Springer Proceedings in Mathematics & Statistics

Jürgen Fuhrmann
Mario Ohlberger
Christian Rohde *Editors*

Finite Volumes for Complex Applications VII - Methods and Theoretical Aspects

FVCA 7, Berlin, June 2014

Springer

Springer Proceedings in Mathematics & Statistics

Volume 77

For further volumes:
http://www.springer.com/series/10533

Springer Proceedings in Mathematics & Statistics

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

Jürgen Fuhrmann · Mario Ohlberger
Christian Rohde
Editors

Finite Volumes for Complex Applications VII - Methods and Theoretical Aspects

FVCA 7, Berlin, June 2014

Springer

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

Mario Ohlberger
Institute for Computational and Applied Mathematics and Center for Nonlinear Sciences (CeNoS)
University of Münster
Münster
Germany

Christian Rohde
Institute of Applied Analysis and Numerical Simulation
University of Stuttgart
Stuttgart
Germany

ISSN 2194-1009 ISSN 2194-1017 (electronic)
ISBN 978-3-319-05683-8 ISBN 978-3-319-05684-5 (eBook)
DOI 10.1007/978-3-319-05684-5
Springer Cham Heidelberg New York Dordrecht London

Library of Congress Control Number: 2014938474

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

Printed on acid-free paper

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

Preface

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

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

The triennial series of conferences "International Symposium on Finite Volumes for Complex Applications—Problems and Perspectives (FVCA)" brings together mathematicians, physicists, and engineers interested in this kind of physically motivated discretizations. Contributions to the further advancement of the theoretical understanding of suitable finite volume, finite element, discontinuous Galerkin and other discretization schemes, and the exploration of new application fields have been welcomed.

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

The present volumes contain the invited and contributed papers presented as posters or talks at the Seventh International Symposium on Finite Volumes for Complex Applications held in Berlin on June 15–20, 2014.

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

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

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

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

Without the financial contributions of the Deutsche Forschungsgemeinschaft (DFG), the Weierstrass Institute for Applied Analysis and Stochastics, the DFG Priority Program 1276 "Metström," the Westfälische Universität Münster, the Stuttgart Research Centre for Simulation Technology (Simtech), and the Czech Technical University of Prague, the organization of the conference and the production of the proceedings would not have been possible.

The Berlin Brandenburgische Akademie der Wissenschaften provided an impressive conference venue in the center of Berlin.

Finally, we have to thank the local organizers and the staff at the Weierstrass Institute for Applied Analysis and Stochastics for carrying the main organizational burden and for providing a friendly atmosphere for the conference.

March 2014

Jürgen Fuhrmann
Mario Ohlberger
Christian Rohde

Organization Committees

Organizing Committee

Peter Bastian
Robert Eymard
Jürgen Fuhrmann
Jiří Fürst
Annegret Glitzky
Volker John
Rupert Klein
Alexander Linke
Mario Ohlberger
Christian Rohde
Jörn Sesterhenn

Proceedings Committee

Remi Abgrall
Brahim Amaziane
Boris Andreianov
Peter Bastian
Fayssal Benkhaldoun
Franck Boyer
Yves Coudière
Andreas Dedner
Vit Dolejsi
Jerome Droniou
Denis Dutykh
Alexandre Ern
Robert Eymard
Jürgen Fuhrmann
Jiří Fürst

Jan Giesselmann
Annegret Glitzky
Khaled Hassouni
Christiane Helzel
Jean-Marc Hérard
Danielle Hilhorst
Florence Hubert
Volker John
Rupert Klein
Robert Kloefkorn
Peter Knabner
Alexander Linke
Konstantin Lipnikov
Andreas Meister
Mario Ohlberger
Christian Rohde
Martin Rumpf
Jörn Sesterhenn
Martin Vohralik
Petra Wittbold

Contents

Part I Invited Papers

Low Mach Number Modeling of Stratified Flows 3
Ann Almgren, John Bell, Andrew Nonaka and Michael Zingale

Entropy Method and Asymptotic Behaviours of Finite Volume Schemes 17
Claire Chainais-Hillairet

Interpolated Pressure Laws in Two-Fluid Simulations and Hyperbolicity 37
Philippe Helluy and Jonathan Jung

Part II Theoretical Aspects

An ALE Formulation for Explicit Runge-Kutta Residual Distribution 57
Remi Abgrall, Luca Arpaia and Mario Ricchiuto

Gradient Schemes for an Obstacle Problem 67
Yahya Alnashri and Jerome Droniou

The Complete Flux Scheme in Cylindrical Coordinates 77
M. J. H. Anthonissen and J. H. M. ten Thije Boonkkamp

A Staggered Scheme with Non-conforming Refinement for the Navier-Stokes Equations 87
Fabrice Babik, Jean-Claude Latché, Bruno Piar and Khaled Saleh

Consistency Analysis of a 1D Finite Volume Scheme for Barotropic Euler Models 97
Florent Berthelin, Thierry Goudon and Sebastian Minjeaud

An Asymptotic-Preserving Scheme for Systems of Conservation Laws with Source Terms on 2D Unstructured Meshes 107
C. Berthon, G. Moebs and R. Turpault

Numerical Dissipation and Dispersion of the Homogeneous and Complete Flux Schemes 117
J. H. M. ten Thije Boonkkamp and M. J. H. Anthonissen

A New Finite Volume Scheme for a Linear Schrödinger Evolution Equation 127
Abdallah Bradji

A Note on a New Second Order Approximation Based on a Low–Order Finite Volume Scheme for the Wave Equation in One Space Dimension 137
Abdallah Bradji

Note on the Convergence of a Finite Volume Scheme Using a General Nonconforming Mesh for an Oblique Derivative Boundary Value Problem 149
Abdallah Bradji

Optimal and Pressure-Independent L^2 Velocity Error Estimates for a Modified Crouzeix-Raviart Element with BDM Reconstructions 159
Christian Brennecke, Alexander Linke, Christian Merdon and Joachim Schöberl

Conservative Finite Differences as an Alternative to Finite Volume for Compressible Flows 169
Jens Brouwer, Julius Reiss and Jörn Sesterhenn

FV Upwind Stabilization of FE Discretizations for Advection–Diffusion Problems 177
Fabian Brunner, Florian Frank and Peter Knabner

Entropy-Diminishing CVFE Scheme for Solving Anisotropic Degenerate Diffusion Equations 187
Clément Cancès and Cindy Guichard

A Finite Volume Scheme with the Discrete Maximum Principle for Diffusion Equations on Polyhedral Meshes 197
Alexey Chernyshenko and Yuri Vassilevski

Continuous Finite-Elements on Non-Conforming Grids Using Discontinuous Galerkin Stabilization 207
Andreas Dedner, Robert Klöfkorn and Mirko Kränkel

