should allow one to analyse the long-time behaviour of the discrete solutions and to get convergence results. These aspects will be the topics of future works. We now state an existence result, which holds only for positive ε. The proof follows the strategy used in [3].

Proposition 2 (Existence of solutions) *Assume that the stabilisation parameter ε in (7) is positive. Then, there exists at least one solution $\left(\underline{\ell}^n_{\mathcal{D}}\right)_{n\geq 1}$ to the scheme (9). The associated densities $\left(\underset{\sim}{u}^n_{\mathcal{D}}\right)_{n\geq 1}$ are positive functions.*

4 Numerical Results

The numerical scheme (9) requires to solve a nonlinear system of equations at each time step. To do so, we use a Newton method, with an adaptative time stepping strategy: if the Newton method does not converge, we try to compute the solution for a smaller time step $0.5 \times \Delta t$. If the method converges, we use for the subsequent time step the value $2 \times \Delta t$. The maximal time step allowed is the initial time step. Each time a linear system has to be solved we perform a static condensation (see [6, Appendix B.3.2]) in order to eliminate (locally) the cell unknowns. Note that the local computations are not implemented in parallel, but only sequentially. In the sequel, we use the following stabilisation parameters: $\varepsilon = \eta_{nl} = \eta_l = 1$.

The tests considered below (on $\Omega =]0, 1[^2$) are the same as in [3], to which we refer for more detailed explanations and descriptions. Given a (face) degree k, the scheme (9) will be denoted by nlhho_k, whereas the HFV scheme of [3] will be denoted by nlhfv. Note that nlhho_0 hinges on affine cell unknowns, whereas the cell unknowns of nlhfv are constant: these two schemes hence do not coincide, and nlhho_0 is expected to be more costly.

4.1 Proof of Concept: Convergence Order and Efficiency

Here, we are interested in the convergence of the scheme when $(h_{\mathcal{D}}, \Delta t) \to (0, 0)$. To do so, we set the advective potential and diffusion tensor as $\phi(x, y) = -x$ and $\Lambda = \begin{pmatrix} l_x & 0 \\ 0 & 1 \end{pmatrix}$ for $l_x > 0$. The exact solution is therefore given by

$$u(t, x, y) = C_1 \,\mathrm{e}^{-\alpha t + \frac{x}{2}} \left(2\pi \cos(\pi x) + \sin(\pi x)\right) + 2C_1 \pi \,\mathrm{e}^{x - \frac{1}{2}},$$

where $C_1 > 0$ and $\alpha = l_x \left(\frac{1}{4} + \pi^2\right)$. Note that u^{in} vanishes on $\{x = 1\}$, but for any $t > 0, u(t, \cdot) > 0$. Here, our experiments are performed using $l_x = 1$ and $C_1 = 10^{-1}$. We compute the solution on the time interval $[0, 0.1]$, and we denote by $(\underset{\sim}{u}^n_{\mathcal{D}})_{1\leq n\leq N_f}$ the corresponding discrete density. Then, we compute the relative $L^2_t(L^2_x)$ error on

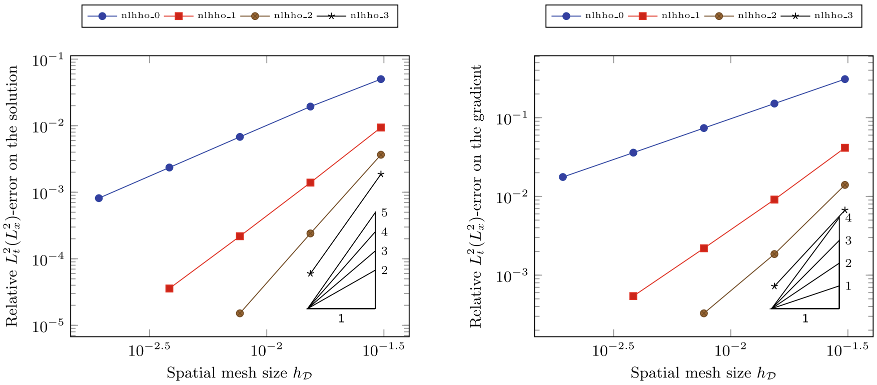


Fig. 1 Accuracy of transient solutions. Relative error on triangular meshes

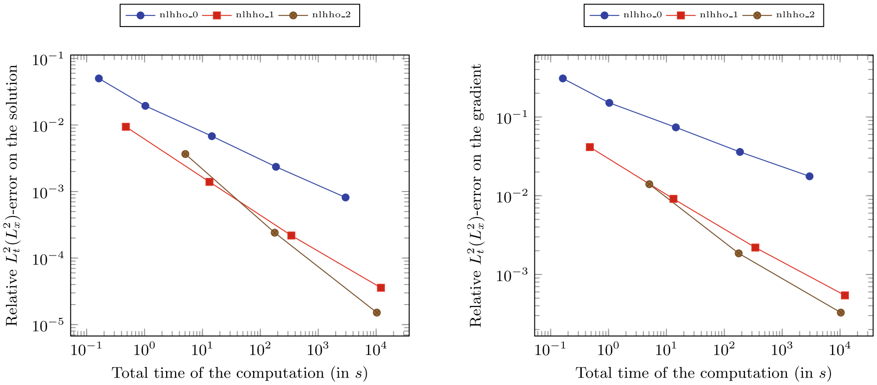


Fig. 2 Accuracy versus computational cost. Relative errors on triangular meshes

the solution and on the gradient of the solution, defined as

$$\frac{\sqrt{\sum_{n=1}^{N_f} \delta t^n \|u_{\mathcal{M}}^n - u(t^n,\cdot)\|_{L^2(\Omega)}^2}}{\|u\|_{L_t^2(L_x^2)}} \text{ and } \frac{\sqrt{\sum_{n=1}^{N_f} \delta t^n \|\mathcal{G}_{\mathcal{M}}(\underset{\sim}{u}_{\mathcal{D}}^n) - \nabla u(t^n,\cdot)\|_{L^2(\Omega)}^2}}{\|\nabla u\|_{L_t^2(L_x^2)}}$$

where $\delta t^n = t^n - t^{n-1}$and the discrete gradient $\mathcal{G}_{\mathcal{M}}(\underset{\sim}{u}_{\mathcal{D}}^n)$ is defined by mimicking the continuous relation $\nabla u = \mathrm{e}^{\ell}\,\nabla\ell$ as a piecewise continuous function satisfying $\mathcal{G}_{\mathcal{M}}(\underset{\sim}{u}_{\mathcal{D}})_{|K} = \exp(\ell_K)\, G_K^k(\underline{\ell}_K)$ on $K \in \mathcal{M}$. The L^2 norms are computed using quadrature formulas of order $2k+5$. Note that, with the chosen definitions, we do not take into account the time $t = 0$. To plot the error graphs, we do simulations on a triangular mesh family $(\mathcal{D}_i)_{1\le i\le 5}$, such that $h_{\mathcal{D}_i}/h_{\mathcal{D}_{i+1}} = 2$. Since the time discretisation is of order one, on the i-th mesh of the family, we use a time step of $\Delta t_i = \Delta t_k / 2^{(i-1)\times(k+2)}$, where $\Delta t_k = 0.05/2^{k+2}$ is the initial time step used on $\mathcal{D}_1$.

In Fig. 1, we see that the scheme, for face unknowns of degree k, converges at order $k+1$ in energy norm and $k+2$ in L^2 norm of the density. In Fig. 2, we plot the errors as functions of the computing time to get the solution. It is remarkable to see that, even with a low order discretisation in time, significant efficiency gains can be reached by using a high value of k. The gain should be even bigger by parallelising the local computations. Of course, the use of higher order time-stepping methods should also lead to significant gains, and this should be investigated in future works. However, the way of getting the entropy dissipation relation is currently unclear for such time discretisations.

4.2 Discrete Long-Time Behaviour

We are now interested in the long-time behaviour of discrete solutions. We use the same test-case as before, but with an anisotropic tensor: we set $l_x = 10^{-2}$. We compute the solution on the time interval [0, 350], with $\Delta t = 10^{-1}$, on two Kershaw meshes of sizes 0.02 and 0.006. In Fig. 3, we show the evolution along time of the L^1 distance between $\underset{\sim}{u}^n_{\mathcal{D}}$ and $u^\infty = 2C_1\pi\, e^{x-\frac{1}{2}}$ computed as $\int_\Omega |u^n_{\mathcal{M}} - u^\infty|$. We observe the exponential convergence towards the steady-state, until some precision is reached. The rates of convergence are similar to the exact one (α), and do not depend on the size of the mesh.

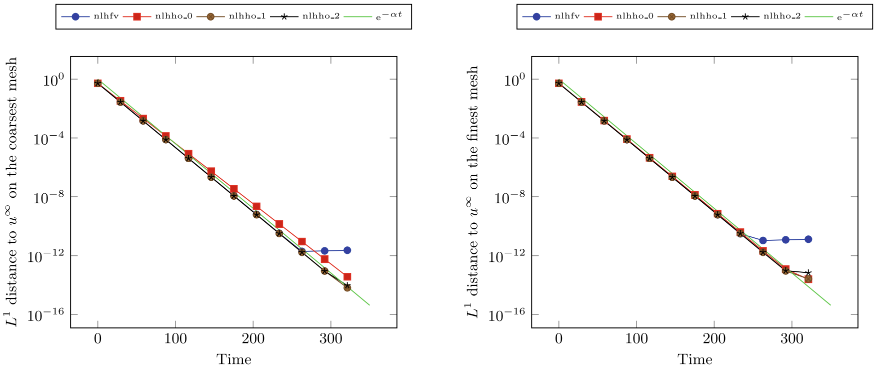


Fig. 3 Long-time behaviour of discrete solutions. Comparison of the long-time behaviour on Kershaw meshes for $T_f = 350$ and $\Delta t = 0.1$

4.3 Positivity

This last section is dedicated to assessing the discrete positivity preservation. We set the advection field as $\phi(x, y) = -\big((x - 0.4)^2 + (y - 0.6)^2\big)$ and the diffusion tensor as $\Lambda = \begin{pmatrix} 0.8 & 0 \\ 0 & 1 \end{pmatrix}$.

Table 1 Positivity of discrete solutions

	Computing time	#resol	Mincells	Minfaces	MincellQN	MinfaceQN
nlhfv	1.77e+01	175	9.93e-04	7.36e-04	9.93e-04	7.36e-04
HMM	2.20e-01	50	−5e-03	−7.74e-02	−5e-03	−7.74e-02
nlhho_0	7.17e+01	224	1.00e-03	1.01-03	2.41e-06	1.01e-03
nlhho_1	4.13e+02	248	6.65e-04	2.05e-05	1.78e-04	3.57e-08
nlhho_2	1.45e+03	251	9.50e-04	5.99e-04	2.67e-07	1.06e-05
nlhho_3	3.87e+03	254	9.85e-04	8.58e-04	1.10e-05	1.79e-05

For the initial data, we take $u^{in} = 10^{-3}\,\mathbb{1}_B + \mathbb{1}_{\Omega\setminus B}$, where B is the Euclidean ball $\big\{(x, y) \in \mathbb{R}^2 \mid (x - 0.5)^2 + (y - 0.5)^2 \leq 0.2^2\big\}$. We perform simulations on the time interval $[0, 5.10^{-4}]$ with $\Delta t = 10^{-5}$ on a refined tilted hexagonal-dominant mesh (4192 cells). In Table 1, we show the minimal values reached by the schemes. The values of "mincells" are defined as $\min\{\frac{1}{|K|}\int_K u^n_{\mathcal{M}} \mid K \in \mathcal{M}, 1 \leq n \leq N_f\}$, whereas "mincellQN" are the minimal values taken by the densities on the cell quadrature nodes. Analogous definitions hold for the faces. The values of "#resol" correspond to the number of linear systems solved during the computation. Note that the size of these systems depends on the value of k. The HMM scheme is a linear one (see [3]), therefore only one LU factorisation was performed to compute the solution, which has 90 (resp. 503) negative cell (resp. face) unknowns.

Acknowledgements The author thanks the anonymous reviewers for their remarks and suggestions, as well as Claire Chainais-Hillairet, Maxime Herda and Simon Lemaire for fruitful discussions about this work. The author acknowledges support by the Labex CEMPI (ANR-11-LABX-0007-01).

References

1. 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)
2. 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. **40**(4), 2473–2504 (2020)

3. Chainais-Hillairet, C., Herda, M., Lemaire, S., Moatti, J.: Long-time behaviour of hybrid finite volume schemes for advection-diffusion equations: linear and nonlinear approaches. Numer. Math. **151**(4), 963–1016 (2022)
4. Cicuttin, M., Ern, A., Pignet, N.: Hybrid High-order Methods. A Primer with Applications to Solid Mechanics. Springer, Cham (2021)
5. 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)
6. Di Pietro, D.A., Droniou, J.: The Hybrid High-order Method for Polytopal Meshes. Design, Analysis, and Applications. Springer, Cham (2020)
7. Droniou, J.: Finite volume schemes for diffusion equations: introduction to and review of modern methods. Math. Models Methods Appl. Sci. **24**, 1575–1619 (2014)

Automatic Solid Reconstruction from 3-D Points Set for Flow Simulation via an Immersed Boundary Method

Gabriel F. Narváez, Martin Ferrand, Thomas Fonty, and Sofiane Benhamadouche

Abstract Dealing with complex geometries for industrial applications is challenging in computational fluid dynamic workflows. Current developments in scan devices offer the possibility to represent very complex solid geometries in fluid dynamic solvers. This paper proposes a novel approach for reconstructing solid geometry from 3-D scans and flow simulation. Based on a 3-D point cloud, the approach automatically reconstructs the solid surface by including local solid planes in any convex computational cell. An immersed boundary method is then used to impose appropriate boundary conditions on the solid surfaces in the co-located finite volume context. The present approach avoids the complex and time-consuming manual/assisted meshing typical of body-fitted mesh workflows while showing satisfactory robustness and accuracy.

Keywords 3-D solid scan · Point cloud · Fluid dynamic simulation · Immersed boundary method

1 Introduction

Mesh paradigm for representing solids in contact with fluid can be classified into the body-fitted methods (BFMs) and immersed boundary methods (IBMs). The BFMs fit the mesh topology on the solid surface while the solid region is not included in the computational mesh; thus, the boundary conditions can be directly imposed on the mesh boundaries. In IBMs, the solid is embedded inside the fluid mesh, and the wall boundary conditions are forced inside the mesh. Consequently, IBMs can avoid the time-consuming manual/assisted meshing step since the mesh can be as simple

G. F. Narváez · M. Ferrand
CEREA, École des Ponts, EDF R&D, Champs-sur-Marne, France

G. F. Narváez (✉) · M. Ferrand · T. Fonty · S. Benhamadouche
EDF R&D, 6, quai Watier, 78400 Chatou, France
e-mail: gabriel.narvaez-campo@enpc.fr

E. Franck et al. (eds.), *Finite Volumes for Complex Applications X—Volume 1, Elliptic and Parabolic Problems*, Springer Proceedings in Mathematics & Statistics 432,
https://doi.org/10.1007/978-3-031-40864-9_30

as a Cartesian mesh. Complex solid geometries can be identified by 3-D scanning techniques, which yield a set of points in the object surface (point cloud), from which the solid surface can be recovered through a process so-called *surface reconstruction*. The reconstruction can be performed by interpolating a parametric surface [1, 6] or implicitly through a signed scalar function [2, 8]. Recently, [2] developed a scalar reconstruction for flow simulation in a finite element fluid solver. In the present work, we propose a novel automatic solid reconstruction approach based on a 3-D point cloud for a co-located finite volumes approach for fluid simulation.

2 Immersed Boundary Method

The set of governing Eq. 1 for a viscous flow is

$$\begin{cases} \dfrac{\partial \rho}{\partial t} + \operatorname{div}\left(\rho \underline{u}\right) = 0, \\ \dfrac{\partial (\rho \underline{u})}{\partial t} + \underline{\operatorname{div}}(\underline{u} \otimes \rho \underline{u}) = -\underline{\nabla} p + \rho \underline{g} + \underline{\operatorname{div}}\left(\underline{\underline{\tau}}\right), \end{cases} \tag{1}$$

where $\underline{u}$ is the velocity, p the pressure and ρ the mass density of the fluid. Initial and boundary conditions must supplement this system on the boundary $\partial\Omega$ for velocity and pressure. Function of the dynamic molecular viscosity μ, the volume viscosity κ and the strain rate tensor $\underline{\underline{\mathrm{S}}} = \dfrac{1}{2}\left(\underline{\underline{\nabla}}\, \underline{u} + \underline{\underline{\nabla}}\, \underline{u}^T\right)$, the viscous tensor for a Newtonian fluid is $\underline{\underline{\tau}} = 2\mu \underline{\underline{\mathrm{S}}} + \left(\kappa - \frac{2}{3}\mu\right) tr\left(\underline{\underline{\mathrm{S}}}\right) \underline{\underline{1}}$. The discretization uses a semi-implicit co-located finite volume scheme with an incremental pressure-correction algorithm [3, 12]. The current developments are implemented in EDF open-source CFD software code_saturne (available in https://www.code-saturne.org/cms/web/). A specific finite volume control (cell) c can be composed of fluid boundaries $\partial\Omega_c^\phi$ and solid boundaries $\partial\Omega_c^w$ (Fig. 1). Our approach considers that only one solid face cuts each cell, dividing the cell in a volume occupied by fluid Ω_c^ϕ and the other by solid Ω_c^s.