A Well-Balanced Scheme for the Euler Equation with a Gravitational Potential 217
Vivien Desveaux, Markus Zenk, Christophe Berthon and Christian Klingenberg

An Explicit Staggered Finite Volume Scheme for the Shallow Water Equations 227
D. Doyen and P. H. Gunawan

A Uniformly Converging Scheme for Fractal Conservation Laws 237
Jérôme Droniou and Espen R. Jakobsen

Uniform-in-Time Convergence of Numerical Schemes for Richards' and Stefan's Models 247
Jérôme Droniou, Robert Eymard and Cindy Guichard

Comparison of Two Couplings of the Finite Volume Method and the Boundary Element Method 255
Christoph Erath

Gradient Schemes for Stokes Problem 265
Robert Eymard and Pierre Feron

Uniform Estimate of the Relative Free Energy by the Dissipation Rate for Finite Volume Discretized Reaction-Diffusion Systems 275
André Fiebach and Annegret Glitzky

Modified Finite Volume Nodal Scheme for Euler Equations with Gravity and Friction 285
Emmanuel Franck

A Linearity-Preserving Cell-Centered Scheme for the Anisotropic Diffusion Equations 293
Zhi-Ming Gao and Ji-Ming Wu

Convergence of Finite Volume Scheme for Degenerate Parabolic Problem with Zero Flux Boundary Condition 303
Boris Andreianov and Mohamed Karimou Gazibo

On Aposteriori Error Analysis of DG Schemes Approximating Hyperbolic Conservation Laws 313
Jan Giesselmann and Tristan Pryer

Estimating the Geometric Error of Finite Volume Schemes for Conservation Laws on Surfaces for Generic Numerical Flux Functions 323
Jan Giesselmann and Thomas Müller

Semi-implicit Alternating Discrete Duality Finite Volume Scheme for Curvature Driven Level Set Equation 333
Angela Handlovičová and Peter Frolkovič

Convergence of the MAC Scheme for the Steady-State Incompressible Navier-Stokes Equations on Non-uniform Grids 343
R. Herbin, J.-C. Latché and K. Mallem

Stochastic Modeling for Heterogeneous Two-Phase Flow 353
M. Köppel, I. Kröker and Christian Rohde

A New Discretization Method for the Convective Terms in the Incompressible Navier-Stokes Equations 363
N. Kumar, J. H. M. ten Thije Boonkkamp and B. Koren

Mimetic Finite Difference Schemes with Conditional Maximum Principle for Diffusion Problems 373
Konstantin Lipnikov

Discrete Relative Entropy for the Compressible Stokes System 383
Thierry Gallouët, David Maltese and Antonín Novotný

A Mixed Explicit Implicit Time Stepping Scheme for Cartesian Embedded Boundary Meshes 393
Sandra May and Marsha Berger

Finite-Volume Analysis for the Cahn-Hilliard Equation with Dynamic Boundary Conditions 401
Flore Nabet

Weak Convergence of Nonlinear Finite Volume Schemes for Linear Hyperbolic Systems 411
Michaël Ndjinga

A-Posteriori Error Estimates for the Localized Reduced Basis Multi-Scale Method 421
Mario Ohlberger and Felix Schindler

Positivity Preserving Implicit and Partially Implicit Time Integration Methods in the Context of the DG Scheme Applied to Shallow Water Flows 431
Sigrun Ortleb

Convergence of a Nonlinear Scheme for Anisotropic Diffusion Equations 439
Christophe Le Potier

A Hydrodynamic Model for Dispersive Waves Generated by Bottom Motion 449
S. R. Pudjaprasetya and S. S. Tjandra

A Conservative Coupling of Level-Set, Volume-of-Fluid and Other Conserved Quantities 457
Matthias Waidmann, Stephan Gerber, Michael Oevermann and Rupert Klein

Author Index 467

Part III Elliptic and Parabolic Problems

Asymptotic-Preserving Methods for an Anisotropic Model of Electrical Potential in a Tokamak 471
Philippe Angot, Thomas Auphan and Olivier Guès

Semi-implicit Second Order Accurate Finite Volume Method for Advection-Diffusion Level Set Equation 479
Martin Balažovjech, Peter Frolkovič, Richard Frolkovič and Karol Mikula

Adaptive Time Discretization and Linearization Based on a Posteriori Estimates for the Richards Equation 489
Vincent Baron, Yves Coudière and Pierre Sochala

Monotone Combined Finite Volume-Finite Element Scheme for a Bone Healing Model 497
Marianne Bessemoulin-Chatard and Mazen Saad

Vertex Approximate Gradient Scheme for Hybrid Dimensional Two-Phase Darcy Flows in Fractured Porous Media 507
Konstantin Brenner, Mayya Groza, Cindy Guichard and Roland Masson

Coupling of a Two Phase Gas Liquid Compositional 3D Darcy Flow with a 1D Compositional Free Gas Flow 517
Konstantin Brenner, Roland Masson, Laurent Trenty and Yumeng Zhang

Gradient Discretization of Hybrid Dimensional Darcy Flows in Fractured Porous Media 527
Konstantin Brenner, Mayya Groza, Cindy Guichard, Gilles Lebeau and Roland Masson

A Gradient Scheme for the Discretization of Richards Equation 537
Konstantin Brenner, Danielle Hilhorst and Huy Cuong Vu Do

Convergence of a Finite Volume Scheme for a Corrosion Model 547
Claire Chainais-Hillairet, Pierre-Louis Colin and Ingrid Lacroix-Violet

High Performance Computing Linear Algorithms for Two-Phase Flow in Porous Media 557
Robert Eymard, Cindy Guichard and Roland Masson

Numerical Solution of Fluid-Structure Interaction by the Space-Time Discontinuous Galerkin Method 567
Miloslav Feistauer, Martin Hadrava, Jaromír Horáček and Adam Kosík

An Anisotropic Diffusion Finite Volume Algorithm Using a Small Stencil 577
Martin Ferrand, Jacques Fontaine and Ophélie Angelini

Coupling of Fluid Flow and Solute Transport Using a Divergence-Free Reconstruction of the Crouzeix-Raviart Element 587
Jürgen Fuhrmann, Alexander Linke and Christian Merdon

Activity Based Finite Volume Methods for Generalised Nernst-Planck-Poisson Systems 597
Jürgen Fuhrmann

Suitable Formulations of Lagrange Remap Finite Volume Schemes for Manycore/GPU Architectures 607
Thibault Gasc and Florian De Vuyst

Efficient Parallel Simulation of Atherosclerotic Plaque Formation Using Higher Order Discontinuous Galerkin Schemes 617
Stefan Girke, Robert Klöfkorn and Mario Ohlberger

A DDFV Scheme for Incompressible Navier-Stokes Equations with Variable Density 627
Thierry Goudon and Stella Krell

An Efficient Implementation of a 3D CeVeFE DDFV Scheme on Cartesian Grids and an Application in Image Processing 637
Niklas Hartung and Florence Hubert

MPFA Algorithm for Solving Stokes-Brinkman Equations on Quadrilateral Grids 647
Oleg Iliev, Ralf Kirsch, Zahra Lakdawala and Galina Printsypar

Nonlinear Monotone FV Schemes for Radionuclide Geomigration and Multiphase Flow Models 655
Ivan Kapyrin, Kirill Nikitin, Kirill Terekhov and Yuri Vassilevski

Numerical Modelling of Viscous and Viscoelastic Fluids Flow in the Channel with T-Junction 665
Radka Keslerová, Karel Kozel and David Trdlička

Gradient Evaluation on a Quadtree Based Finite Volume Grid 675
Zuzana Krivá, Angela Handlovičová and Karol Mikula

3D Lagrangian Segmentation with Simultaneous Mesh Adjustment . . . 685
Karol Mikula and Mariana Remešíková

A Model Reduction Framework for Efficient Simulation of Li-Ion Batteries 695
Mario Ohlberger, Stephan Rave, Sebastian Schmidt and Shiquan Zhang

Coupling Free Flow and Porous Medium Flow Systems Using Sharp Interface and Transition Region Concepts 703
Iryna Rybak

Convergence Analysis of a FV-FE Scheme for Partially Miscible Two-Phase Flow in Anisotropic Porous Media 713
Bilal Saad and Mazen Saad

Piecewise Linear Transformation in Diffusive Flux Discretizations . . . 723
D. Vidović, M. Dotlić, B. Pokorni, M. Pušić and M. Dimkić

Comparison of Two Approaches for Treatment of the Interface Conditions in FV Discretization of Pore Scale Models for Li-Ion Batteries 731
Shiquan Zhang, Oleg Iliev, Sebastian Schmidt and Jochen Zausch

Part IV Hyperbolic Problems

A Finite Volume Method for Large-Eddy Simulation of Shallow Water Equations 741
Rajaa Abdellaoui, Fayssal Benkhaldoun, Imad Elmahi and Mohammed Seaid

An Asymptotic Preserving Scheme for the Barotropic Baer-Nunziato Model 749
Rémi Abgrall and Sophie Dallet

Numerical Simulations of a Fluid-Particle Coupling 759
Nina Aguillon

A Simple Finite Volume Approach to Compute Flows in Variable Cross-Section Ducts 769
Bruno Audebert, Jean-Marc Hérard, Xavier Martin and Ouardia Touazi

A 1D Stabilized Finite Element Model for Non-hydrostatic Wave Breaking and Run-up 779
P. Bacigaluppi, M. Ricchiuto and P. Bonneton

A Quasi-1D Model of Biomass Co-Firing in a Circulating Fluidized Bed Boiler 791
Michal Beneš, Pavel Strachota, Radek Máca, Vladimír Havlena and Jan Mach

Simulation of Diluted Flow Regimes in Presence of Unsteady Boundaries 801
Florian Bernard, Angelo Iollo and Gabriella Puppo

On the Use of the HLL-Scheme for the Simulation of the Multi-Species Euler Equations 809
Phillip Berndt

A Conservative Well-Balanced Hybrid SPH Scheme for the Shallow-Water Model 817
Christophe Berthon, Matthieu de Leffe and Victor Michel-Dansac

Asymptotic-Preserving Scheme Based on a Finite Volume/Particle-In-Cell Coupling for Boltzmann-BGK-Like Equations in the Diffusion Scaling 827
Anaïs Crestetto, Nicolas Crouseilles and Mohammed Lemou

Some Applications of a Two-Fluid Model 837
Fabien Crouzet, Frédéric Daude, Pascal Galon, Jean-Marc Hérard, Olivier Hurisse and Yujie Liu

Numerical Simulation of Flow in a Meridional Plane of Multistage Turbine 847
Jiří Fürst, Jaroslav Fořt, Jan Halama, Jiří Holman, Jan Karel, Vladimír Prokop and David Trdlička

Application of a Two-Fluid Model to Simulate the Heating of Two-Phase Flows 857
Jean-Marc Hérard, Olivier Hurisse, Antoine Morente and Khaled Saleh

Modeling Phase Transition and Metastable Phases 865
François James and Hélène Mathis

Almost Parallel Flows in Porous Media 873
Alaa Armiti-Juber and Christian Rohde

Towards a Stochastic Closure Approach for Large Eddy Simulation 883
Th. von Larcher, R. Klein, I. Horenko, P. Metzner, M. Waidmann, D. Igdalov, A. D. Beck, G. Gassner and C. D. Munz

A Well Balanced Scheme for a Transport Equation with Varying Velocity Arising in Relativistic Transfer Equation 891
T. Leroy, C. Buet and B. Després

An Arbitrary Space-Time High-Order Finite Volume Scheme for Gas Dynamics Equations in Curvilinear Coordinates on Polar Meshes 901
Bertrand Meltz, Stéphane Jaouen and Frédéric Lagoutière

A Combined Finite Volume Discontinuous Galerkin Approach for the Sharp-Interface Tracking in Multi-Phase Flow 911
Stefan Fechter and Claus-Dieter Munz

Numerical Simulation of an Incompressible Two-Fluid Model 919
Michael Ndjinga, Thi-Phuong-Kieu Nguyen and Christophe Chalons

On Boundary Approximation for Simulation of Granular Flow 927
David Neusius, Sebastian Schmidt and Axel Klar

Comparison of Realizable Schemes for the Eulerian Simulation of Disperse Phase Flows 935
Macole Sabat, Adam Larat, Aymeric Vié and Marc Massot

Shock Capturing for Discontinuous Galerkin Methods using Finite Volume Subcells 945
Matthias Sonntag and Claus-Dieter Munz

A Simple Well-Balanced, Non-negative and Entropy-Satisfying Finite Volume Scheme for the Shallow-Water System 955
Emmanuel Audusse, Christophe Chalons and Philippe Ung

Well-Balanced Inundation Modeling for Shallow-Water Flows with Discontinuous Galerkin Schemes 965
Stefan Vater and Jörn Behrens

Comparison of Cell-Centered and Staggered Pressure-Correction Schemes for All-Mach Flows 975
Nicolas Therme and Chady Zaza

Author Index 985

Part I
Invited Papers

Low Mach Number Modeling of Stratified Flows

Ann Almgren, John Bell, Andrew Nonaka and Michael Zingale

Abstract Low Mach number equation sets approximate the equations of motion of a compressible fluid by filtering out the sound waves, which allows the system to evolve on the advective rather than the acoustic time scale. Depending on the degree of approximation, low Mach number models retain some subset of possible compressible effects. In this paper we give an overview of low Mach number methods for modeling stratified flows arising in astrophysics and atmospheric science as well as low Mach number reacting flows. We discuss how elements from the different fields are combined to form MAESTRO, a code for modeling low Mach number stratified flows with general equations of state, reactions and time-varying stratification.