The pressure gradient at cell c with immersed boundary is

$$\Omega_c^\phi \underline{\nabla}_c p = \int_{\partial\Omega_c^\phi \cup \partial\Omega_c^w} p \mathrm{d}\underline{S} = \sum_{f \in \mathcal{F}_c} p_f \underline{S}_{c>f}^\phi + p_w \underline{S}_c^w \tag{2}$$

where $\underline{S}_{c>f}^\phi$ (resp. $\underline{S}_c^w$) is the outwarding normal of the fluid (resp. immersed wall) face, and $p_w = p_c$, to impose an homogeneous Neumann boundary condition.

The velocity boundary condition, namely the no-slip Dirichlet boundary condition, is transformed into a Neumann boundary condition where the wall shear stress is an imposed function of the local flow velocities. The distance $h_{w/c}$ of the cell centre to the wall is given by $h_{w/c} = \left(\underline{x}_c^w - \underline{x}_c^\phi\right) \cdot \underline{n}_c^w$, where $\underline{x}_c^w$ is the immersed solid face

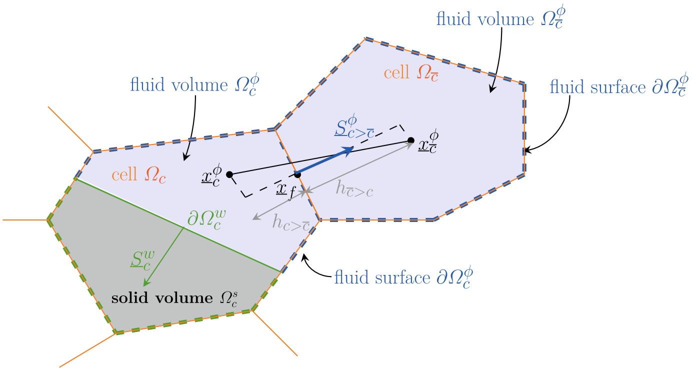

fluid volume $\Omega_{\bar{c}}^{\phi}$
cell $\Omega_{\bar{c}}$
fluid surface $\partial\Omega_{\bar{c}}^{\phi}$
fluid volume Ω_c^{ϕ}
$\underline{S}_{c>\bar{c}}^{\phi}$
$\underline{x}_{\bar{c}}^{\phi}$
cell Ω_c
$\underline{x}_c^{\phi}$
$\underline{x}_f$
$h_{\bar{c}>c}$
$\partial\Omega_c^w$
$h_{c>\bar{c}}$
$\underline{S}_c^w$
solid volume Ω_c^s
fluid surface $\partial\Omega_c^{\phi}$

Fig. 1 2-D Sketch of a cell Ω_c with a solid volume Ω_c^s and connected to the cell $\Omega_{\bar{c}}$ through a fluid face $\underline{S}_{c>\bar{c}}^{\phi}$ separated by a distance $h_{c>\bar{c}}$ from the fluid cell centre $\underline{x}_c^{\phi}$. The solid wall $\partial\Omega_c^w$ separates the fluid volume Ω_c^{ϕ} from the solid volume Ω_c^s

centre, $\underline{x}_c^{\phi}$ is the fluid cell centre, and $\underline{n}_c^w$, the normal of the solid face pointing outward the fluid. The shear stress estimation on the wall, in laminar or DNS turbulent simulations of a Newtonian incompressible flow, can be estimated by a two-point flux approximation $\underline{\underline{\tau}}_c^w \cdot \underline{S}_c^w = \mu \dfrac{\underline{u}_w - \underline{u}_c}{h_{w/c}} S_c^w$, where the no-slip condition $\underline{u}_w = \underline{0}$, for fixed wall can be applied and the velocity at the cell $\underline{u}_c$ is implicitly applied in time.

3 Solid Reconstruction from 3-D Scan Point Cloud

We propose a straightforward *hybrid surface reconstruction* which uses the scan points in a cell to reconstruct a plane that minimises the mean square error (step ① of Fig. 2). The wall plane is used to recompute the cell/face geometrical quantities, such as the volume/surface and centre of gravity of the fluid part (steps ②, ③, and ④ of Fig. 2). Then, the distance to the wall plane can be computed to impose the wall boundary conditions (steps ⑤ and ⑥ of Fig. 2). It should be noted that the algorithm works for any convex cell and it converges to the actual wall area as the cell size is reduced (Fig. 3).

For a given local point cloud in the cell Ω_c, a plane $\underline{n}_c \cdot \left(\underline{x} - \underline{x}_{p_c}\right) = 0$ is fitted by least square adjustment, relative to the centroid of the point cloud ($\underline{x}_{p_c}$). The wall-normal vector is normalised $\underline{n}_c^w = \frac{\underline{n}_c}{\|\underline{n}_c\|}$, but the orientation has to be set towards the solid region.

Solid plane orientation: The plane orientation definition is essential to identify solid and fluid regions. Then a dedicated algorithm has been developed and integrated to code_saturne to find the solid region. The algorithm numerically reproduces what

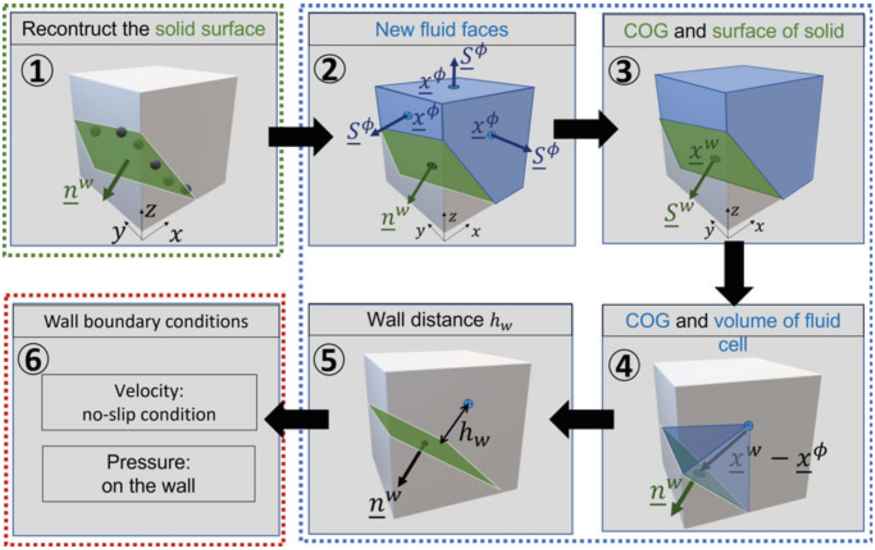


Fig. 2 Flux chart of the algorithm to update fluid part of cell cut by a solid wall and impose boundary conditions. Example with a hexahedral cell

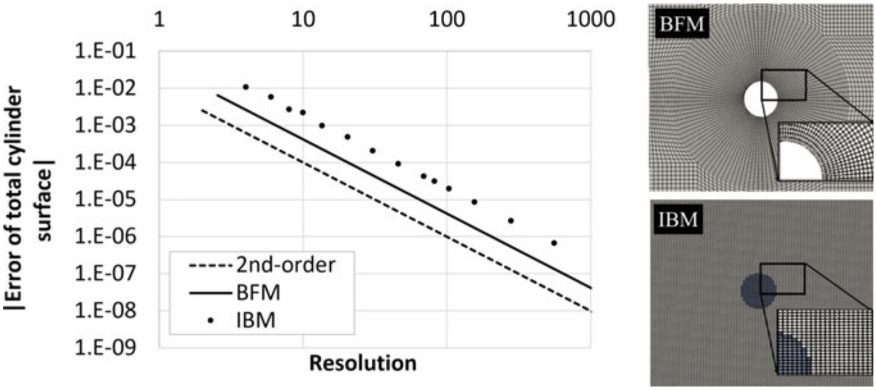


Fig. 3 Convergence of the estimated wall area of current immersed boundary method (IBM), with dense point cloud, and body-fitted method (BFM)

the scanner does: from a point source (where the scanner was located), numerical rays are traced in all directions (like the Discrete Ordinate Method to solve the radiative transfer equation). It corresponds to a steady pure convection equation on ϕ, whose continuous and discrete formulations, respectively, are:

$$\underline{\nabla}\phi \cdot \left(\underline{x} - \underline{x}_0\right) = (1-\phi)\,\delta_{\underline{x}_0}\left(\underline{x}\right) \tag{3}$$

$$\sum_{f \in \mathcal{F}_c} \left[\phi_f - \phi_c\right] \left(\underline{x}_f - \underline{x}_0\right) \cdot \underline{S}^{\phi,0}_{c>f} = (1-\phi_c)\,\Omega_c\,\mathbb{1}_{\underline{x}_0 \in c}, \tag{4}$$

with $\underline{x}_0$ the source position and the indicator function $\mathbb{1}_{\underline{x}_0 \in c} = \int_{\Omega_c} \delta_{\underline{x}_0}\left(\underline{x}\right) \mathrm{d}\Omega = 1$ if the cell contains the source, and zero otherwise. The scalar at the faces ϕ_f is estimated by an upwind scheme. Cells with more than three scan points are penalised and stop the rays (i.e., are fully solid cells with $\underline{S}^{\phi,0}_{c>f} = \underline{0}$). Thus, ϕ fills the fluid domain. As many loops as available scans are performed, storing the maximum of ϕ. If it is bigger than one-half, the cell is considered fluid. It means that $\underline{\nabla}_c\phi$ points towards the fluid region in solid cells in contact with the fluid. Given that $\underline{n}^w_c$ should point outward the fluid, it is inverted if $\underline{\nabla}_c\phi \cdot \underline{n}^w_c > 0$.

Note that $\underline{S}^{\phi,0}_{c>f}$ is the fluid face first guess to run the orientation algorithm. However, considering that cells with a solid plane (i.e., with scan points) are split into solid and fluid parts, the fluid cell and face geometric quantities are then updated using the deduced solid plane information.

Fluid cell quantities: The fluid and wall faces bound the fluid cell. The fluid volume Ω^ϕ_c and centre of gravity $\underline{x}^\phi_c$ are computed by integrating the resulting fluid cell c (polyhedron-shaped) composed by discrete pyramids with apex $\underline{G}_c$ and a base face $\underline{S}^\phi_{c>\bar{c}}$ or $\underline{S}^w_c$, where $\underline{G}_c$ is estimated by the sum of faces centres ($\underline{x}^w_c$, $\underline{x}^\phi_f$) weighted by the face surfaces (S^w_c, $S^\phi_{f_{c|\bar{c}}}$). **Fluid face quantities:** The wall unit vector $\underline{n}^w$ and a point on the wall ($\underline{x}_p$) are employed to identify if a vertex of the cell is inside the fluid or the solid. The algorithm creates a new fluid face by removing the solid vertices from the original face and, if required, adding fluid-solid interface vertices (Fig. 4). Planes of neighbour cells $c - \bar{c}$ with a different intersection on the joint face create a discontinuity (Fig. 5), which is overcome by setting the minimum fluid surface value $\underline{S}^\phi_{c>f}$, with the corresponding centre of gravity $\underline{x}_f$, and then use it in the following

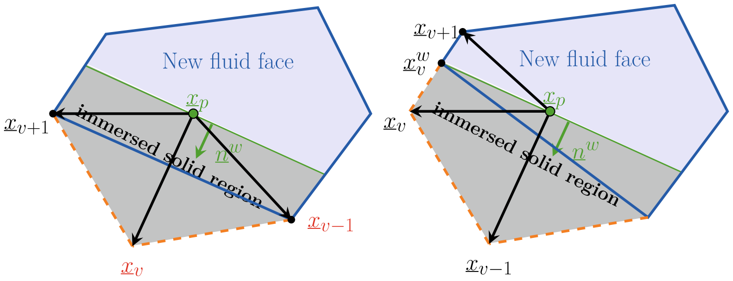


Fig. 4 Based on the sign of the inner product of vector $\left(\underline{x}_v - \underline{x}_p\right) \cdot \underline{n}^w$, the algorithm loops over the original face vertices (clockwise oriented) to obtain the vertices of the new fluid face. **Left:** solid vertex with no phase transition between neighbours is removed. **Right:** solid to fluid change, the solid vertex is projected onto the wall along the edge. The vertex is maintained if it is in the fluid or on the wall

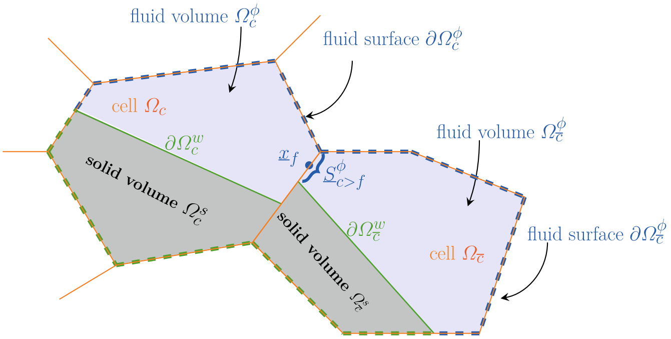

fluid volume Ω_c^ϕ
fluid surface $\partial\Omega_c^\phi$
cell Ω_c
$\partial\Omega_c^w$
solid volume Ω_c^s
$\underline{x}_f$
$\underline{S}^\phi_{c>f}$
fluid volume $\Omega_{\overline{c}}^\phi$
$\partial\Omega_{\overline{c}}^w$
solid volume $\Omega_{\overline{c}}^s$
cell $\Omega_{\overline{c}}$
fluid surface $\partial\Omega_{\overline{c}}^\phi$

Fig. 5 2-D sketch of possible discontinuity due to local wall reconstruction

correction. **Solid face geometric quantities:** Previously identified wall vertices $\underline{x}_v^w$ are employed for computing solid face quantities. Then, we propose the following correction of the solid face that preserves the fluid quantities previously computed and deals with the discontinuity presented in Fig. 5.

$$\begin{aligned} \underline{S}_c^w &= -\sum_{f\in\mathcal{F}_c} \underline{S}^\phi_{c>f} \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad , \\ \underline{x}_c^w &= \underline{x}_c^\phi + \left[\Omega_c^\phi \underline{\underline{\mathrm{I}}} - \sum_{f\in\mathcal{F}_c} \left(\underline{x}_f^\phi - \underline{x}_c^\phi \right) \otimes \underline{S}^\phi_{c>f} \right] \cdot \frac{\underline{S}_c^w}{(S_c^w)^2} \ . \end{aligned} \tag{5}$$

4 Study Case

The flow around a circular cylinder is employed to validate the aforementioned implementations, in a domain $(L_x \times L_y \times L_z) = 50D \times 30D \times 0.625D$, with uniform stream-wise velocity inlet U, pressure imposed outlet and free-slip lateral boundaries. The cylinder centre is located $20D$ from the inlet and $15D$ from the lateral boundaries. We consider the body-fitted (BFM) and immersed boundary methods (IBM), where BFM is applied in a cylindrical high-refined mesh and used to ver-

Table 1 Uniform Cartesian meshes for IBM simulations with a 3-D point cloud size of $N_p = 144\ 000$ on the cylinder surface. The point cloud density ensures at least 16 points per cell

Mesh	$n_x \times n_y$	$\Delta x/D$	$\Delta y/D$	$\#cells/D$
Coarse	401 × 241	0.2	0.2	8
Medium	801 × 481	0.1	0.1	16
Fine	1601 × 961	0.05	0.05	32

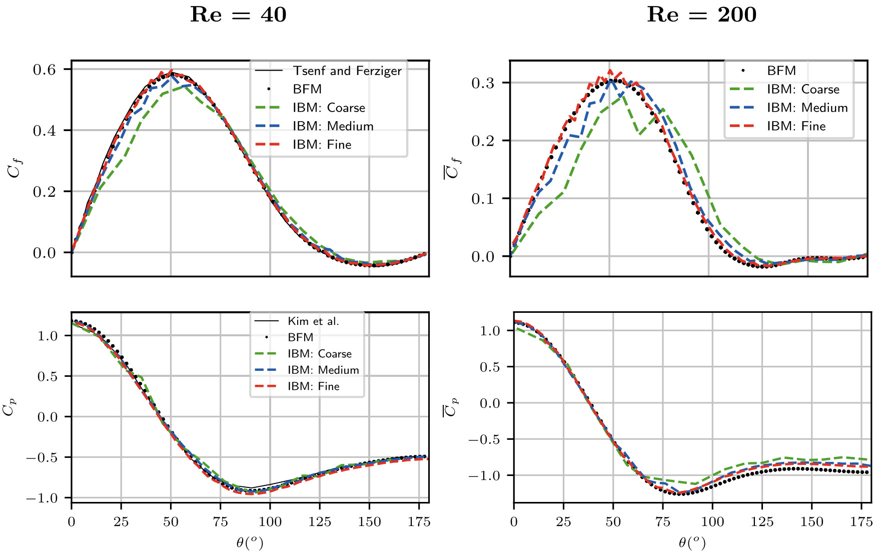


Fig. 6 Azimuthal variation of the skin-friction $C_f = \frac{\tau_w}{\frac{1}{2}\rho U^2}$ and the pressure $C_p = \frac{p_w}{\frac{1}{2}\rho U^2}$ coefficients for viscous flow around a circular cylinder. Results are also compared to references [7, 13]

ify the IBM implementations in a Cartesian mesh with three different resolutions (Table 1).