1 Introduction

Physical phenomena encompassing a wide range of length and time scales occur in a large number of fluid dynamical areas. In the atmosphere, for example, we want to understand flows on the scale of local regions, continents, or the entire globe. In astrophysics we want to understand not only how the nuclear energy release from thin flame fronts may unbind a star, resulting in a dramatic supernova, but also how the

A. Almgren (✉) · J. Bell · A. Nonaka
Lawrence Berkeley National Laboratory, Berkeley, CA, USA
e-mail: ASAlmgren@lbl.gov

J. Bell
e-mail: JBBell@lbl.gov

A. Nonaka
e-mail: AJNonaka@lbl.gov

M. Zingale
Stony Brook University, Stony Brook, NY, USA
e-mail: Michael.Zingale@stonybrook.edu

J. Fuhrmann et al. (eds.), *Finite Volumes for Complex Applications VII - Methods and Theoretical Aspects*, Springer Proceedings in Mathematics & Statistics 77,
DOI: 10.1007/978-3-319-05684-5_1,

large-scale convective flows that precede the explosion lead to ignition. In laboratory-scale turbulent combustion, we need to understand not only the detailed chemistry of how the flame burns, but also how it influences and responds to the turbulent flow around it. In principle these scenarios can all be modeled using the equations of motion that govern compressible reacting fluids. Even with today's (and tomorrow's) supercomputers, however, in each of these areas there are phenomena for which simulating all the fluid dynamical scales of these problems over the time periods of interest is beyond our reach. For many of these problems, our understanding of the flow does not require tracking the acoustic waves that carry relatively little energy and travel much faster than the fluid itself.

We are interested in numerical methods that can accurately model the phenomena of interest but are not limited by needing to resolve the acoustic time scale. One approach is to advance the acoustic signal in time with an implicit time discretization or to treat the acoustic signal explicitly yet separately from the rest of the flow, for example as used in cloud modeling in [41] and [22], respectively. Another approach is to modify the governing equations so that acoustic waves are no longer supported by the equations. In this paper we explore the latter approach for low Mach number flows, i.e. flows for which we can exploit the separation of scales that occurs when the Mach number, M (the ratio of the fluid velocity to the sound speed), is much less than unity. Physically, one can think of the solution to a low Mach number model as supporting infinitely fast acoustic equilibration rather than finite-velocity acoustic wave propagation. Mathematically, this is manifest in the addition of a constraint on the velocity field to the system of otherwise hyperbolic evolution equations. This velocity constraint can be translated into an elliptic equation for pressure that expresses the equilibration process. Because the time step in explicit discretization schemes for a low Mach number system is limited by the fluid velocity and not by the sound speed, these methods often gain several orders of magnitude in computational efficiency over traditional compressible approaches.

Fundamental to the traditional low Mach number approach is that we can decompose the pressure as

$$p(\mathbf{x}, t) = p_0 + p'(\mathbf{x}, t)$$

where p_0 is the ambient thermodynamic (or "reference" or "background") pressure, and p' is the perturbational pressure, where $p'/p_0 \sim \mathcal{O}\left(M^2\right)$. For small-scale reacting flows in an open environment, p_0 reduces to a constant in space and time; in a closed combustion chamber, $p_0 = p_0(t)$. For reacting, stratified flow on a larger scale, p_0 is a function of both the radius (distance from the center of the star for stellar applications, or elevation in the atmospheric case) and time, when large-scale net heating/cooling is present.

2 Background

The simplest low Mach number model is expressed by the incompressible Navier-Stokes equations for a constant density fluid. This can be generalized to variable density incompressible flow, in which the density varies spatially across the domain but the density of an individual fluid element does not change over time. Low Mach number models for chemical combustion [10, 31, 40] and nuclear burning [6] incorporate large compressibility effects due to chemical/nuclear reactions and thermal processes with a spatially constant background pressure. The Boussinesq approximation [8] includes a heating-induced buoyancy term in the momentum equation but requires that the flow itself be incompressible. The anelastic equations (see, e.g., [4, 5, 14, 20, 27–29, 37, 44] for atmospheric flows, and [18, 19, 26] for astrophysical flows; see also the references in [9]), capture volumetric expansion due to motion relative to a stratified background in addition to buoyancy due to deviation of the local state from the background state, and have been widely used in modeling of atmospheric and astrophysical flows. The pseudo-incompressible (PI) approximation, introduced by [12, 13] and rigorously derived using low Mach number asymptotics by [7], generalizes anelastic models by allowing larger variations in density and temperature in response to localized heat release, but is restricted to an ideal gas equation of state. A low Mach number model for astrophysical flows [1–3, 35, 46], implemented in a code named MAESTRO, has generalized the pseudo-incompressible approximation to more general equations of state for use in astrophysical modeling, and has extended its applicability by allowing time variation of the background stratification to accommodate an expanding stellar atmosphere. Recently, a modification of the momentum equation to improve the accuracy of the buoyancy term in the low Mach number equation set was proposed, in a general form by [21], and then by [43] using an alternate derivation based on Lagrangian analysis.

3 Fully Compressible Equations for Stratified Flow

The fully compressible (FC) equations for a reacting multicomponent gas in the presence of gravity, neglecting Coriolis terms, viscosity, thermal and species diffusion, and weak nuclear interactions, can be written in conservation form as

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho U) = 0, \tag{1}$$

$$\frac{\partial (\rho h)}{\partial t} + \nabla \cdot (\rho U h) = \frac{Dp}{Dt} - \rho \sum_k q_k \dot{\omega_k}, \tag{2}$$

$$\frac{\partial (\rho U)}{\partial t} + \nabla \cdot (\rho U U) + \nabla p = -\rho g \mathbf{e}_r, \tag{3}$$

$$\frac{\partial (\rho X_k)}{\partial t} + \nabla \cdot (\rho U X_k) = \rho \dot{\omega_k}, \tag{4}$$

where ρ, U, p and h are the density, velocity, pressure and enthalpy (per unit mass), respectively, and g is the gravitational acceleration in the direction of $\mathbf{e}_r$, the unit vector in the radial direction. The species are represented by their mass fractions, X_k, along with their associated production rates, $\dot{\omega}_k$, and specific binding energies, q_k. The equation of state can be written in a general form as $p = \widehat{p}\,(\rho, T, X_k)$ or $\rho = \widehat{\rho}\,(p, T, X_k)$ where T is the temperature.

Still allowing the fluid to be fully compressible, we can define a hydrostatic base state pressure, $p_0(r, t)$, and corresponding base state density, $\rho_0(r, t)$ such that $\nabla p_0 \equiv -\rho_0 g \mathbf{e}_r$. We define the deviation from the reference value, $p' = p - p_0$, but make no assumptions about the size of the deviation. Then, with no approximation, Eq. (3) can be written instead as

$$\frac{\partial(\rho U)}{\partial t} + \nabla \cdot (\rho U U) + \nabla p' = -\rho' g \mathbf{e}_r, \tag{5}$$

where $\rho' \equiv \rho - \rho_0$. Even though we have decomposed the variables into reference state values and perturbational values, no assumptions about the magnitude of the perturbations have been made at this point, and the perturbational form is algebraically equivalent to the non-perturbational form.

4 Low Mach Number Approach

The methodology developed for low Mach number modeling of astrophysical flows, and implemented in a code named MAESTRO [1–3, 35, 46], represents the synthesis of ideas from three separate fields into a new algorithmic approach. First, the anelastic and pseudo-incompressible approximations, first derived in the context of atmospheric science, suggest how to filter sound waves for environments in which the background stratification due to gravity plays a significant role in the dynamics. These approximations differ from those in incompressible and low Mach number combustion modeling in that the background pressure and density are in hydrostatic equilibrium rather than spatially constant. Second, methods developed for low Mach number combustion inform how to incorporate local expansion effects due to reactions and thermal diffusion in a low Mach number setting. Finally, formulation of the equation of state, reaction networks, and other thermodynamical characterizations of stellar material require detailed astrophysical expertise. Contributions from all three fields were used to devise a method that

- captures the same large-scale motions as the fully compressible equations
- allows for local expansion due to reactions, thermal diffusion, and compositional mixing
- allows for local expansion due to movement relative to background stratification
- allows the background state to evolve in time in response to large-scale heating and/or mixing

- does not require an ideal gas equation of state
- removes acoustic wave propagation.

Low Mach number equations in all settings are typically derived by first expanding all of the variables asymptotically in the Mach number, M, which is assumed to be small. Using standard asymptotic techniques, one can show that in the momentum equations we retain the zeroth order terms of velocity and density, but the zeroth order term of the pressure gradient must be either zero or, in the case of stratified flows, balanced by the hydrostatic gravitational forcing. The dynamic component of the pressure gradient must be $O(M^2)$; this is typically translated into writing the pressure, p, as $p = p_0 + p'$, where p_0 is the background pressure, p' is the dynamic pressure, and $p'/p_0 \sim O(M^2)$. Standard techniques would also dictate that the density variation from ambient must be small at all times or the buoyancy term would lead to too strong an acceleration. A slightly more general approach, outlined in [2], replaces the restriction on the magnitude of the buoyancy term itself by a restriction on the effect of the buoyancy term, namely the magnitude of the velocity itself.

The fundamental approximation made in the the low Mach number equations is that the compressible pressure can be approximated by the background pressure in the equation of state. This differs from earlier derivations of similar equations (e.g, the anelastic approximation) which required that density and temperature variations from the ambient be small in order to ensure the pressure perturbation be small; here we require only that the pressure perturbation itself be small. In the low Mach number system, then, we write the equation of state in a general form as $p_0 = \widehat{p}\,(\rho, T, X_k)$ or $\rho = \widehat{\rho}\,(p_0, T, X_k)$.

To model a full star, for example, we would start by defining a radial background state in hydrostatic equilibrium. In practice, this profile would come from a one-dimensional stellar evolution code, which provides us with a model in hydrostatic balance. Thus our base state satisfies

$$\frac{\partial p_0(r,t)}{\partial r} = -\rho_0(r,t)g(r,t), \tag{6}$$

where $g(r,t)$ can be computed from $\rho_0(r,t)$ as

$$g(r,t) = \frac{GM_{\text{encl}}(r,t)}{r^2} \tag{7}$$

with the mass enclosed within a radius r defined as

$$M_{\text{encl}}(r,t) = 4\pi \int_0^r \rho_0(r',t)r'^2 dr'. \tag{8}$$

Here G is the gravitational constant. For modeling the earth's atmosphere, one can remove the time dependence of ρ_0 and p_0, and consider g to be spatially and temporally constant.

Given the base state, standard asymptotic analysis yields the following system of equations:

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (U\rho) = 0, \tag{9}$$

$$\frac{\partial (\rho h)}{\partial t} + \nabla \cdot (U\rho h) = \frac{Dp_0}{Dt} - \rho \sum_k q_k \dot{\omega_k}, \tag{10}$$

$$\frac{\partial U}{\partial t} + U \cdot \nabla U = -\frac{1}{\rho} \nabla p' - \frac{\rho'}{\rho} g \mathbf{e}_r, \tag{11}$$

$$\frac{\partial (\rho X_k)}{\partial t} + \nabla \cdot (\rho U X_k) = \rho \dot{\omega_k}. \tag{12}$$

The first and second equations are unchanged from the fully compressible versions with the exception that p is replaced by p_0 in the enthalpy evolution equation. The only source terms in the velocity evolution equation (21) are due to the dynamic pressure and the buoyancy. However, if one asymptotically examines not the difference between the solution and the base state, but the difference between the solution to the low Mach number system and the solution to the compressible system, one finds a correction to the buoyancy term of the form,

$$\left(\frac{\rho_0}{\rho^2} \frac{\partial \rho}{\partial p_0} \bigg|_s p' \right) g \mathbf{e}_r, \tag{13}$$

where the derivative is taken at constant entropy, s, so that the velocity equation now has the form

$$\frac{\partial U}{\partial t} + U \cdot \nabla U = -\frac{1}{\rho} \nabla p' - \frac{1}{\rho} \left(\rho' + \frac{\rho_0}{\rho} \frac{\partial \rho}{\partial p_0} \bigg|_s p' \right) g \mathbf{e}_r. \tag{14}$$

This additional term, which can be viewed as modifying the density appearing in the buoyancy term to have a correction due to the perturbational pressure, was introduced first in [21] and then in [43]; both demonstrated that the inclusion of this term enables the system to conserve a low Mach number form of total energy. This modified system yields a solution that is closer to the solution to the fully compressible system of equations than the original low Mach number system.