Flow is evaluated for Reynolds numbers $Re = UD/\nu$ equal to 40 and 200, respectively, involving steady laminar separation and laminar unsteady separation regimes. Although small spurious values appears for the skin-friction, it and the pressure coefficient of the IBM solution converge to the BFM solution and fit with the reference results (see Fig. 6 and Table 2). Consequently, the force coefficients, flow separation and wake length also show a good agreement.

To summarise, we have developed a novel method for automatic solid surface reconstruction from a 3-D points cloud applied to flow simulation by immersed boundary method capability in the fully parallel *open source* solver code_saturne. The solid reconstruction algorithm avoids the time-consuming manual (or assisted) meshing step of the body-fitted methods workflow, and it can deal with any convex cell geometry of fluid solvers based on finite-volume formulation. For the cylinder study case, manual/assisted meshing by an expert could require about 10 minutes, while the current algorithm can do it in fractions of a second, depending on the point cloud size. Besides, current developments open perspectives of application, such as scalar transport (e.g., temperature, sediments and pollutants) and real 3-D scan applications for complex indoor and exterior spaces, which will be included in subsequent publications.

Table 2 Drag C_D and lift C_L coefficients, separation angle θ_s and recirculating wake length L_w/D and Strouhal number of the unsteady wake *St*

$Re = 40$			
Study	C_D (C_{D_p}, C_{D_τ})	θ_s	L_w/D
[5]	1.49	126.4°	2.24
[10]	1.56	127.3°	2.14
BFM	1.560 (1.02, 0.54)	127.3°	2.22
IBM - Coarse	1.545 (0.996, 0.546)	131.2°	2.10
IBM - Medium	1.555 (1.003, 0.552)	128.3°	2.24
IBM - Fine	1.560 (1.010, 0.549)	128.2°	2.25
$Re = 200$			
Study	$\overline{C}_D$ $(\overline{C}_{D_p}, \overline{C}_{D_\tau})$	$C_{L_{max}}$ $(C_{L_p}, C_{L_\tau})_{max}$[a]	*St*
[9]	1.34 ± 0.044	0.69	0.197
[11]	1.35 ± 0.048	0.68	0.196
BFM	1.386 ± 0.051 (1.138, 0.248)	0.753 (0.689, 0.092)	0.191
IBM - Coarse	1.240 ± 0.022 (0.981, 0.259)	0.462 (0.438, 0.033)	0.191
IBM - Medium	1.324 ± 0.046 (1.059, 0.265)	0.623 (0.575, 0.057)	0.192
IBM - Fine	1.342 ± 0.046 (1.078, 0.264)	0.680 (0.622, 0.066)	0.192

[a] The pressure $C_{L_{p\ max}}$ and the viscous part $C_{L_{\tau\ max}}$ are not in phase

References

1. Amenta, N., Choi, S., Kolluri, R.K.: The power crust. In: Proceedings of the Sixth ACM Symposium on Solid Modeling and Applications, pp. 249–266 (2001). https://doi.org/10.1145/376957.376986
2. Bouchiba, H., Santoso, S., Deschaud, J.E., Rocha-Da-Silva, L., Goulette, F., Coupez, T.: Computational fluid dynamics on 3D point set surfaces. J. Comput. Phys. X **7**, 100,069 (2020). https://doi.org/10.1016/j.jcpx.2020.100069
3. Chorin, A.J.: Numerical solution of the Navier-Stokes equations. Math. Comput. **22**(104), 745–762 (1968). https://doi.org/10.1090/S0025-5718-1968-0242392-2
4. Colas, C., Ferrand, M., Hérard, J.M., Latché, J.C., Le Coupanec, E.: An implicit integral formulation to model inviscid fluid flows in ob- structed media. Comput. Fluids **188**, 136–163 (2019). https://hal.archives-ouvertes.fr/hal-01969129
5. Gautier, R., Biau, D., Lamballais, E.: A reference solution of the flow over a circular cylinder at Re = 40. Comput. Fluids **75**, 103–111 (2013)
6. Kazhdan, M., Bolitho, M., Hoppe, H.: Poisson surface reconstruction. In: Proceedings of the Fourth Eurographics Symposium on Geometry Processing, vol. 7 (2006). https://doi.org/10.5555/1281957.1281965
7. Kim, J., Kim, D., Choi, H.: An immersed-boundary finite-volume method for simulations of flow in complex geometries. J. Comput. Phys. **171**(1), 132–150 (2001)

8. Kovalčíková Duračíková, K., Bugáňová, A., Cimrák, I.: Modelling of arbitrary shaped channels and obstacles by distance function. In: International Work-Conference on Bioinformatics and Biomedical Engineering, pp. 28–41. Springer (2022). https://doi.org/10.1007/978-3-031-07704-3_3
9. Linnick, M.N., Fasel, H.F.: A high-order immersed interface method for simulating unsteady incompressible flows on irregular domains. J. Comput. Phys. **204**(1), 157–192 (2005)
10. Patil, D.V., Lakshmisha, K.: Finite volume TDV formulation of Lattice Boltzmann simulation on unstructured mesh. J. Comput. Phys. **228**(14), 5262–5279 (2009)
11. Taira, K., Colonius, T.: The immersed boundary method: A projection approach. J. Comput. Phys. **225**(2), 2118–2137 (2007)
12. Temam, R.: Sur l'approximation de la solution des équations de Navier-Stokes par la méthode des pas fractionnaires (I). Arch. Ration. Mech. Anal. **32**(5), 135–153 (1969)
13. Tseng, Y.H., Ferziger, J.H.: A ghost-cell immersed boundary method for flow in complex geometry. J. Comput. Phys. **192**(2), 593–623 (2003)

Stokes–Brinkman–Darcy Models for Coupled Free-Flow and Porous-Medium Systems

Linheng Ruan and Iryna Rybak

Abstract Coupled systems involving free flow and porous medium have gained significant attention in recent years due to their prevalence in environment and industry. Most of the coupling approaches are suitable only for flows parallel to the fluid–porous interface, and a generalization of the coupling concept is required. In this work, we consider a thin transition region between the free-flow and porous-medium domains, which stores and transports mass, momentum, and energy. The flow system of interest is incompressible and single-phase. The model comprises the Stokes equations in the free-flow domain, the Brinkman equations in the transition region, and Darcy's law in the porous medium. These models are coupled through suitable interface conditions. Numerical simulation results for the coupled full-dimensional Stokes–Brinkman–Darcy model are provided. A dimensionally reduced formulation for the coupled model is proposed in the case of a thin transition region. This model consists of the averaged Brinkman equations of co-dimension one, which are coupled to the full-dimensional Stokes and Darcy's equations in the free flow and porous medium, respectively.

Keywords Stokes equations · Brinkman equations · Darcy's law · Porous media · Finite volume method

MSC (2010) 76D07 · 76S05 · 76M12

L. Ruan (✉) · I. Rybak
Institute of Applied Analysis and Numerical Simulation, University of Stuttgart, Pfaffenwaldring 57, 70569 Stuttgart, Germany
e-mail: linheng.ruan@ians.uni-stuttgart.de

I. Rybak
e-mail: iryna.rybak@ians.uni-stuttgart.de

E. Franck et al. (eds.), *Finite Volumes for Complex Applications X—Volume 1, Elliptic and Parabolic Problems*, Springer Proceedings in Mathematics & Statistics 432,
https://doi.org/10.1007/978-3-031-40864-9_31

1 Introduction

Coupled free-flow and porous-medium systems appear in many environmental settings and industrial applications, e.g., evaporation from soil influenced by wind, industrial filtration, and drying processes. In a macroscale view, these flow systems consist of two distinct sub-regions, and many coupling strategies have been developed in the past few years. Generally, the Stokes equations describe laminar flows in the free-flow domain, and Darcy's law describes the fluid flow in the porous-medium domain. Flow behavior between the free-flow and porous-medium regions can be described using a sharp interface or a transition region concept. For the sharp interface approach, there exist coupled Stokes–Darcy models with different sets of interface conditions, e.g., [2, 3, 5, 8]. Most of the coupling conditions are based on the Beavers–Joseph approach in the tangential direction [2], and therefore they are only suitable for the flows parallel to the interface. Recently, generalized coupling conditions for arbitrary flow directions at the sharp interface have been developed in [4] using homogenization and boundary layer theory.

An alternative approach is to consider a thin transition region between two flow domains, which resolves storage and transfer of mass, momentum, and energy in the tangential direction. The transition region approach can be regarded as a generalization of the sharp interface concept. The Brinkman equations, which are an extension of Darcy's law, are applied to describe fluid flow in the transition region and take the dissipation of the kinetic energy by viscous shear into account. In this paper, we couple three different models (Stokes equations in the free-flow domain, Brinkman equations in the transition region, and Darcy's law in the porous-medium domain) using suitable interface conditions. At the interface between the free flow and transition region, we consider the coupling conditions developed in [1, 9], where a stress jump in the normal and tangential directions is involved. At the interface between the transition region and porous medium, we use the classical coupling strategy based on the Beavers–Joseph–Saffman condition [2, 3, 12].

The thickness of the transition region is significantly smaller than the size of the whole flow domain. Therefore, the transition zone can be modeled as a lower-dimensional inclusion in the coupled system. Several dimensionally reduced models exist in the literature for flows in fractured porous media, e.g., [6, 7, 11]. We use a similar approach in this work and propose a reduced Stokes–Brinkman–Darcy model, which is derived by averaging the Brinkman equations across the transition region and coupling these equations of co-dimension one to the full-dimensional Stokes and Darcy's equations.

The paper is arranged as follows. In Sect. 2, we propose the Stokes–Brinkman–Darcy model for coupled free-flow and porous-medium systems In Sect. 3, we derive the reduced formulation with suitable closure relations. The numerical simulation results are presented in Sect. 4. Conclusions and future work are provided in Sect. 5.

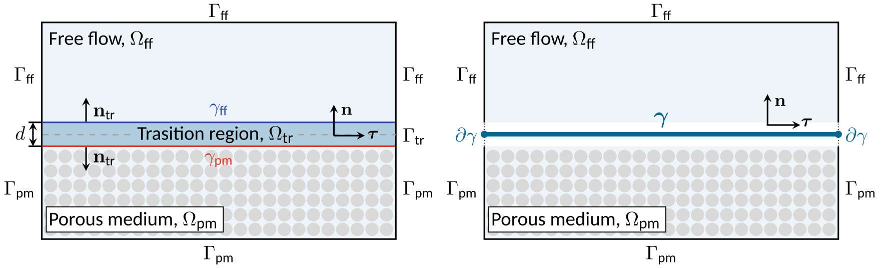


Fig. 1 Schematic representation of the coupled free-flow and porous-medium systems with a full-dimensional transition region (left) and a lower-dimensional interface (right)

2 Coupled Stokes–Brinkman–Darcy Models

In this section, we propose the coupled Stokes–Brinkman–Darcy model for the free-flow and porous-medium systems. The flow domain $\overline{\Omega} = \overline{\Omega}_{\rm ff} \cup \overline{\Omega}_{\rm tr} \cup \overline{\Omega}_{\rm pm} \subset \mathbb{R}^2$ consists of the free-flow region $\Omega_{\rm ff}$, transition region $\Omega_{\rm tr}$, and porous-medium region $\Omega_{\rm pm}$ (Fig. 1, left). We introduce a local coordinate system with the corresponding unit normal vector $\mathbf{n}$ and unit tangential vector $\boldsymbol{\tau}$ (Fig. 1). The interface between the free-flow domain and transition region $\gamma_{\rm ff} = \overline{\Omega}_{\rm ff} \cap \overline{\Omega}_{\rm tr} \setminus \partial\Omega$ and the interface between the transition region and porous-medium domain $\gamma_{\rm pm} = \overline{\Omega}_{\rm tr} \cap \overline{\Omega}_{\rm pm} \setminus \partial\Omega$ are assumed to be smooth enough. The thickness of the transition region $d > 0$ is much smaller than the size of the flow domain Ω, which motivates us to model the transition region as a lower-dimensional interface γ (Fig. 1, right).

Both the transition region and porous-medium domain are assumed to be homogeneous. We consider single-phase and steady-state fluid flows at low Reynolds numbers in domain Ω. The fluid is supposed to be isothermal and incompressible with constant viscosity. The same fluid occupies the free-flow domain and fully saturates the porous medium. We consider the Stokes equations in the free-flow region, the Brinkman equations in the transition zone, and Darcy's law in the porous medium. The flow models in these regions are coupled using suitable coupling conditions at the interfaces $\gamma_{\rm ff}$ and $\gamma_{\rm pm}$, respectively.

2.1 Stokes Equations

Fluid flow in the free-flow region is described by the Stokes equations. Since the fluid is incompressible, we have mass conservation

$$\nabla \cdot \mathbf{v}_{\rm ff} = 0 \quad \text{in } \Omega_{\rm ff}, \tag{1}$$

where $\mathbf{v}_{\rm ff}$ is the free-flow velocity. In the case of laminar flows, we disregard the convective acceleration and use Newton's law to get the momentum conservation equations

$$-\nabla \cdot \mathbf{T}\left(\mathbf{v}_{\mathrm{ff}}, p_{\mathrm{ff}}\right) = \mathbf{f}_{\mathrm{ff}} \quad \text{in } \Omega_{\mathrm{ff}}, \tag{2}$$

where p_{ff} is the free-flow pressure, $\mathbf{f}_{\mathrm{ff}}$ is the momentum source term, $\mathbf{T}\left(\mathbf{v}, p\right) = \mu \nabla \mathbf{v} - p\mathbf{I}$ is the stress tensor, μ is the dynamic viscosity, and $\mathbf{I}$ is the identity tensor. On the external boundary of the free-flow domain $\Gamma_{\mathrm{ff}} = \partial\Omega_{\mathrm{ff}} \setminus \gamma_{\mathrm{ff}}$, we consider the Dirichlet boundary conditions

$$\mathbf{v}_{\mathrm{ff}} = \overline{\mathbf{v}}_{\mathrm{ff}} \quad \text{on } \Gamma_{\mathrm{ff}}, \tag{3}$$

where $\overline{\mathbf{v}}_{\mathrm{ff}}$ is a given function.

2.2 Brinkman Equations

To describe the flow in the transition region, the Brinkman equations are used

$$\nabla \cdot \mathbf{v}_{\mathrm{tr}} = 0 \quad \text{in } \Omega_{\mathrm{tr}}, \tag{4}$$

$$\mu \mathbf{K}_{\mathrm{tr}}^{-1} \mathbf{v}_{\mathrm{tr}} - \nabla \cdot \mathbf{T}_{\mathrm{eff}}(\mathbf{v}_{\mathrm{tr}}, p_{\mathrm{tr}}) = \mathbf{f}_{\mathrm{tr}} \quad \text{in } \Omega_{\mathrm{tr}}, \tag{5}$$

where $\mathbf{v}_{\mathrm{tr}}$ and p_{tr} represent the velocity and pressure in the transition region, respectively, $\mathbf{K}_{\mathrm{tr}}$ is the permeability of the transition region, $\mathbf{f}_{\mathrm{tr}}$ is the momentum source term, and $\mathbf{T}_{\mathrm{eff}}\left(\mathbf{v}, p\right) = \mu_{\mathrm{eff}} \nabla \mathbf{v} - p\mathbf{I}$ represents the stress tensor in the transition region with the effective viscosity μ_{eff}. The permeability tensor $\mathbf{K}_{\mathrm{tr}}$ is symmetric, positive definite, and bounded.

On the external boundary of the transition region $\Gamma_{\mathrm{tr}} = \partial\Omega_{\mathrm{tr}} \setminus \{\gamma_{\mathrm{ff}} \cup \gamma_{\mathrm{pm}}\}$, the following Dirichlet boundary conditions

$$\mathbf{v}_{\mathrm{tr}} = \overline{\mathbf{v}}_{\mathrm{tr}} \quad \text{on } \Gamma_{\mathrm{tr}} \tag{6}$$

with a given function $\overline{\mathbf{v}}_{\mathrm{tr}}$ are imposed.

2.3 Darcy's Law

Darcy's equations are used to describe the slow flow in the porous-medium region

$$\nabla \cdot \mathbf{v}_{\mathrm{pm}} = q, \quad \mathbf{v}_{\mathrm{pm}} = -\frac{\mathbf{K}_{\mathrm{pm}}}{\mu} \nabla p_{\mathrm{pm}} \quad \text{in } \Omega_{\mathrm{pm}}, \tag{7}$$

where $\mathbf{v}_{\mathrm{pm}}$ is the fluid velocity through the porous medium, p_{pm} is the fluid pressure, $\mathbf{K}_{\mathrm{pm}}$ is the intrinsic permeability, and q is the source term. The permeability tensor $\mathbf{K}_{\mathrm{pm}}$ is symmetric, positive definite, and bounded.

On the external boundary $\Gamma_{\mathrm{pm}} = \partial\Omega \setminus \gamma_{\mathrm{pm}}$, we consider the following Dirichlet boundary condition

$$p_{\mathrm{pm}} = \overline{p}_{\mathrm{pm}} \quad \text{on } \Gamma_{\mathrm{pm}}, \tag{8}$$

where $\overline{p}_{\mathrm{pm}}$ is a given function.

2.4 Interface Conditions

In coupled Stokes–Brinkman–Darcy models, suitable interface conditions have to be chosen on the interfaces γ_{ff} and γ_{pm}.

On the interface between the free flow and the transition region, we consider continuity of velocity

$$\mathbf{v}_{\mathrm{tr}} = \mathbf{v}_{\mathrm{ff}} \quad \text{on } \gamma_{\mathrm{ff}}. \tag{9}$$

According to [1], a stress jump exists between the free-flow domain and transition region

$$\mathbf{T}(\mathbf{v}_{\mathrm{ff}}, p_{\mathrm{ff}}) \cdot \mathbf{n}_{\mathrm{tr}} - \mathbf{T}_{\mathrm{eff}}(\mathbf{v}_{\mathrm{tr}}, p_{\mathrm{tr}}) \cdot \mathbf{n}_{\mathrm{tr}} = \frac{\mu}{\sqrt{K_{\mathrm{tr}}}} \boldsymbol{\beta} \mathbf{v}_{\mathrm{ff}} \quad \text{on } \gamma_{\mathrm{ff}}, \tag{10}$$

where K_{tr} is the permeability component, $\boldsymbol{\beta}$ denotes the symmetric positive semi-definite friction tensor, and $\mathbf{n}_{\mathrm{tr}} = \mathbf{n}$ is the unit outward normal vector from Ω_{tr} on interfaces (Fig. 1, left).

Across the interface between the transition region and the porous medium, we consider the mass conservation

$$\mathbf{v}_{\mathrm{tr}} \cdot \mathbf{n}_{\mathrm{tr}} = \mathbf{v}_{\mathrm{pm}} \cdot \mathbf{n}_{\mathrm{tr}} \quad \text{on } \gamma_{\mathrm{pm}}, \tag{11}$$

and the balance of normal forces

$$-\mathbf{n}_{\mathrm{tr}} \cdot \mathbf{T}_{\mathrm{eff}}(\mathbf{v}_{\mathrm{tr}}, p_{\mathrm{tr}}) \cdot \mathbf{n}_{\mathrm{tr}} = p_{\mathrm{pm}} \quad \text{on } \gamma_{\mathrm{pm}}, \tag{12}$$

where $\mathbf{n}_{\mathrm{tr}} = -\mathbf{n}$. The tangential velocity on the interface γ_{pm} satisfies the well-known Beavers–Joseph–Saffman condition

$$\mathbf{v}_{\mathrm{tr}} \cdot \boldsymbol{\tau} = \frac{\sqrt{K_{\mathrm{pm}}}}{\alpha} \frac{\partial \mathbf{v}_{\mathrm{tr}}}{\partial \mathbf{n}} \cdot \boldsymbol{\tau} \quad \text{on } \gamma_{\mathrm{pm}}. \tag{13}$$

Here, $\alpha > 0$ is the slip coefficient and $K_{\mathrm{pm}} = \boldsymbol{\tau} \cdot \mathbf{K}_{\mathrm{pm}} \boldsymbol{\tau}$.

3 Dimensionally Reduced Model

We consider the thickness in the transition region d to be much smaller than the length of the flow domain Ω. This motivates us to model the transition region as a complex interface γ of co-dimension one (Fig. 1, right), which allows storage and transport of mass and momentum in and along the interface, respectively. The transition region is described as $\Omega_{\text{tr}} = \left\{\mathbf{x} \in \mathbb{R}^2 \middle| \mathbf{x} = \mathbf{s} + \xi \frac{d}{2}\mathbf{n}, \mathbf{s} \in \gamma, \xi \in [-1, 1]\right\}$. To derive a reduced model, we project the Brinkman equations (4)–(5) on the local orthogonal reference system in normal and tangential directions on γ and average them in the vertical direction.

Integrating the incompressibility condition (4) in the vertical direction and considering the mass conservation equation in (9) and (11) across γ_{ff} and γ_{pm}, respectively, we obtain

$$\mathbf{v}_{\text{ff}} \cdot \mathbf{n}\big|_{\gamma_{\text{ff}}} - \mathbf{v}_{\text{pm}} \cdot \mathbf{n}|_{\gamma_{\text{pm}}} + d\frac{\partial V_{\boldsymbol{\tau}}}{\partial \boldsymbol{\tau}} = 0 \quad \text{on } \gamma, \tag{14}$$

where the averaged tangential velocity is defined as $V_{\boldsymbol{\tau}} := \left(\frac{1}{d}\int_{-\frac{d}{2}}^{\frac{d}{2}} \mathbf{v}_{\text{tr}} \cdot \boldsymbol{\tau}\, dn\right)$.

Integrating the momentum conservation equations (5) in the vertical direction and taking interface conditions (10) on γ_{ff} and (12)–(13) on γ_{pm} into account, we have

$$\left(p_{\text{ff}} - \mu\frac{\partial \mathbf{v}_{\text{ff}}}{\partial \mathbf{n}} \cdot \mathbf{n} + \frac{\mu}{\sqrt{K_{\text{tr}}}}\mathbf{n} \cdot \boldsymbol{\beta}\mathbf{v}_{\text{ff}}\right)\Bigg|_{\gamma_{\text{ff}}} - p_{\text{pm}}|_{\gamma_{\text{pm}}} \tag{15}$$

$$= d\left(F_{\mathbf{n}} - \mu\mathbf{M}_{\mathbf{nn}}V_{\mathbf{n}} - \mu\mathbf{M}_{\mathbf{n}\boldsymbol{\tau}}V_{\boldsymbol{\tau}} + \mu_{\text{eff}}\frac{\partial^2 V_{\mathbf{n}}}{\partial \boldsymbol{\tau}^2}\right) \quad \text{on } \gamma,$$

$$\left(-\mu\frac{\partial \mathbf{v}_{\text{ff}}}{\partial \mathbf{n}} \cdot \boldsymbol{\tau} + \frac{\mu}{\sqrt{K_{\text{tr}}}}\boldsymbol{\tau} \cdot \boldsymbol{\beta}\mathbf{v}_{\text{ff}}\right)\Bigg|_{\gamma_{\text{ff}}} + \mu_{\text{eff}}\frac{\alpha}{\sqrt{K_{\text{pm}}}}\mathbf{v}_{\text{tr}} \cdot \boldsymbol{\tau}\Bigg|_{\gamma_{\text{pm}}} \tag{16}$$

$$= d\left(F_{\boldsymbol{\tau}} - \mu\mathbf{M}_{\boldsymbol{\tau}\mathbf{n}}V_{\mathbf{n}} - \mu\mathbf{M}_{\boldsymbol{\tau}\boldsymbol{\tau}}V_{\boldsymbol{\tau}} + \mu_{\text{eff}}\frac{\partial^2 V_{\boldsymbol{\tau}}}{\partial \boldsymbol{\tau}^2} - \frac{\partial P}{\partial \boldsymbol{\tau}}\right) \quad \text{on } \gamma.$$

Here, we define $\mathbf{M}_{\mathbf{ab}} := \mathbf{a}^T\mathbf{K}_{\text{tr}}^{-1}\mathbf{b}$ for vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^2$, the averaged normal velocity $V_{\mathbf{n}} := \left(\frac{1}{d}\int_{-\frac{d}{2}}^{\frac{d}{2}} \mathbf{v}_{\text{tr}} \cdot \mathbf{n}\, dn\right)$, the averaged pressure $P := \left(\frac{1}{d}\int_{-\frac{d}{2}}^{\frac{d}{2}} p_{\text{tr}}\, dn\right)$, and the averaged source terms $F_{\mathbf{n}} := \left(\frac{1}{d}\int_{-\frac{d}{2}}^{\frac{d}{2}} \mathbf{f}_{\text{tr}} \cdot \mathbf{n}\, dn\right)$, $F_{\boldsymbol{\tau}} := \left(\frac{1}{d}\int_{-\frac{d}{2}}^{\frac{d}{2}} \mathbf{f}_{\text{tr}} \cdot \boldsymbol{\tau}\, dn\right)$.

To close the model, we need to express $\mathbf{v}_{\text{tr}} \cdot \boldsymbol{\tau}|_{\gamma_{\text{pm}}}$ in (16) and $\frac{\partial \mathbf{v}_{\text{tr}}}{\partial \mathbf{n}} \cdot \boldsymbol{\tau}\big|_{\gamma_{\text{ff}}}$,

$$\frac{\partial \mathbf{v}_{\text{tr}}}{\partial \mathbf{n}} \cdot \mathbf{n}\big|_{\gamma_{\text{ff}}/\gamma_{\text{pm}}}$$

in (10) and (12) in terms of $(V_{\mathbf{n}}, V_{\boldsymbol{\tau}}, P)$. Assuming a quadratic profile in the tangential direction and applying the tangential velocity continuity in Eq. (9) and the Beavers–Joseph–Saffman condition (13), we obtain the closure condition

$$\mathbf{v}_{\text{tr}} \cdot \boldsymbol{\tau}\Big|_{\gamma_{\text{pm}}} = \frac{\sqrt{K_{\text{pm}}}\left(6V_{\boldsymbol{\tau}} - 2\mathbf{v}_{\text{ff}} \cdot \boldsymbol{\tau}|_{\gamma_{\text{ff}}}\right)}{\left(\alpha d + 4\sqrt{K_{\text{pm}}}\right)}, \tag{17}$$

$$\frac{\partial \mathbf{v}_{\text{tr}}}{\partial \mathbf{n}} \cdot \boldsymbol{\tau}\Bigg|_{\gamma_{\text{ff}}} = \frac{-\left(6\alpha d + 12\sqrt{K_{\text{pm}}}\right) V_{\boldsymbol{\tau}} + \left(4\alpha d + 12\sqrt{K_{\text{pm}}}\right) \mathbf{v}_{\text{ff}} \cdot \boldsymbol{\tau}|_{\gamma_{\text{ff}}}}{d\left(\alpha d + 4\sqrt{K_{\text{pm}}}\right)}. \tag{18}$$

Assuming linear normal velocity in the transition region and applying continuity of normal velocities in (9) and (11), we obtain the closure condition

$$\frac{\partial \mathbf{v}_{\text{tr}}}{\partial \mathbf{n}} \cdot \mathbf{n}\Bigg|_{\gamma_{\text{ff}}} = -\frac{\left(V_{\mathbf{n}} - \mathbf{v}_{\text{ff}} \cdot \mathbf{n}|_{\gamma_{\text{ff}}}\right)}{0.5d}, \quad \frac{\partial \mathbf{v}_{\text{tr}}}{\partial \mathbf{n}} \cdot \mathbf{n}\Bigg|_{\gamma_{\text{pm}}} = \frac{\left(V_{\mathbf{n}} - \mathbf{v}_{\text{pm}} \cdot \mathbf{n}|_{\gamma_{\text{pm}}}\right)}{0.5d}. \tag{19}$$

Remark 1 If there is no transition region ($d = 0$), Eq. (14) becomes the mass conservation and Eq. (15) reverts to the normal stress jump between the free-flow domain and porous-medium region.

4 Numerical Simulations

In this section, we present numerical results for the full-dimensional model (1)–(13) and the proposed reduced model (1)–(3), (7)–(8), (10), (12), (14)–(19). We discretize both models using the second-order finite volume method on staggered grids (MAC scheme) as in [10]. In all regions we consider uniform rectangular grids conforming at the interfaces. The computation domain is $\Omega = [0, 1] \times [0, 2]$ with interfaces $\gamma_{\text{ff}} = (0, 1) \times \{1.2\}$ and $\gamma_{\text{pm}} = (0, 1) \times \{0.8\}$. We choose the parameters $\mu = \mu_{\text{eff}} = 1$, $\boldsymbol{\beta} = \mathbf{0}$, $\mathbf{K}_{\text{tr}} = \mathbf{K}_{\text{pm}} = \mathbf{I}$, and $\alpha = 1$. The local basis is set as $\boldsymbol{\tau} = \mathbf{e}_1$ and $\mathbf{n} = \mathbf{e}_2$.

In this section, we present numerical simulation results for both flow models. The analytical solution that satisfies the incompressibility conditions (1), (4) and the coupling conditions (9)–(13) on both interfaces is chosen as follows

$$\begin{aligned}
u_{\text{ff}} &= \cos(x_1)\mathrm{e}^{x_2-0.8}, \quad v_{\text{ff}} = \sin(x_1)\mathrm{e}^{x_2-0.8}, \quad p_{\text{ff}} = \sin(x_1 + x_2 - 0.8),\\
u_{\text{tr}} &= \cos(x_1)\mathrm{e}^{x_2-0.8}, \quad v_{\text{tr}} = \sin(x_1)\mathrm{e}^{x_2-0.8}, \quad p_{\text{tr}} = \sin(x_1 + x_2 - 0.8),\\
& \qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad p_{\text{pm}} = (0.8 - x_2)\sin(x_1),
\end{aligned} \tag{20}$$

where $\mathbf{v}_{\text{ff}} = (u_{\text{ff}}, v_{\text{ff}})$ and $\mathbf{v}_{\text{tr}} = (u_{\text{tr}}, v_{\text{tr}})$. The source terms and the boundary conditions are obtained by substituting the chosen parameters and the exact solution (20) into Eqs. (2), (3), and (5)–(8).