The more fundamental change in the structure of the system of equations results from replacing the pressure in the equation of state by the background pressure. Differentiating the equation of state along particle paths then converts the algebraic equation of state into a constraint on the divergence of the velocity. Differentiating $p_0 = \widehat{p}(\rho, T, X_k)$ along particle paths and rearranging terms yields

$$-\nabla \cdot U = \frac{1}{\rho} \frac{D\rho}{Dt} = \frac{1}{\rho p_\rho} \left(\frac{Dp_0}{Dt} - p_T \frac{DT}{Dt} - \sum_k p_{X_k} \dot{\omega_k} \right), \tag{15}$$

with $p_\rho = \partial p/\partial\rho|_{X_k,T}$, $p_{X_k} = \partial p/\partial X_k|_{T,\rho,(X_j,j\neq k)}$, and $p_T = \partial p/\partial T|_{\rho,X_k}$. Expanding and simplifying this expression as in [1] results in

$$\nabla\cdot U + \frac{1}{\Gamma_1 p_0}\left(\frac{\partial p_0}{\partial t} + U\cdot\nabla p_0\right) = -\sigma\sum_k \xi_k\dot{\omega}_k + \frac{1}{\rho p_\rho}\sum_k p_{X_k}\dot{\omega}_k - \sigma\sum_k q_k\dot{\omega}_k \equiv S, \tag{16}$$

where $\xi_k \equiv \partial h/\partial X_k|_{T,p,(X_j,j\neq k)}$, $\Gamma_1 \equiv d(\log p)/d(\log\rho)|_s$. and

$$\sigma = \frac{p_T}{\rho c_p p_\rho}, \tag{17}$$

which, for a gamma law gas, reduces to $\sigma = 1/(c_p T)$. We see that the first two terms in S capture the effect of compositional changes, while the third represents heat release from the reactions. If we now allow Γ_1 to be replaced by its lateral average, $\overline{\Gamma}_1(r,t)$, then, as shown in [2], $\nabla\cdot U + (1/(\overline{\Gamma}_1 p_0))U\cdot\nabla p_0$ can be rewritten as $(1/\beta_0)\nabla\cdot(\beta_0 U)$ where

$$\beta_0(r,t) = \beta(0,t)\exp\left(\int_0^r \frac{1}{(\overline{\Gamma}_1 p_0)}\frac{\partial p_0}{\partial r'}\,dr'\right). \tag{18}$$

Thus we can write the constraint as

$$\nabla\cdot(\beta_0 U) = \beta_0\left(S - \frac{1}{\overline{\Gamma}_1 p_0}\frac{\partial p_0}{\partial t}\right), \tag{19}$$

which allows us to use a variable density projection method analogous to that used to solve the incompressible Navier-Stokes equations. This constraint on the velocity field controls the degree to which the fluid can expand, forcing the evolution of the thermodynamic quantities to be consistent with the equation of state. It is the presence of $\beta_0 \neq 1$ that allows the fluid to expand as it rises; the magnitude of S determines the degree to which the fluid expands due to heat release and compositional changes.

The resulting Poisson equation for p', in the absence of the additional term, (13), in the velocity evolution equation, can be written

$$\nabla\cdot\left(\frac{\beta_0}{\rho}\nabla p'\right) = RHS, \tag{20}$$

where RHS includes the divergence of the advective terms as well as S and the time derivative of p_0. With the substitution of $\overline{\Gamma}_1$ for Γ_1, it was shown in [43] that the velocity evolution equation with the additional term, (13), can be written as

$$\frac{\partial U}{\partial t} + U\cdot\nabla U = -\frac{\beta_0}{\rho}\nabla\left(\frac{p'}{\beta_0}\right) - \rho' g\mathbf{e}_r, \tag{21}$$

instead of (14). This results in a Poisson equation for the perturbational pressure in the form

$$\nabla \cdot \left(\frac{\beta_0^2}{\rho} \nabla \frac{p'}{\beta_0} \right) = RHS. \tag{22}$$

For simulations of full stars where the background stratification varies with both radius and time, we must evolve the base state pressure and density in time in response to large scale heating and convection while retaining hydrostatic equilibrium. The velocity field used to update the density includes a local component, $\tilde{U}$, which accounts for localized convective and compressibility effects, and a base state component, w_0, which accounts for large-scale equilibration of the atmosphere. We can calculate w_0 by deriving a one-dimensional expression for the divergence of w_0, containing terms representing the average of S, and integrating that expression in the radial direction. Once we have advanced the density field, we can compute the new base state density as the average stratification, and compute the new base state pressure using hydrostatic equilibrium. Details are given in [35].

5 Numerical Approach

Low Mach number formulations replace the compressible flow equations with a constrained system of partial differential equations similar in structure to the incompressible Navier-Stokes equations. A number of projection-type methods have been developed to simulate incompressible and other low Mach number flows using a time step based on the fluid velocity and not the sound speed. Projection methods are fractional step schemes in which the solution is first advanced using a lagged approximation to the constraint, then, in a second step, a projection is applied to enforce the constraint. To solve the low Mach number equations for astrophysics in the MAESTRO code, we use an explicit second-order upwind discretization for advection, and Strang splitting to incorporate the contributions of reactions to the species and enthalpy. The projection step solves a second-order, self-adjoint, variable-coefficient elliptic equation for an update to the perturbational pressure, which is then used to correct the velocity. To include the base state evolution, in the predictor step we use an estimate of the expansion term, S, to compute a preliminary solution at the new time level, and in the corrector step we use the results from the predictor step to compute a more accurate expansion term, and compute the final solution at the new time level. The resulting algorithm advances the fluid evolution equations using a time step constrained by the fluid velocity rather than the acoustic wave speed, resulting in a significant increase in efficiency over traditional fully compressible methods.

We have implemented the entire low Mach number algorithm in MAESTRO in an adaptive mesh refinement (AMR) framework. Our approach to AMR uses a nested

hierarchy of logically rectangular grids with successively finer grids at higher levels. The key difference between our method and most block-structured AMR methods stems from the presence of a one-dimensional base state whose time evolution is coupled to that of the full solution. The algorithm does not subcycle in time, i.e., the solution at all levels is advanced with the same time step. Complete details are available in [35].

6 Case Study: Type Ia Supernovae

Type Ia supernovae (SNe Ia) are important distance indicators in cosmology, responsible for the discovery of the acceleration of the expansion of the Universe. They are also important sites of nucleosynthesis, making half of the iron in our galaxy. Despite their great importance, there are major uncertainties in the theoretical understanding of SNe Ia, even as to what progenitor systems give rise to these explosions. One of the theoretically favored models is the explosion of a carbon/oxygen white dwarf (a compact star about the volume of Earth weighing roughly 1.4 solar masses, the Chandrasekhar limit) which accretes mass from a stellar companion. As its mass grows, the central temperature increases and carbon fusion reactions begin, driving convection throughout the white dwarf interior. This convection, during which heated parcels of fluid buoyantly rise away from the center and cool as they expand, can last for centuries [45]. Eventually the reactions proceed so vigorously that the hot parcels do not cool fast enough and a thermonuclear burning front (flame) is formed. This burning front propagates through the white dwarf in seconds, converting the majority of carbon and oxygen into heavy elements [including silicon, iron, and nickel (Ni)], releasing enough energy to unbind the star. The brightness of the event depends on how much radioactive ^{56}Ni is produced, and this in turn depends on how complete the burning is, the composition of the white dwarf, and at what densities it occurs (see e.g. [23, 42]). While there are great uncertainties in all phases of this picture, a critical uncertainty is the nature of the convection preceding ignition, and how that affects ignition of the first flame. Computationally it has been shown that variations in the location of the ignition lead to great differences in the explosion outcome [15, 30, 34, 39]. We conducted a computational study of this convective phase preceding explosion using MAESTRO.