Starting from $h = \frac{1}{10}$, the grid size is decreased by a factor of two at each refinement level, where seven refinement levels are considered. For the convergence analysis, we compute the L_2-errors for all primary variables

$$\varepsilon_f = \|f - f_h\|_2, \quad f \in \{u_{\text{ff}}, v_{\text{ff}}, p_{\text{ff}}, u_{\text{tr}}, v_{\text{tr}}, p_{\text{tr}}, p_{\text{pm}}\}, \tag{21}$$

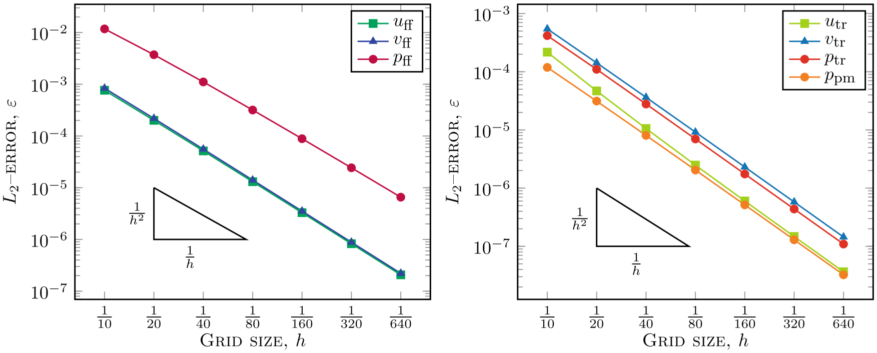


Fig. 2 Convergence analysis for all primary variables (full-dimensional model)

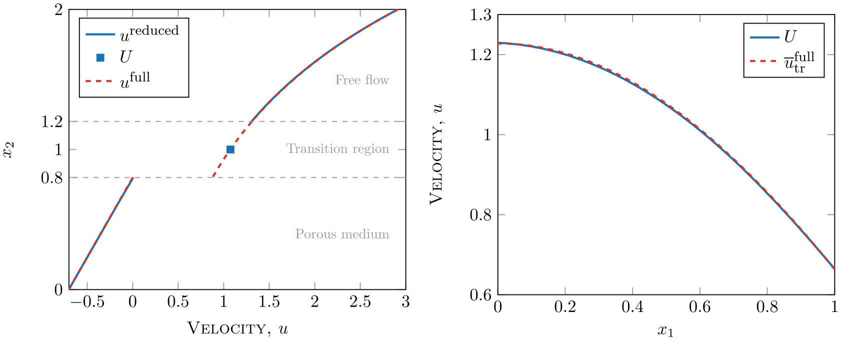


Fig. 3 Comparison between the full- and reduced-dimensional models: tangential velocity profile at $x_1 = 0.5$ (left) and $x_2 = 1$ (right)

where f_h is the numerical solution. The simulation results are visualized in Fig. 2 and demonstrate the second-order convergence of the discretization scheme.

We validate the reduced model (1)–(3), (7)–(8), (10), (12), (14)–(19) against the full-dimensional model (1)–(13). The averaged source terms $F_{\mathbf{n}}$, $F_{\boldsymbol{\tau}}$ and the boundary conditions on $\partial\gamma$ for the reduced model are obtained by averaging the corresponding terms across the transition region. Numerical simulation results for both models are obtained with grid size $h = \frac{1}{320}$ and presented in Fig. 3. We compare velocity profiles in the middle of the domain at $x_1 = 0.5$ (Fig. 3, left) and along γ (Fig. 3, right). In the latter case, we average the velocity of the full-dimensional model across transition region $\overline{u}_{\mathrm{tr}}^{\mathrm{full}} = \frac{1}{d}\int_{-\frac{d}{2}}^{\frac{d}{2}} u_{\mathrm{tr}}^{\mathrm{full}} dx_2$.

5 Conclusions

In this paper, we proposed the full- and reduced-dimensional Stokes–Brinkman–Darcy models for coupled free-flow and porous-medium systems with transition region. This concept describes storage and transfer of mass and momentum in the tangential direction. The coupled model consists of the Stokes equations in the free-flow domain, the Brinkman equations in the transition region, and Darcy's law in the porous-medium domain. In the reduced formulation, the Brinkman equations are averaged across the transition region. Suitable coupled conditions are chosen at the respective interfaces. Both models are discretized with the MAC scheme and the numerical simulation results are provided. Well-posedness of the full-dimensional and reduced-dimensional Stokes–Brinkmann–Darcy models will be presented at the conference.

Acknowledgements The work is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Project Number 490872182.

References

1. Angot, P., Goyeau, B., Ochoa-Tapia, J.A.: Nonlinear asymptotic model for the inertial flow at a fluid-porous interface. Adv. Water Res. **149**, 103798 (2021). https://doi.org/10.1016/j.advwatres.2020.103798
2. Beavers, G., Joseph, D.: Boundary conditions at a naturally permeable wall. J. Fluid Mech. **30**, 197–207 (1967). https://doi.org/10.1017/S0022112067001375
3. Discacciati, M., Quarteroni, A.: Navier-Stokes/Darcy coupling: modeling, analysis, and numerical approximation. Rev. Mat. Complut. **22**, 315–426 (2009). https://doi.org/10.5209/rev_REMA.2009.v22.n2.16263
4. Eggenweiler, E., Rybak, I.: Effective coupling conditions for arbitrary flows in Stokes-Darcy systems. Multiscale Model. Simul. **19**, 731–757 (2021). https://doi.org/10.1137/20M1346638
5. 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). https://doi.org/10.1017/jfm.2016.838
6. Lesinigo, M., D'Angelo, C., Quarteroni, A.: A multiscale Darcy-Brinkman model for fluid flow in fractured porous media. Numer. Math. **117**, 717–752 (2011). https://doi.org/10.1007/s00211-010-0343-2
7. 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). https://doi.org/10.1137/S1064827503429363
8. Mikelić, A., Jäger, W.: On the interface boundary condition of Beavers, Joseph, and Saffman. SIAM J. Appl. Math. **60**, 1111–1127 (2000). https://doi.org/10.1137/S003613999833678X
9. Ochoa-Tapia, J. A., Whitaker, S.: Momentum transfer at the boundary between a porous medium and a homogeneous fluid – I. Theoretical development. Int. J. Heat Mass Transf. **38**, 2635–2646 (1995). https://doi.org/10.1016/0017-9310(94)00346-W
10. Rybak, I., Magiera, J., Helmig, R., et al.: Multirate time integration for coupled saturated/unsaturated porous medium and free flow systems. Comput. Geosci. **19**, 299–309 (2015). https://doi.org/10.1007/s10596-015-9469-8
11. Rybak, I., Metzger, S.: A dimensionally reduced Stokes-Darcy model for fluid flow in fractured porous media. Appl. Math. Comput. **384**, 125260 (2020). https://doi.org/10.1016/j.amc.2020.125260
12. Saffman, P.G.: On the boundary condition at the surface of a porous medium. Stud. Appl. Math. **50**, 93–101 (1971). https://doi.org/10.1002/sapm197150293

Robust and Efficient Preconditioners for Stokes–Darcy Problems

Paula Strohbeck, Cedric Riethmüller, Dominik Göddeke, and Iryna Rybak

Abstract Coupled systems of porous media and free flow can be modelled by the Stokes equations in the free-flow domain, Darcy's law in the porous medium, and an appropriate set of coupling conditions on the fluid–porous interface. Discretisation of the coupled Stokes–Darcy problem leads to a large, sparse, ill-conditioned, and nonsymmetric linear system. We discretise the system using the MAC scheme, i.e., the finite volume method on staggered grids. To accelerate convergence of the GMRES method, efficient preconditioners are needed. We propose a block diagonal, a block triangular and a constraint preconditioner for the Stokes–Darcy problem with the classical set of coupling conditions based on the Beavers–Joseph condition and the generalised coupling conditions which were developed for arbitrary flow to the interface. We show the robustness and efficiency of the proposed preconditioners in numerical experiments.

Keywords Stokes equations · Darcy's law · Fluid–porous interface · Finite volume method · Preconditioner

P. Strohbeck (✉) · C. Riethmüller · D. Göddeke · I. Rybak
Institute of Applied Analysis and Numerical Simulation, University of Stuttgart, Pfaffenwaldring 57, 70569 Stuttgart, Germany
e-mail: paula.strohbeck@ians.uni-stuttgart.de

C. Riethmüller
e-mail: cedric.riethmueller@ians.uni-stuttgart.de

D. Göddeke
e-mail: dominik.goeddeke@ians.uni-stuttgart.de

I. Rybak
e-mail: iryna.rybak@ians.uni-stuttgart.de

C. Riethmüller · D. Göddeke
Stuttgart Centre for Simulation Science, University of Stuttgart, Allmandring 5b, 70569 Stuttgart, Germany

E. Franck et al. (eds.), *Finite Volumes for Complex Applications X—Volume 1, Elliptic and Parabolic Problems*, Springer Proceedings in Mathematics & Statistics 432,
https://doi.org/10.1007/978-3-031-40864-9_32

1 Introduction

Coupled free-flow and porous-medium flow problems appear routinely in science and engineering, e.g., interaction between surface and groundwater, water-gas management in fuel cells, industrial filtration, etc. The most widely studied problem in the literature is the Stokes–Darcy problem for coupled single-fluid-phase flows with different sets of interface conditions on the fluid–porous interface, see e.g., [1, 10, 11, 13]. Solving such coupled flow problems in a monolithic way is challenging because the system of linear equations arising from discretisation is ill-conditioned. However, for validation purposes, the monolithic approach is the method of choice. Thus, proper preconditioning techniques are needed.

The Stokes–Darcy problem with the classical set of interface conditions (conservation of mass, balance of normal forces, the Beavers–Joseph–Saffman condition on the tangential velocity) has been well-studied in the last two decades. Several preconditioning strategies have been developed recently for this problem, e.g., [4, 6–8, 12, 15]. However, the Beavers–Joseph–Saffman condition [2] is only applicable for flows parallel to the fluid–porous interface. That restricts the amount of applications which can be accurately modelled. Recently, generalised coupling conditions suitable for arbitrary flow directions have been proposed in [11]. The purpose of this work is to develop robust and efficient preconditioners for the Stokes–Darcy problem with these generalised interface conditions and with the Beavers–Joseph condition without Saffman simplification.

Discretisation of the Stokes–Darcy problem leads to a saddle-point type matrix. Therefore, we adjust preconditioning techniques developed for saddle-point systems [3, 5]. We extend the results available in literature [4, 7, 8, 12] and propose three different preconditioners for the Stokes–Darcy problem with the Beavers–Joseph and the generalised interface conditions: block diagonal, block triangular and constraint preconditioners. We study the effectiveness and robustness of these preconditioners and provide numerical simulation results.

The paper is organised as follows. In Sect. 2, we describe the coupled Stokes–Darcy problem. In Sect. 3, we briefly describe the discretisation scheme and propose three different preconditioners for the discrete problem. The benchmark problem and the numerical simulations results are presented in Sect. 4. Conclusions and future work follow in Sect. 5.

2 Problem Formulation

The coupled domain $\Omega = \Omega_{\mathrm{pm}} \cup \Omega_{\mathrm{ff}}$ consists of the free-flow region Ω_{ff} and the porous-medium domain Ω_{pm} coupled at the sharp fluid–porous interface Γ (Fig. 1). In this paper, we restrict ourselves to a two-dimensional setting. We consider a steady-state, single-phase flow of an incompressible and isothermal fluid at low Reynolds

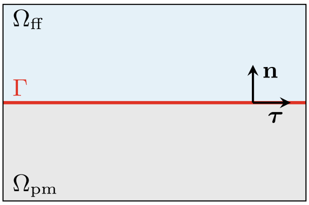


Fig. 1 Flow system description

numbers ($Re \ll 1$). The solid phase is supposed to be nondeformable and rigid leading to a constant porosity.

The fluid flow in the free-flow domain Ω_{ff} is described by the Stokes equations

$$\nabla \cdot \mathbf{v}_{\mathrm{ff}} = 0, \qquad -\nabla \cdot \mathbf{T}(\mathbf{v}_{\mathrm{ff}}, p_{\mathrm{ff}}) = \mathbf{0} \quad \text{in } \Omega_{\mathrm{ff}}, \tag{1}$$

where $\mathbf{v}_{\mathrm{ff}}$ is the fluid velocity, p_{ff} is the fluid pressure, $\mathbf{T}(\mathbf{v}_{\mathrm{ff}}, p_{\mathrm{ff}}) = \mu \nabla \mathbf{v}_{\mathrm{ff}} - p_{\mathrm{ff}} \mathbf{I}$ is the nonsymmetric stress tensor, μ is the dynamic viscosity, and $\mathbf{I}$ is the identity tensor.

Fluid flow in the porous medium is described by Darcy's law

$$\nabla \cdot \mathbf{v}_{\mathrm{pm}} = 0, \qquad \mathbf{v}_{\mathrm{pm}} = -\frac{\mathbf{K}}{\mu} \nabla p_{\mathrm{pm}} \quad \text{in } \Omega_{\mathrm{pm}}, \tag{2}$$

where $\mathbf{K}$ is the intrinsic permeability tensor, which is symmetric, positive definite, and bounded.

Equations (1) and (2) are of different types. To couple them on the interface Γ, various sets of interface conditions have been proposed in the literature. In this paper, we consider the classical set of coupling conditions which are valid for parallel flows to the interface as well as generalised conditions which have been recently developed in [11] and are applicable to arbitrary flows.

The following coupling conditions—the conservation of mass across the interface (3), the balance of normal forces (4) and the Beavers–Joseph condition (5) on the tangential velocity [2]—are typically used in the literature

$$\mathbf{v}_{\mathrm{ff}} \cdot \mathbf{n} = \mathbf{v}_{\mathrm{pm}} \cdot \mathbf{n} \qquad \text{on } \Gamma, \tag{3}$$

$$-\mathbf{n} \cdot \mathbf{T}(\mathbf{v}_{\mathrm{ff}}, p_{\mathrm{ff}})\, \mathbf{n} = p_{\mathrm{pm}} \qquad \text{on } \Gamma, \tag{4}$$

$$\left(\mathbf{v}_{\mathrm{ff}} - \mathbf{v}_{\mathrm{pm}}\right) \cdot \boldsymbol{\tau} - \frac{\sqrt{K}}{\alpha_{\mathrm{BJ}}} \left(\nabla \mathbf{v}_{\mathrm{ff}}\, \mathbf{n}\right) \cdot \boldsymbol{\tau} = 0 \qquad \text{on } \Gamma. \tag{5}$$

Here, $\mathbf{n} = -\mathbf{n}_{\mathrm{ff}} = \mathbf{n}_{\mathrm{pm}}$ is the unit vector normal to the fluid–porous interface Γ pointing outwards from the porous-medium domain Ω_{pm}, $\boldsymbol{\tau}$ is the unit vector tangential to the interface (Fig. 1), $\alpha_{\mathrm{BJ}} > 0$ is the Beavers–Joseph slip coefficient, and $\sqrt{K} = \sqrt{\boldsymbol{\tau} \cdot \mathbf{K} \boldsymbol{\tau}}$.

The generalised coupling conditions consist of the conservation of mass (3), an extension of the balance of normal forces (6) and a generalisation of the Beavers–

Joseph condition (7):

$$-\mathbf{n}\cdot\mathbf{T}\left(\mathbf{v}_{\mathrm{ff}},\,p_{\mathrm{ff}}\right)\mathbf{n}-\mu N_s^{\mathrm{bl}}\left(\nabla\mathbf{v}_{\mathrm{ff}}\,\mathbf{n}\right)\cdot\boldsymbol{\tau}=p_{\mathrm{pm}}\qquad\text{on } \varGamma, \tag{6}$$

$$\left(\mathbf{v}_{\mathrm{ff}}-\mathbf{v}_{\mathrm{pm}}^{\mathrm{int}}\right)\cdot\boldsymbol{\tau}-\varepsilon\left(\mathbf{N}^{\mathrm{bl}}\cdot\boldsymbol{\tau}\right)\left(\nabla\mathbf{v}_{\mathrm{ff}}\,\mathbf{n}\right)\cdot\boldsymbol{\tau}=0\qquad\text{on } \varGamma. \tag{7}$$

Here, the interfacial velocity is defined as

$$\mathbf{v}_{\mathrm{pm}}^{\mathrm{int}}=-\frac{\varepsilon^2\mathbf{M}^{\mathrm{bl}}}{\mu}\nabla p_{\mathrm{pm}},$$

and the boundary layer coefficients $N_s^{\mathrm{bl}}\in\mathbb{R}$, $\mathbf{N}^{\mathrm{bl}}=(N_1^{\mathrm{bl}},N_2^{\mathrm{bl}})^\top\in\mathbb{R}^2$ and $\mathbf{M}^{\mathrm{bl}}=(\mathbf{M}_i^{j,\mathrm{bl}})_{i,j=1,2}\in\mathbb{R}^{2\times 2}$ are computed numerically based on the theory of homogenisation and boundary layers [11].

3 Discretisation and Preconditioners

The coupled Stokes–Darcy problems (1)–(5) and (1)–(3), (6), (7) are discretised using the finite volume method on staggered grids (MAC scheme) [16]. The resulting systems of linear equations are of the form

$$\mathcal{A}\mathbf{x}=\mathbf{b},\qquad \mathcal{A}\in\{\mathcal{A}_{\mathrm{BJ}},\,\mathcal{A}_{\mathrm{ER}}\},\qquad \mathbf{x}=(\mathbf{v}_{\mathrm{ff}},\,p_{\mathrm{ff}},\,p_{\mathrm{pm}})^\top, \tag{8}$$

where $\mathcal{A}_{\mathrm{BJ}}$ corresponds to the discretised Stokes–Darcy problem with the classical coupling conditions (3)–(5) including the Beavers–Joseph condition and $\mathcal{A}_{\mathrm{ER}}$ corresponds to the generalised coupling conditions (3), (6), (7) derived in [11]. Both matrices are large, sparse and ill-conditioned.