In order to model convection in the (roughly) spherical self-gravitating white dwarf, the core algorithm in MAESTRO, which solves the low Mach number equations on a Cartesian grid, was modified to allow the base state to vary in the radial direction, which is not aligned with any of the coordinate axes [46]. In addition, we used AMR in order to focus spatial refinement on the regions of most intense heating where ignition was most likely to occur. This required a novel spatial mapping technique between the time-evolving one-dimensional radial base state and the hierarchical Cartesian mesh [35]. The procedure to expand the base state in response to large-scale heating is also considerably more complex for a self-gravitating fluid in a spherical geometry.

In a series of three papers [36, 46, 48] we presented results from a suite of simulations of the last few hours of convection preceding ignition. Once a burning front ignited we ended each simulation. However, by running many simulations and looking at the spatial and temporal distribution of "failed" hotspots (plumes that approached the ignition temperature but cooled before actually igniting) we built up statistics on the likelihood of ignition at a given radius from the center. Our major findings were:

- A strong jet-like feature dominates the outward-moving convective flow. This is similar to the dipole feature reported in previous studies [24], but more collimated and with a rapidly changing direction.
- Ignition most likely occurs at a single location, not multiple distinct locations all at once.
- Although the strongest heating occurs at the center of the star, ignition itself is most likely off-center, with a typical radius of 75 km from the center.
- The turbulent field in the convective layer follows Kolmogorov scaling with a smaller turbulent intensity and larger integral scale than assumed in previous works.
- The turbulent field is likely too weak to affect the initial flame propagation.

Figure 1 shows a snapshot of the convective region in the star. We see that a strong outflow feature (colored red) dominates the flow, and that the nuclear energy generation is strongly peaked toward the center of the star.

7 Future Work

MAESTRO in its current form can be used for a variety of astrophysical applications beyond that of modeling the Chandrasekhar-mass progenitor of SNe Ia. To date, MAESTRO has been used to study the sub-Chandra progenitor model for SNe Ia (in which burning begins in an accreted helium layer on the surface of a white dwarf) [47], core convection in massive stars [16, 17], and X-ray bursts [32]. MAESTRO simulation results have also provided the initial conditions for fully compressible simulations of the explosion phase of SNe Ia, as in [33]. Potential future applications include classical novae, proto-neutron star cooling, and convection in exoplanetary interiors.

Modeling warm, moist, non-precipitating flows in the earth's atmosphere with MAESTRO is easily achieved by assuming constant gravity, neglecting base state expansion, and including an appropriate equation of state for moist microphysics. A representation of phase change that is suitably accurate at the larger time steps of a low Mach number model must be used; see [11] for a discussion of the role of the time step in the accuracy of moist compressible models. An alternative, pseudo-incompressible, model for moist flows has been developed by [38].

Future developments of the MAESTRO code include the extension of the base state to include long-wavelength lateral variation. For future stellar modeling, we plan

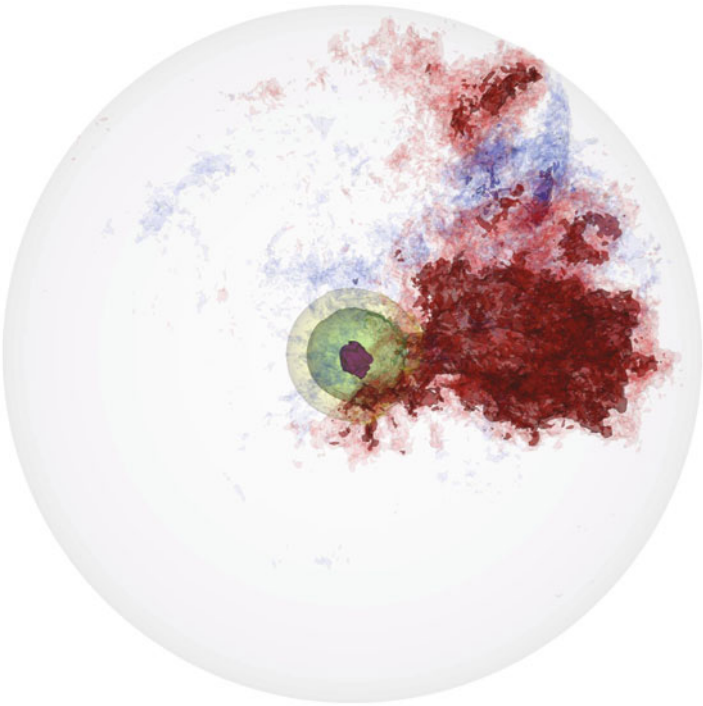


Fig. 1 Snapshot of the convection region in a white dwarf seconds before a supernova explosion. The *red* and *blue* contours show the radial velocity field (*red* is outflow, *blue* is inflow) and the *yellow* to *green* to *purple* contours show the nuclear energy generation rate. The radius of this region is ∼1000 km. Figure adopted from [36]

to include rotation of the star, which generates additional terms in the momentum equation as well as breaking the spherical symmetry of the base state. Finally, following recent studies in [25] of a generalized anelastic model compared to a standard anelastic model for moist flows, we plan to investigate further issues about the potential role of the pressure perturbation in the thermodynamics for both astrophysical and atmospheric applications.

Acknowledgments The work at LBNL was supported by the Applied Mathematics Program of the DOE Office of Advance Scientific Computing Research under U.S. Department of Energy under contract No. DE-AC02-05CH11231. The work at Stony Brook was supported by a DOE/Office of Nuclear Physics grant Nos. DE-FG02-06ER41448 and DE-FG02-87ER40317 to Stony Brook. An award of computer time was provided by the Innovative and Novel Computational Impact on Theory and Experiment (INCITE) program. This research used resources of the Oak Ridge Leadership Computing Facility located in the Oak Ridge National Laboratory, which is supported by the Office of Science of the Department of Energy under Contract DE-AC05-00OR22725. The MAESTRO code is freely available from http://bender.astro.sunysb.edu/Maestro/.