While the discrete coupled Stokes–Darcy equations are of the more favourable double saddle point form for the Beavers–Joseph–Saffman interface condition [4, 7, 8], the matrices $\mathcal{A}_{\mathrm{BJ}}$ and $\mathcal{A}_{\mathrm{ER}}$ are nonsymmetric:

$$\mathcal{A}_{\mathrm{BJ}}=\begin{pmatrix} A_{\mathrm{BJ}} & B_{\mathrm{BJ},2}^\top & C_{\mathrm{BJ},2}^\top\\ B_{\mathrm{BJ},1} & 0 & 0\\ C_{\mathrm{BJ},1} & 0 & -D_{\mathrm{BJ}}\end{pmatrix},\qquad \mathcal{A}_{\mathrm{ER}}=\begin{pmatrix} A_{\mathrm{ER}} & B_{\mathrm{ER},2}^\top & C_{\mathrm{ER},2}^\top\\ B_{\mathrm{ER},1} & 0 & 0\\ C_{\mathrm{ER},1} & 0 & -D_{\mathrm{ER}}\end{pmatrix}. \tag{9}$$

To solve system (8) efficiently, we use flexible GMRES [14, Chap. 9.4.1]. For this algorithm, we have to use the right preconditioning

$$\mathcal{A}\mathcal{P}^{-1}\overline{\mathbf{x}}=\mathbf{b},\qquad \overline{\mathbf{x}}=\mathcal{P}\mathbf{x}. \tag{10}$$

We consider the block diagonal preconditioner

$$\mathcal{P}_{\text{diag}} = \begin{pmatrix} A & 0 & 0 \\ 0 & S_B & 0 \\ 0 & 0 & -\left(D + \sigma\, C_1 A^{-1} C_2^\top\right) \end{pmatrix} \tag{11}$$

based on the preconditioner developed in [7], the block triangular preconditioner

$$\mathcal{P}_{\text{triang}} = \begin{pmatrix} A & B_2^\top & 0 \\ 0 & -S_B & 0 \\ 0 & 0 & -\left(D + \sigma\, C_1 A^{-1} C_2^\top\right) \end{pmatrix} \tag{12}$$

based on [3] and the constraint preconditioner

$$\mathcal{P}_{\text{con}} = \begin{pmatrix} A & B_2^\top & 0 \\ B_1 & 0 & 0 \\ 0 & 0 & -\left(D + \sigma\, C_1 A^{-1} C_2^\top\right) \end{pmatrix} \tag{13}$$

based on [4, 8]. Here, $S_B := B_1 A^{-1} B_2^\top$ is the Schur complement, $A \in \{A_{\text{BJ}}, A_{\text{ER}}\}$, $B_i \in \{B_{\text{BJ},i}, B_{\text{ER},i}\}$, $C_i \in \{C_{\text{BJ},i}, C_{\text{ER},i}\}$, for $i = 1, 2$, and $D \in \{D_{\text{BJ}}, D_{\text{ER}}\}$. The constant $\sigma \geq 0$ is fitted for each problem formulation.

4 Numerical Results

4.1 Benchmark Model

We study a flow scenario where the flow is arbitrary to the fluid–porous interface $\varGamma$. The coupled domain is divided into the free-flow region $\varOmega_{\text{ff}} = (0, 1) \times (0, 0.5)$ and the porous-medium domain $\varOmega_{\text{pm}} = (0, 1) \times (0, -0.5)$ that are separated by the interface $\varGamma = (0, 1) \times \{0\}$. To get a closed model, the following boundary conditions are implemented on the external boundary,

$$\begin{aligned} \mu(\nabla \mathbf{v} - p\mathbf{I})\mathbf{n} &= 0 && \text{on } \varGamma_{\text{out}}, \\ \mu(\nabla \mathbf{v} - p\mathbf{I})\mathbf{n} &= -p_b \mathbf{n} && \text{on } \{y = -0.5\}, \\ \mathbf{v} &= \mathbf{0} && \text{on } \varGamma_{\text{wall}}, \\ \mathbf{v} &= (0, -0.2\sin(\pi x)) && \text{on } \varGamma_{\text{in}}, \end{aligned} \tag{14}$$

for $\varGamma_{\text{out}} = \{x = 0\} \times (0, 0.1) \cup \{x = 1\} \times (0, 0.5)$, $\varGamma_{\text{in}} = \{y = 0.5\}$, $\varGamma_{\text{wall}} = (\{x = 0\} \cup \{x = 1\}) \backslash \varGamma_{\text{out}}$ and $p_b = 10^{-6} - x$ (Fig. 2). Due to the flow being arbitrary to the fluid–porous interface $\varGamma$, the Beavers–Joseph coupling conditions (3)–(5) are not suitable and the generalised coupling conditions (3), (6), (7) are recommended.

We consider an isotropic porous medium constructed by periodically distributed circular solid inclusions of radius $r = 0.25\varepsilon$, where ε is the scale separation parameter. The permeability k_{11} and the boundary layer constants $M_1^{1,\text{bl}}$ and N_1^{bl} from the

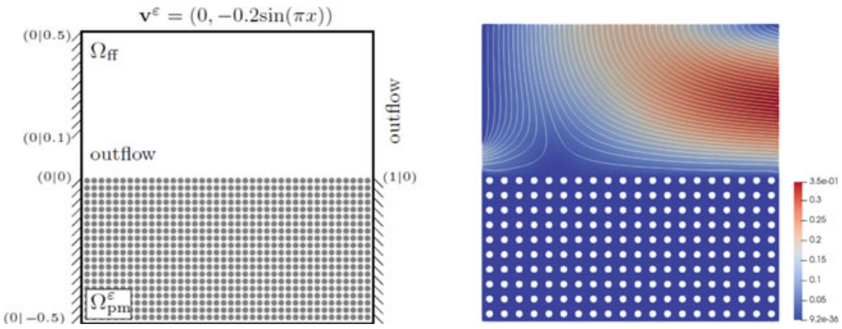
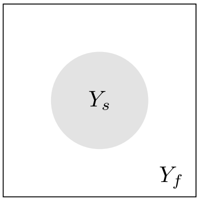


Fig. 2 Flow problem for the numerical results

Fig. 3 Permeability k_{11} and boundary layer constants $M_1^{1,\mathrm{bl}}$ and N_1^{bl} of the geometry



k_{11}	1.99×10^{-2}
$M_1^{1,\mathrm{bl}}$	-3.02×10^{-3}
N_1^{bl}	-5.48×10^{-2}

generalised coupling conditions (3), (6), (7) corresponding to the considered geometry are given in Fig. 3. The permeability k_{11} and the boundary layer constant $M_1^{1,\mathrm{bl}}$ have to be scaled by ε^2.

4.2 Robustness and Efficiency Analysis

To show the robustness and efficiency of the exact versions of the preconditioners $\mathcal{P}_{\mathrm{diag}}$ given in (11), $\mathcal{P}_{\mathrm{triang}}$ given in (12), and $\mathcal{P}_{\mathrm{con}}$ given in (13) we consider three different values for the viscosity $\mu \in \{10^{-5}, 10^{-3}, 1\}$ and two different scale separation parameters $\varepsilon \in \{1/20, 1/200\}$ which yields different permeability tensors $\mathbf{K} = k_{11}\mathbf{I}$ and different boundary layer constants $M_1^{1,\mathrm{bl}}$. We consider the grid width $h = 1/80$ and set $\alpha_{\mathrm{BJ}} = 1$ for the Beaver–Joseph coupling condition (5) as routinely used in the literature. We fit the parameter σ by determining the optimal $\sigma \in \{0, 1, \ldots, 10\}$ using a brute-force search. The Stokes–Darcy problem is discretised using our in-house C++ software. We solve the preconditioned system (8) with the restarted flexible GMRES method using 20 restarts in Matlab. The initial solution $\mathbf{x}_0$ is the zero vector. The iterations are stopped once $\|\mathcal{A}\mathbf{x}_n - \mathbf{b}\|_2 \le 10^{-8}\|\mathbf{b}\|_2$ is reached or after $n_{\max} = 2000$ iteration steps. All computations were carried out on a laptop with an AMD Ryzen™ 5 2500U processor and 12.0GB RAM using MATLAB.R2019b. The number of iterations until the flexible GMRES algorithm reaches the given tolerance are displayed in Table 1. In Fig. 4 we plot the relative

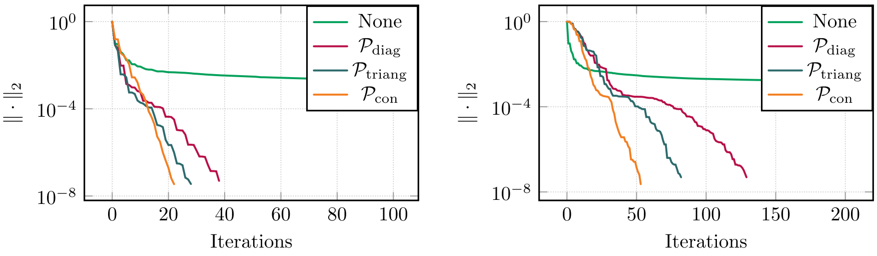


Fig. 4 Relative residuals for the classical conditions (3)–(5) (left) and the generalised coupling conditions (3), (6), (7) (right)

Table 1 Iterations for different values of μ and ε appearing in k_{11} and $M_1^{1,\text{bl}}$ in the coupled Stokes–Darcy system with $h = 1/80$

		Preconditioners for $\mathcal{A}_{\text{BJ}}$						Preconditioners for $\mathcal{A}_{\text{ER}}$					
ε	μ	σ	$\mathcal{P}_{\text{diag}}$	σ	$\mathcal{P}_{\text{triang}}$	σ	$\mathcal{P}_{\text{con}}$	σ	$\mathcal{P}_{\text{diag}}$	σ	$\mathcal{P}_{\text{triang}}$	σ	$\mathcal{P}_{\text{con}}$
1/20	10^{-5}	2	49	1	33	2	27	0	139	0	91	0	60
1/20	10^{-3}	2	38	1	28	2	22	0	129	0	82	0	53
1/20	1	1	41	1	27	2	23	0	124	0	75	0	61
1/200	10^{-5}	1	31	1	24	2	20	1	135	10	107	3	69
1/200	10^{-3}	1	31	1	23	2	19	3	135	3	100	0	65
1/200	1	2	35	1	25	2	22	6	135	3	100	2	74

residuals $\|\mathcal{A}\mathbf{x}_n - \mathbf{b}\|_2/\|\mathcal{A}\mathbf{x}_0 - \mathbf{b}\|_2$ against the number of iterations n for $\varepsilon = 1/20$ and $\mu = 10^{-3}$.

It can be seen in Fig. 4 that all three preconditioners significantly reduce the number of iteration steps. The needed CPU time to solve problem (8) is decreased using the developed preconditioners. Here, we choose the parameters $\varepsilon = 1/20$ and $\mu = 10^{-3}$. To solve the nonpreconditioned system in the case of the Beavers–Joseph coupling condition 1207 seconds are needed. The diagonal preconditioner $\mathcal{P}_{\text{diag}}$ reduces the CPU time to 54.01 s, the triangular preconditioner $\mathcal{P}_{\text{triang}}$ needs 52.53 s, and the constraint preconditioner $\mathcal{P}_{\text{con}}$ requires 42.64 s. For the generalised coupling conditions 1413 s are necessary to solve the system without preconditioning. It takes 97.22 s to solve the system with the diagonal, 78.70 s with the triangular and 72.22 s with the constraint preconditioner. Furthermore, all preconditioners show a high robustness with respect to changes in the viscosity μ, the permeability k_{11} and the boundary layer constant $M_1^{1,\text{bl}}$ as shown in Table 1. While all preconditioners provide an improvement of the convergence, the constraint-preconditioned system $\mathcal{A}\mathcal{P}_{\text{con}}^{-1}$ needs the fewest number of iteration steps and the smallest CPU time until convergence for every considered case. Therefore, in practical applications the use of $\mathcal{P}_{\text{con}}$ is recommended.

5 Conclusions and Future Work

In this paper, we have considered the coupled Stokes–Darcy system with the classical set of interface conditions comprising the Beavers–Joseph coupling condition and with generalised interface conditions that were recently developed for arbitrary flows to the interface. To discretise the coupled problem the MAC scheme was used. We have suggested and evaluated three preconditioners: a block diagonal $\mathcal{P}_{\text{diag}}$, a block triangular $\mathcal{P}_{\text{triang}}$, and a constraint preconditioner $\mathcal{P}_{\text{con}}$. The robustness and efficiency of these preconditioners were shown in numerical experiments for a benchmark model with arbitrary flow to the interface. This work was especially focused on the monolithic solution of the governing coupled system and the development of suitable preconditioners to the latter. In a further step we will compare the efficiency of the monolithic approach to a partitioned coupling one. For the system resulting from the classical conditions this has already been done in a comparative study in [15] where the coupling library preCICE [9] was used for the partitioned case. The incorporation of the generalised coupling conditions into preCICE and a comparison of the solution approaches are part of future work. Furthermore, we strive for applying the preconditioners to systems resulting from real-world scenarios. Therefore, we want to investigate their scalability with respect to both, runtime and memory consumption.

Acknowledgements The work is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project Number 327154368 – SFB 1313, Project Number 490872182 and under Germany's Excellence Strategy - EXC 2075 - 390740016. We acknowledge the support by the Stuttgart Center for Simulation Science (SC SimTech).

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). https://doi.org/10.1103/PhysRevE.95.063302
2. Beavers, G., Joseph, D.: Boundary conditions at a naturally permeable wall. J. Fluid Mech. **30**, 197–207 (1967). https://doi.org/10.1017/S0022112067001375
3. Beik, F.P.A., Benzi, M.: Iterative methods for double saddle point systems. SIAM J. Matrix Anal. Appl. **39**, 902–921 (2018). https://doi.org/10.1137/17M1121226
4. Beik, F.P.A., Benzi, M.: Preconditioning techniques for the coupled Stokes-Darcy problem: spectral and field-of-values analysis. Numer. Math. **150**, 257–298 (2022). https://doi.org/10.1007/s00211-021-01267-8
5. Benzi, M., Golub, G., Liesen, J.: Numerical solution of saddle point problems. Acta Numer **14**, 1–137 (2005). https://doi.org/10.1017/S0962492904000212
6. Boon, W.M., Koch, T., Kuchta, M., et al.: Robust monolithic solvers for the Stokes-Darcy problem with the Darcy equation in primal form. SIAM J. Sci. Comput. **44**, B1148–B1174 (2022). https://doi.org/10.1137/21M1452974
7. Cai, M., Mu, M., Xu, J.: Preconditioning techniques for a mixed Stokes/Darcy model in porous media applications. J. Comput. Appl. Math. **233**, 346–355 (2009). https://doi.org/10.1016/j.cam.2009.07.029

8. Chidyagwai, P., Ladenheim, S., Szyld, D.B.: Constraint preconditioning for the coupled Stokes-Darcy system. SIAM J. Sci. Comput. **38**, A668–A690 (2016). https://doi.org/10.1137/15M1032156
9. Chourdakis, G., Davis, K., Rodenberg, B. et al.: preCICE v2: A sustainable and user-friendly coupling library. Open Res. Europe **51** (2022). https://doi.org/10.12688/openreseurope.14445.2
10. Discacciati, M., Quarteroni, A.: Navier-Stokes/Darcy coupling: modeling, analysis, and numerical approximation. Rev. Mat. Complut. **22**, 315–426 (2009). https://doi.org/10.5209/rev_REMA.2009.v22.n2.16263
11. Eggenweiler, E., Rybak, I.: Effective coupling conditions for arbitrary flows in Stokes-Darcy systems. Multiscale Model. Simul. **19**, 731–751 (2021). https://doi.org/10.1137/20M1346638
12. Holter, K.E., Kuchta, M., Mardal, K.-A.: Robust preconditioning for coupled Stokes-Darcy problems with the Darcy problem in primal form. Comput. Math. Appl. **91**, 53–66 (2021). https://doi.org/10.1016/j.camwa.2020.08.021
13. Layton, W.J., Schieweck, F., Yotov, I.: Coupling fluid flow with porous media flow. SIAM J. Numer. Anal. **40**, 2195–2218 (2003). https://doi.org/10.1137/S0036142901392766
14. Saad, Y.: Iterative Methods for Sparse Linear Systems. SIAM (2003)
15. Schmalfuss, J., Riethmüller, C., Altenbernd, M., Weishaupt, K., Göddeke, D.: Partitioned coupling vs. monolithic block-preconditioning approaches for solving Stokes–Darcy systems. In: Proceedings of International Conference on Computational Methods for Coupled Problems in Science and Engineering (COUPLED PROBLEMS) (2021). https://doi.org/10.23967/coupled.2021.043
16. Versteeg, H.K., Malalasekera, W.: An Introduction to Computational Fluid Dynamics: The Finite Volume Method. Pearson Education (1995)

A DDFV Scheme for Incompressible Two-Phase Flow Degenerate Problem in Porous Media

Thomas Crozon, El-Houssaine Quenjel, and Mazen Saad

Abstract We propose a Discrete Duality Finite Volume (DDFV) scheme to discretize the degenerate system modeling incompressible immiscible two-phase Darcy flow in porous media problem. This method allows general meshes with fewer limitations, the upwind mobility term in the discretization combined with minimum mobility in the cross term and the degeneracy give a maximum principle. Moreover, we establish some energy estimates on the approximate solutions, an existence result, and we give numerical tests to illustrate the efficiency of the proposed scheme.

Keywords Degenerate incompressible two-phase · Darcy flow · Finite volume scheme · DDFV scheme

1 Introduction

In this paper, we build a Discrete Duality Finite Volume scheme for the incompressible immiscible two-phase Darcy flow continuous model [3]. One particular aspect is to cope with degeneracy coming from the mobilities. It has been done in various works, with convergence results. For instance in [7, 11] the authors use a Two Point Flux Approximation on orthogonal meshes. The fluxes are approximated using upwind approach with regard to the phase pressures or centered approximation for the capillary term and upwind for the convective one. A similar idea is developed in

T. Crozon (✉) · M. Saad
Ecole Centrale Nantes, Laboratoire de mathématiques Jean Leray, CNRS UMR 6629, 1 rue de la Noë, 44300 Nantes, France
e-mail: thomas.crozon@ec-nantes.fr

M. Saad
e-mail: mazen.saad@ec-nantes.fr

E.-H. Quenjel
Université Paris-Saclay, CentraleSupélec, LGPM, CEBB, 3 rue des Rouges Terres, 51110 Pomacle, France
e-mail: el-houssaine.quenjel@centralesupelec.fr

E. Franck et al. (eds.), *Finite Volumes for Complex Applications X—Volume 1, Elliptic and Parabolic Problems*, Springer Proceedings in Mathematics & Statistics 432,
https://doi.org/10.1007/978-3-031-40864-9_33

[8] with a Control Volume Finite Elements discretization on simplicial conforming meshes. It uses a sub-upwinding scheme with regard to the stiffness coefficient and the phase pressure to keep the physical bounds and energy estimates, also showing existence and convergence results. In [1] a Vertex Approximate Gradient scheme is constructed for general meshes using upwind approach, this scheme is convergent but it does not guarantee the maximum principle on the saturation. In this paper, we build a scheme on general meshes. The originality of our scheme is to propose two different upwind approximations, in the normal direction we use an upwind approximation of mobilities with respect to the discrete gradient of the phase pressure and in the tangential direction the mobilities are split into two parts in which an upwind and a minimum of approximation are used. These approximations allow the maximum principle on the saturations. We then show some energy estimates and the existence of the approximate solution. Finally, we exhibit numerical tests to show the efficiency of our method.

2 Continuous Model

Let $t_f > 0$ and Ω be a bounded connected subset of $\mathbb{R}^2$. The incompressible two-phase Darcy flow reads

$$\begin{cases} \phi(\mathbf{x})\partial_t s_\alpha + \operatorname{div}\left(\vec{V}^\alpha\right) = 0, \\ \vec{V}^\alpha = -\dfrac{K_{\mathrm{r}\alpha}(s_\alpha)}{\mu_\alpha}\Lambda(\mathbf{x})\nabla p_\alpha, \qquad \forall\alpha \in \{nw, w\}, \\ P_c(s_{nw}) = p_{nw} - p_{\mathrm{w}}, \\ s_{\mathrm{w}} + s_{nw} = 1 \end{cases} \tag{1}$$

with nw, w denoting resp. the non-wetting and aqueous 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, $\vec{V}^\alpha$ the phase velocity, and P_c the capillary pressure. We neglect the gravity effects (as in a horizontal slice of the medium) and source-sink rate terms since their contributions can be added without technical difficulties. The phase mobility function $M_\alpha(s_\alpha)$ is defined as the ratio of the relative permeability of the phase $K_{\mathrm{r}\alpha}$ over its viscosity μ_α. The mobility is increasing with the saturation and we will consider the continuous constant extension outside of $[0, 1]$. Notice $M_\alpha(s_\alpha = 0) = 0$ is known as the degeneracy issue. It is assumed that P_c is increasing and $P_c(s_{\mathrm{nw}} = 0) = 0$; moreover the total mobility function $M(s) = M_{nw}(s) + M_{\mathrm{w}}(1 - s)$ verifies $M(s) \geq m_0 > 0$, where s is the non-wetting saturation. The porosity verifies $0 < \phi_0 \leq \phi(x) \leq \phi_1$ a.e. (almost everywhere). Also, the permeability is a symmetric positive-definite matrix and essentially bounded, in addition to uniformly elliptic i.e. there exist constants $\underline{\Lambda} > 0$ and $\overline{\Lambda}$ such that $\underline{\Lambda}|z|^2 \leq \Lambda(x)z \cdot z \leq \overline{\Lambda}|z|^2$ for all $z \in \mathbb{R}^2$ and for almost every x. The system (1) is completed by some initial distribution of p_α, in addition to the boundary conditions

$$\mathbf{V}^{\alpha}\cdot\mathbf{n}=0 \quad \text{on } \Gamma_N\times(0,t_f), \quad p_\alpha=p_{\alpha,\mathrm{Dir}} \quad \text{on } \Gamma_{\mathrm{Dir}}\times(0,t_f) \quad \text{for } \alpha\in\{w,nw\},$$

where $\mathbf{n}$ is the outward normal to Γ_N, and $\partial\Omega=\Gamma_N\cup\Gamma_{\mathrm{Dir}}$ with $|\Gamma_{\mathrm{Dir}}|>0$.

The global pressure p introduced in [3] is given by $p=0$ when $s_{nw}=0$ and the relation $M(s_{nw})\nabla p=M_w(s_w)\nabla p_w+M_{nw}(s_{nw})\nabla p_{nw}$. Used with the capillary term $\xi(s_{nw})=\displaystyle\int_0^{s_{nw}}\frac{\sqrt{M_w(1-u)M_{nw}(u)}}{M(u)}P_c'(u)\,\mathrm{d}u$, we find energy estimates to overcome the degeneracies.

3 DDFV Discretization and Discrete Operators

In the DDFV framework, we use three different meshes of $\Omega\subset\mathbb{R}^2$. Let us recall briefly the meshes and their notations [2, 4, 6, 9].

The primal mesh denoted by $\overline{\mathfrak{M}}$ is composed of $\mathfrak{M}$ the interior primal mesh, a partition of Ω made of disjoint open polygonal cells, and $\partial\mathfrak{M}$ the set of boundary edges seen as degenerated cells. For each cell $K\in\overline{\mathfrak{M}}$ one defines a unique point x_K called its center. Usually we take the barycenter, for the boundary cells we take the midpoint as the center.

The dual mesh is constructed from the vertices of the primal mesh. To any interior vertex x_{K^*}, we define a polygonal control volume K^* built by connecting all the centers of the primal cells sharing x_{K^*} as a vertex. The set of cells is called the dual mesh and written $\mathfrak{M}^*$. When the vertex x_{K^*} is on $\partial\Omega$, we construct a dual cell connecting x_{K^*} with of the primal cell centers x_K and the midpoint of the boundary edges sharing x_{K^*} as vertex. This last collection is the boundary dual mesh, denoted $\partial\mathfrak{M}^*$. Finally, one has the dual mesh $\overline{\mathfrak{M}^*}=\mathfrak{M}^*\cup\partial\mathfrak{M}^*$.

For two cells K and L, we assume that $\partial K\cap\partial L$ is either empty, a vertex, or a segment. In this last case, we write $\sigma=K|L$, this is the case of convex cells for simplicity. We denote $\mathcal{E}$ the set of the edges of the primal mesh, and $\mathcal{E}_K$ stands for the edges of K. Likewise, we define $\mathcal{E}^*$ the edges of the dual mesh, and following $\mathcal{E}^*_{K^*}$ are the edges of K^*.

The diamond mesh is based on the segments of those two previous meshes. For each edge $\sigma=K|L$ having x_{K^*} and x_{L^*} as vertices, denoting $\sigma^*=K^*|L^*$, we define the quadrilateral diamond $\mathcal{D}_{\sigma,\sigma^*}$ joining x_K, x_{K^*}, x_L and x_{L^*} (if $\sigma\subset\partial\Omega$ then it is a triangle see Fig. 1). The set of the diamond is a partition of Ω, it gives the diamond mesh $\mathfrak{D}$. In the end the DDFV mesh is composed of $\mathcal{T}=(\overline{\mathfrak{M}},\overline{\mathfrak{M}^*})$ and $\mathfrak{D}$ (see Fig. 1).

For every cell $K\in\overline{\mathfrak{M}}$ or $K^*\in\overline{\mathfrak{M}^*}$, m_K and m_{K^*} designate their measure. In addition d_K and d_{K^*} are the diameters of the control volumes. For a diamond $\mathcal{D}=\mathcal{D}_{\sigma,\sigma^*}$, with x_K, x_L, x_{K^*} and x_{L^*} as vertices, we define m_σ and m_{σ^*} the lengths of the edges, $m_{\mathcal{D}}$ its measure, $d_{\mathcal{D}}$ its diameter, $\alpha_{\mathcal{D}}$ is the angle between (x_K,x_L) and (x_{K^*},x_{L^*}). We have $m_{\mathcal{D}}=\frac{1}{2}m_\sigma m_{\sigma^*}\sin(\alpha_{\mathcal{D}})$. We will use $\mathbf{n}_{\sigma,K}$ the outward unit normal to σ, as well we have $\mathbf{n}_{\sigma^*,K^*}$.

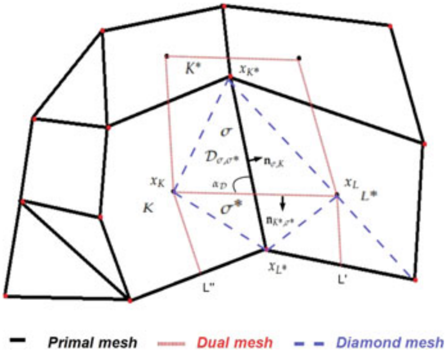


Fig. 1 DDFV mesh example with notations

Furthermore we define some regularity quantification of the mesh that gives information on its shape. In other words, it ensures that any two edges are comparable and excludes small interfaces or degenerate diamond subdomains. One defines $\alpha_\mathcal{T}$ the real number in $]0, \frac{\pi}{2}]$ satisfying $\sin(\alpha_\mathcal{T}) = \min\limits_{\mathcal{D}\in\mathfrak{D}} |\sin(\alpha_\mathcal{D})|$. We define and assume some regularities conditions according to [10], and the mesh size $h_\mathcal{T} = \max_{A\in\mathcal{T}} d_A$.

We denote by $\mathbb{R}^\mathcal{T}$ the space comprising the elements of the vector form $u_\mathcal{T} = ((u_K)_{K\in\overline{\mathfrak{M}}}, (u_{K^*})_{K^*\in\overline{\mathfrak{M}^*}})$. In the DDFV approach, the discrete gradient operator is a linear mapping from $\mathbb{R}^\mathcal{T}$ to $(\mathbb{R}^2)^\mathfrak{D}$, its purpose is to mimic a gradient [4]. $\nabla^\mathfrak{D} u_\mathcal{T}$ is defined, for every $u_\mathcal{T} \in \mathbb{R}^\mathcal{T}$, constant on each diamonds cells $\mathcal{D}$ by

$$\nabla^\mathcal{D} u_\mathcal{T} = \frac{1}{\sin(\alpha_\mathcal{D})}\left(\frac{u_L - u_K}{m_{\sigma^*}}\mathbf{n}_{\sigma,K} + \frac{u_{L^*} - u_{K^*}}{m_\sigma}\mathbf{n}_{\sigma^*,K^*}\right) \quad , \quad \forall\mathcal{D}\in\mathfrak{D}.$$

We consider the space $X_{\mathcal{T},\delta t}$ of piecewise constant time dependent vectors of $\mathbb{R}^\mathcal{T}$. One writes $u_{\mathcal{T},\delta t}(t) = u^n_\mathcal{T} \in \mathbb{R}^\mathcal{T}$ for t in $(t^n, t^{n+1}]$. Then, we can define a piecewise constant in time gradient $\nabla^\mathfrak{D} u_{\mathcal{T},\delta t}$ and the following discrete L^2-norm

$$\left\|\nabla^\mathfrak{D} u_{\mathcal{T},\delta t}\right\|_2 = \left(\sum_{n=0}^{N-1}\delta t \sum_{\mathcal{D}\in\mathfrak{D}} m_\mathcal{D}|\nabla^\mathfrak{D} u^n_\mathcal{T}|^2\right)^{\frac{1}{2}} \quad \text{with} \quad t_f = N\delta t.$$

4 DDFV Scheme for System (1)

In this paper, we want to take into account Dirichlet boundary conditions on the pressures. We suppose that the vertices of Γ_Dir are vertices of the primal mesh, then the centers of the boundary primal cells are exclusively in $\overline{\Gamma_\text{Dir}}$ or exclusively in Γ_N. We will separate $\partial\mathfrak{M}$ and $\partial\mathfrak{M}^*\backslash\overline{\Gamma_\text{Dir}}$

$$\partial\mathfrak{M}_\text{Dir} = \left\{K \in \partial\mathfrak{M}, x_K \in \overline{\Gamma_\text{Dir}}\right\} , \partial\mathfrak{M}_\text{N} = \left\{K \in \partial\mathfrak{M}, x_K \in \Gamma_\text{N}\backslash\overline{\Gamma_\text{Dir}}\right\} ,$$

$$\partial\mathfrak{M}^*_{\mathrm{Dir}} = \left\{K^* \in \partial\mathfrak{M}^*, x_{K^*} \in \overline{\Gamma_{\mathrm{Dir}}}\right\} \text{ and } \partial\mathfrak{M}^*_N = \left\{K^* \in \partial\mathfrak{M}^*, x_{K^*} \in \Gamma_{\mathrm{N}}\backslash\overline{\Gamma_{\mathrm{Dir}}}\right\}.$$

For $N \in \mathbb{N}^*$, we consider the time subdivision $t^0 = 0 < t^1 < ... < t^{n-1} < t^n < ... < t^N = t_f$ of $[0, t_f[$. The will take the time step $\delta t = t^n - t^{n-1}$ constant. We perform an implicit method with a DDFV discretization of each phase flow.

Furthermore we first discretize the initial condition by taking the mean values of $s_{nw,0}$ and $s_{w,0}$ on the primal and dual cells. Then, for all $n \geq 0$ we look for $(p_{nw}^{\mathrm{n+1}}, p_w^{\mathrm{n+1}})$ in $X_{\mathcal{T},\delta t}$, or equivalently (s_{nw}, p_w), solution to the system of discrete equations (2)–(7). One writes the discrete primal equations, where the $V_{KL}^{\alpha,n+1}$ are the projected velocity at the interface $\sigma = K|L$ (resp. $V_{K^*L^*}^{\alpha,n+1}$ for $\sigma^* = K^*|L^*$):

$$\phi_K(s_{\alpha,K}^{n+1} - s_{\alpha,K}^n) - \frac{\delta t}{m_K} \sum_{\sigma=K|L\in\mathcal{E}_K} V_{KL}^{\alpha,n+1} = 0 \quad , \forall\alpha \in \{nw, w\}, \forall K \in \mathfrak{M}, \tag{2}$$

and its dual counterpart for every $K^* \in \mathfrak{M}^* \cup \partial\mathfrak{M}^*_N$

$$\phi_{K^*}(s_{\alpha,K^*}^{n+1} - s_{\alpha,K^*}^n) - \frac{\delta t}{m_{K^*}} \sum_{\sigma=K^*|L^*\in\mathcal{E}^*_{K^*}} V_{K^*L^*}^{\alpha,n+1} = 0 \quad , \forall\alpha \in \{nw, w\}. \tag{3}$$

We keep the relations between the saturations:

$$s_{nw,K}^{n+1} + s_{w,K}^{n+1} = 1, \quad s_{nw,K^*}^{n+1} + s_{w,K^*}^{n+1} = 1 \quad , \forall K \in \overline{\mathfrak{M}}, \forall K^* \in \overline{\mathfrak{M}^*}. \tag{4}$$

Then we close the discrete system with link between the pressures and the saturations $\forall K \in \overline{\mathfrak{M}}, \forall K^* \in \overline{\mathfrak{M}^*}$

$$P_c(s_{nw,K}^{n+1}) = p_{nw,K}^{n+1} - p_{w,K}^{n+1} \quad \text{and} \quad P_c(s_{nw,K^*}^{n+1}) = p_{nw,K^*}^{n+1} - p_{w,K^*}^{n+1}. \tag{5}$$

In addition we have the discrete Neumann (6) and Dirichlet (7) boundary conditions.

$$V_{\alpha,KL}^{n+1} = 0 \quad , \forall\alpha \in \{nw, w\}, \forall K \in \partial\mathfrak{M}_{\mathrm{N}}, \tag{6}$$

$$p_{\alpha,K}^{n+1} = 0 \quad \text{and} \quad p_{\alpha,K^*}^{n+1} = 0 \quad , \forall\alpha \in \{nw, w\}, \forall K \in \partial\mathfrak{M}_{Dir}, \forall K^* \in \partial\mathfrak{M}^*_{Dir}. \tag{7}$$

The novelty of our contribution is the expression of the projected velocity at the interface $\sigma = K|L$ (resp. for $\sigma^* = K^*|L^*$)

$$\begin{aligned} &\int_{\sigma=K|L} M_\alpha(s_\alpha)\Lambda\nabla p_\alpha \cdot n_{KL}\, \mathrm{d}\sigma(s) \approx V_{KL}^{\alpha,n+1} := \\ &M_{\alpha,KL}^{up,n+1}\tau_{KL}(p_{\alpha,L}^{n+1} - p_{\alpha,K}^{n+1}) + \sqrt{M_{\alpha,KL}^{min,n+1}}\sqrt{M_{\alpha,K^*L^*}^{up,n+1}}\eta_{\mathcal{D}}(p_{\alpha,L^*}^{n+1} - p_{\alpha,K^*}^{n+1}) \end{aligned}, \tag{8}$$

$$V_{K^*L^*}^{\alpha,n+1} = M_{\alpha,K^*L^*}^{up,n+1}\tau_{K^*L^*}(p_{\alpha,L^*}^{n+1} - p_{\alpha,K^*}^{n+1}) + \sqrt{M_{\alpha,K^*L^*}^{min,n+1}}\sqrt{M_{\alpha,KL}^{up,n+1}}\eta_{\mathcal{D}}(p_{\alpha,L}^{n+1} - p_{\alpha,K}^{n+1}). \tag{9}$$

The mobility is distributed following the normal and the tangential components of the approximate gradient. We apply different upwind approaches with respect to the sign of the flow of each phase. The presence of the square root is to force the coercivity of the scheme i.e. there exists $\gamma > 0$ depending only on the mesh regularity such that

$$V_{KL}^{\alpha,n+1}(p_{\alpha,L}^{n+1} - p_{\alpha,K}^{n+1}) + V_{K^*L^*}^{\alpha,n+1}(p_{\alpha,L^*}^{n+1} - p_{\alpha,K^*}^{n+1}) \geq \gamma(\tau_{KL}M_{\alpha,KL}^{up,n+1}(p_{\alpha,L}^{n+1} - p_{\alpha,K}^{n+1})^2 \\ + \tau_{K^*L^*}M_{\alpha,K^*L^*}^{up,n+1}(p_{\alpha,L^*}^{n+1} - p_{\alpha,K^*}^{n+1})^2).$$

In the crossed term the modified mobilities $M_{\alpha,KL}^{min,n+1}$ are particularly chosen to keep the maximum principle (resp. for $M_{\alpha,K^*L^*}^{up,n+1}$ and $M_{\alpha,K^*L^*}^{min,n+1}$)

$$M_{\alpha,KL}^{up,n+1} := \begin{cases} M_\alpha(s_{\alpha,L}^{n+1}) \text{ if } p_{\alpha,L}^{n+1} - p_{\alpha,K}^{n+1} \geq 0 \\ \\ M_\alpha(s_{\alpha,K}^{n+1}) \text{ otherwise} \end{cases}, M_{\alpha,KL}^{min,n+1} := M_\alpha(\min(s_{\alpha,K}^{n+1}, s_{\alpha,L}^{n+1})). \tag{10}$$

We take for the permeability or stiffness tensor on the diamond its mean-value on $\mathcal{D}$. For the porosity approximation, one takes as mean value on the cell

$$\Lambda_{\mathcal{D}} = \frac{1}{m_{\mathcal{D}}}\int_{\mathcal{D}} \Lambda(x)\,\mathrm{d}x, \quad \phi_K = \frac{1}{m_K}\int_K \phi(x)\,\mathrm{d}x. \tag{11}$$

We now can give the transmissibility coefficients $\quad \tau_{KL} = \dfrac{m_\sigma}{m_{\sigma^*}}\dfrac{\langle\Lambda_{\mathcal{D}}\mathbf{n}_{KL}, \mathbf{n}_{KL}\rangle}{\sin(\alpha_{\mathcal{D}})}$,

$$\tau_{K^*L^*} = \frac{m_{\sigma^*}}{m_\sigma}\frac{\langle\Lambda_{\mathcal{D}}\mathbf{n}_{K^*L^*}, \mathbf{n}_{K^*L^*}\rangle}{\sin(\alpha_{\mathcal{D}})} \quad , \quad \eta_{\mathcal{D}} = \frac{\langle\Lambda_{\mathcal{D}}\mathbf{n}_{KL}, \mathbf{n}_{K^*L^*}\rangle}{\sin(\alpha_{\mathcal{D}})}. \tag{12}$$

5 Energy Estimates

We now state two properties of the scheme (2)–(12) and an existence result. We refer to the forthcoming article [5] for more details.

Lemma 1 (Maximum principle) *Let $(p_{nw,\mathcal{T},\delta t}, p_{w,\mathcal{T},\delta t})$ be a solution to the numerical scheme (see Sect. 4). Then, for $\alpha \in \{nw, w\}$, the discrete saturation of the α-phase obeys its physical bounds i.e.*

$$0 \leq s_{\alpha,K}^{n+1} \leq 1, \quad \forall K \in \mathcal{T}, \forall n \in [\![0\,;\, N-1]\!].$$

We multiply the discrete equation by $(s_{nw,K}^{n+1})^- = -\min(s_{nw,K}^{n+1}, 0) \geq 0$, for the minimizing cell of the mesh for s_{nw}. One has, thanks to the degeneracy $M_{nw}(s_{nw,K})(s_{nw,K})^- = 0$. For the cross-term of the flow through $\sigma = K|L$ we have:

$$\sqrt{M_{nw,KL}^{min,n+1}}\sqrt{M_{nw,K^*L^*}^{up,n+1}}(p_{nw,L^*}^{n+1} - p_{nw,K^*}^{n+1})(s_{nw,K}^{n+1})^- = \sqrt{M_{nw}(s_{nw,K}^{n+1})}(s_{nw,K}^{n+1})^- \\ \times \sqrt{M_{nw,K^*L^*}^{up,n+1}}(p_{nw,L^*}^{n+1} - p_{nw,K^*}^{n+1}) = 0$$

Then, handling the rest as in [11], we deduce the physical bound. The following lemma in addition to a norm equivalence [10] leads to the energy estimates. We adapt the pathway of [8] to the DDFV-framework to find (13).

Lemma 2 *For every K, L neighbours in $\mathfrak{M}$ (resp. K^*, L^* in $\overline{\mathfrak{M}^*}$), there holds:*

$$m_0((p_L - p_K)^2 + (\xi_L - \xi_K))^2 \leq M_{nw,KL}^{up}\left(p_{nw,L} - p_{nw,K}\right)^2 + M_{w,KL}^{up}\left(p_{w,L} - p_{w,K}\right)^2.$$

Proposition 1 (Energy estimates) *Let $(p_{\alpha,\mathcal{T},\delta t})_{\alpha\in\{nw,w\}}$ be solution of our scheme* (2)–(12). *Then there exists a constant C independent of the discretization parameters $h_{\mathcal{T}}$ and δt, depending on the mesh regularity, such that*

$$\left\|\nabla^{\mathfrak{D}} p_{\mathcal{T},\delta t}\right\|_2^2 + \left\|\nabla^{\mathfrak{D}} \xi_{\mathcal{T},\delta t}\right\|_2^2 \leq C. \tag{13}$$

Proposition 2 (Existence) *The scheme* (2)–(12) *admits at least a solution.*

6 Numerical Results

We test our scheme on a framework similar to the one used in [8, 11]. We simulate second recovery of oil in an isotropic and anisotropic 2D reservoirs. Our domain of interest is $\Omega = (0, 1m)^2$, with the uniform porosity $\phi = 0.206$. The relative permeabilities are squared functions of the saturation $k_{r,\alpha}(s_\alpha) = (s_\alpha)^2$, with fluid viscosities given: $\mu_{nw} = 9 \times 10^{-5}\ Pa\ s$ and $\mu_w = 10^{-3}\ Pa\ s$. The capillary pressure is a linear function in terms of the oil saturation $P_c(s_{nw}) = P_{max}s_{nw}$ with $P_{max} = 10^5\ Pa$. The domain is initially occupied by oil with uniform saturation, where the oil is at the standard atmospheric pressure $P_{atm} = 1.013 \times 10^5\ Pa$. Under the pressure $p_w^{left} = 4.6732 \times 10^5\ Pa$, water with few oil ($s_w^{left} = 0.99$) is injected in the under left corner ($x = 0, 0 \leq y \leq 0.2$ and $0 \leq x \leq 0.2, y = 0$). This pressure displaces the oil which flows freely outside the medium where the extraction zone is located in the upper right corner ($x = 1$, $0.8 \leq y \leq 1$ and $0.8 \leq x \leq 1, y = 1$) and is at the atmospheric pressure (see Fig. 2). The rest of the boundary is impervious. We take $\delta t = 0.001\ s$, with 50000 iterations. To solve the nonlinear system (2)–(12) we use a Newton method with a preconditioning of the Jacobian matrix. We take $s = s_{nw}$ and $p = p_w$ as primary variables.

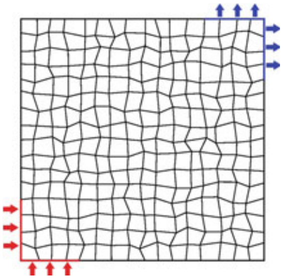
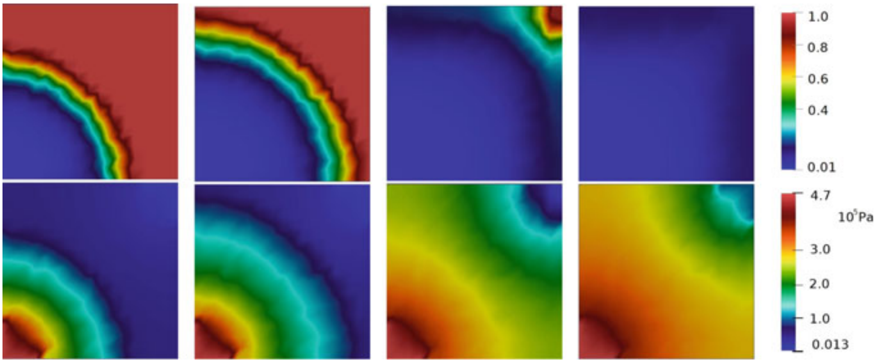


Fig. 2 Quadrangle mesh with boundary conditions



(a) $t = 10\ s$ (b) $t = 20\ s$ (c) $t = 40\ s$ (d) $t = 50\ s$ (e) Scale

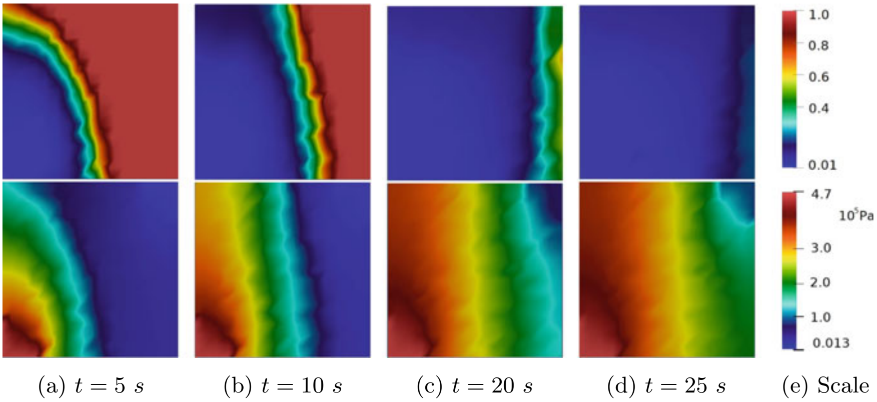
Fig. 3 Isotropic flow, upside s_{nw} and downside p_w

Case 1: isotropic **Case 2: homogeneous anisotropic**

$$\Lambda = 10^{-10} \times \begin{bmatrix} 0.15 & 0 \\ 0 & 0.15 \end{bmatrix} [m^2].\ \Lambda = 0.15 \times 10^{-10} \times R_{\theta_0} \times \begin{bmatrix} 1 & 0 \\ 0 & 5 \end{bmatrix} \times R_{\theta_0}^{-1}\ [m^2].$$

R_{θ_0} denotes the rotation matrix where $\theta_0 = \pi/6$.

In the first case, water flows in the medium regardless of the direction. Figure 3 illustrates the behavior of the oil saturation and water pressure with respect to mesh deformation at $t = 10,\ 20$, 40 and 50 s. Since there is no direction preferred, the oil is diagonally pushed to the extraction corner. At the final time, most of the oil is extracted. We indicate that the approximated saturation verifies its physical bounds between 0 and 1. In the second case, the permeability tensor is globally anisotropic. The favorised direction is the y-axis rotated with an angle of $\pi/6$. Figure 4 depicts the oil saturation and water pressure subject to this anisotropy and distortion of the mesh at the times $t = 5,\ 10$, 20 and 25 s. Compared to the first test, the y-eigenvalue is bigger in this test so the water pushes the oil through the vertical direction faster than through the horizontal direction. We see at $t = 25\ s$ most of the oil is extracted,



(a) $t = 5\ s$ (b) $t = 10\ s$ (c) $t = 20\ s$ (d) $t = 25\ s$ (e) Scale

Fig. 4 Uniform anisotropic flow with a rotation, upside s_{nw} and downside p_w

compared to the first test where there is still oil at $t = 40\ s$. We highlight that we also have the saturation in [0, 1] which confirms the theoretical result proved in Lemma 1.

References

1. Brenner, K., Masson, R.: Inter. J. Finite Volumes. **10**, 1–37 (2013)
2. Chainais-Hillairet, C., Krell, S., Mouton, A.: Num. Methods Partial. Differ. Equ., **31**(3), 723–760, De Gruyter (2015)
3. Chavent, G., Jaffre, J. Dtuf. Math. Appl. vol. 17. North-Holland, Amsterdam (1986)
4. Coudiere, Y., Manzini, G.: SIAM J. Numer. Anal. **47**(6), 4163–4192 (2010)
5. Crozon, T., Quenjel, E.H., Saad, M.: Submitted (2023)
6. Domelevo, K., Omnes, P.: ESAIM: Math. Model. Numer. Anal. **39**(6), 1203–1249 (2005)
7. Eymard, R., Herbin, R., Michel, A.: ESAIM: Math. Model. Numer. Anal. **37**(6), 937–972 (2003)
8. Ghilani, M., Quenjel, E.H. Saad, M.J.: Comput. Phys. **407** (2020)
9. Krell, S.: Numer. Methods Partial Differ. Equ. **27**, 1666–1706 (2011)
10. Quenjel, E.H.: Appl. Numer. Math. **161**, 148–168 (2021)
11. Saad, B., Saad, M.: SIAM J. Num. Anal. **51**(1), 716–751 (2013)

Author Index

E. Franck et al. (eds.), *Finite Volumes for Complex Applications X—Volume 1, Elliptic and Parabolic Problems*, Springer Proceedings in Mathematics & Statistics 432,
https://doi.org/10.1007/978-3-031-40864-9