References

1. Almgren, A.S., Bell, J.B., Nonaka, A., Zingale, M.: Low mach number modeling of type ia supernovae. iii. reactions. Astrophys. J. **684**, 449–470 (2008). doi:10.1086/590321
2. Almgren, A.S., Bell, J.B., Rendleman, C.A., Zingale, M.: Low mach number modeling of type ia supernovae. i. hydrodynamics. Astrophys. J. **637**, 922–936 (2006)
3. Almgren, A.S., Bell, J.B., Rendleman, C.A., Zingale, M.: Low mach number modeling of type ia supernovae. ii. energy evolution. Astrophys. J. **649**, 927–938 (2006)
4. Bannon, P.: Nonlinear hydrostatic adjustment. J. Atmos. Sci. **53**(23), 3606–3617 (1996)
5. Batchelor, G.K.: The conditions for dynamical similarity of motions of a frictionless perfect-gas atmosphere. Quart. J. R. Meteor. Soc. **79**, 224–235 (1953)
6. Bell, J.B., Day, M.S., Rendleman, C.A., Woosley, S.E., Zingale, M.A.: Adaptive low mach number simulations of nuclear flame microphysics. J. Comp. Phys. **195**(2), 677–694 (2004)
7. Botta, N., Klein, R., Almgren, A.: Asymptotic analysis of a dry atmosphere. In: Neittaanmäki et al. (eds.) ENUMATH 99, Numerical Mathematics and Advanced Applications, p. 262. World Scientific, Singapore (1999)
8. Boussinesq, J.: Theorie Analytique de la Chaleur, vol. 2. Gauthier-Villars, Paris (1903)
9. Brown, B.J., Vasil, G.M., Zweibel, E.G.: Energy conservation and gravity waves in sound-proof treatments of stellar interiors: part i. anelastic approximations. Astrophys. J. **756**(109), 1–20 (2012)
10. Day, M.S., Bell, J.B.: Numerical simulation of laminar reacting flows with complex chemistry. Combust. Theory Model. **4**(4), 535–556 (2000)
11. Duarte, M., Almgren, A.S., Balakrishnan, K., Bell, J.B.: A Numerical Study of Methods for Moist Atmospheric Flows: Compressible Equations. Submitted for publication, arXiv:1311.4265 (2014)
12. Durran, D.R.: Improving the anelastic approximation. J. Atmos. Sci. **46**(11), 1453–1461 (1989)
13. Durran, D.R.: A physically motivated approach for filtering acoustic waves from the equations governing compressible stratified flow. J. Atmos. Sci. **601**, 365–379 (2008)
14. Dutton, J.A., Fichtl, G.H.: Approximate equations of motion for gases and liquids. J. Atmos. Sci. **26**, 241–254 (1969)
15. García-Senz, D., Bravo, E.: Type ia supernova models arising from different distributions of igniting points. Astron. Astrophys. **430**, 585–602 (2005). doi:10.1051/0004-6361:20041628
16. Gilet, C., Almgren, A.S., Bell, J.B., Nonaka, A., Woosley, S., Zingale, M.: Low mach number modeling of core convection in massive stars. APJ **773**, 137 (2013)
17. Gilet, C.E.: Low Mach Number simulation of core convection in massive stars. Ph.D. thesis, University of California, Berkeley (2012)
18. Gilman, P.A., Glatzmaier, G.A.: Compressible convection in a rotating spherical shell. i. anelastic equations. Astrophys. J. Supp. **45**, 335–349 (1981)
19. Glatzmaier, G.A.: Numerical simulation of stellar convective dynamos i. The model and method. J. Comp. Phys. **55**, 461–484 (1984)
20. Gough, D.O.: The anelastic approximation for thermal convection. J. Atmos. Sci. **26**, 448–456 (1969)
21. Klein, R., Pauluis, O.: Thermodynamic consistency of a pseudoincompressible approximation for general equations of state. J. Atmos. Sci. **69**:961–968 (2012)
22. Klemp, J.B., Wilhelmson, R.B.: The simulation of three-dimensional convective storm dynamics. J. Atmos. Sci. **35**, 1070–1096 (1978)
23. Krueger, B.K., Jackson, A.P., Calder, A.C., Townsley, D.M., Brown, E.F., Timmes, F.X.: Evaluating systematic dependencies of type ia supernovae: the influence of central density. Astrophys. J. **757**, 175 (2012). doi:10.1088/0004-637X/757/2/175
24. Kuhlen, M., Woosley, S.E., Glatzmaier, G.A.: Carbon ignition in type ia supernovae. ii. a three-dimensional numerical model. Astrophys. J. **640**, 407–416 (2006). doi:10.1086/500105
25. Kurowski, M., Grabowski, W., Smolarkiewicz, P.: Towards multiscale simulation of moist flows with soundproof equations. J. Atmos. Sci. **70**, 3995–4011 (2013)

26. Latour, J., Spiegel, E.A., Toomre, J., Zahn, J.P.: Stellar convection theory. i. the anelastic modal equations. Astrophys. J. **207**, 233–243 (1976)
27. Lipps, F.: On the anelastic approximation for deep convection. J. Atmos. Sci. **47**, 1794–1798 (1990)
28. Lipps, F., Hemler, R.: A scale analysis of deep moist convection and some related numerical calculations. J. Atmos. Sci. **39**, 2192–2210 (1982)
29. Lipps, F., Hemler, R.: Another look at the scale analysis for deep moist convection. J. Atmos. Sci. **42**, 1960–1964 (1985)
30. Livne, E., Asida, S.M., Höflich, P.: On the sensitivity of deflagrations in a chandrasekhar mass white dwarf to initial conditions. Astrophys. J. **632**, 443–449 (2005). doi:10.1086/432975
31. Majda, A., Sethian, J.A.: Derivation and numerical solution of the equations of low mach number combustion. Comb. Sci. Tech. **42**, 185–205 (1985)
32. Malone, C., Nonaka, A., Almgren, A., Bell, J., Zingale, M.: Multidimensional modeling of type i x-ray bursts. i. two-dimensional convection prior to the outburst of a pure he accretor. APJ **728**, 118 (2011)
33. Malone, C., Nonaka, A., Woosley, S., Almgren, A.S., Bell, J.B., Dong, S., Zingale, M.: The deflagration stage of chandrasekhar mass models for type ia supernovae: i. early evolution. APJ **782**(1), 11 (2014)
34. Niemeyer, J.C., Hillebrandt, W., Woosley, S.E.: Off-center deflagrations in chandrasekhar mass type IA supernova models. Astrophys. J. **471**, 903−+ (1996). doi: 10.1086/178017
35. Nonaka, A., Almgren, A.S., Bell, J.B., Lijewski, M.J., Malone, C.M., Zingale, M.: Maestro:an adaptive low mach number hydrodynamics algorithm for stellar flows. Astrophys. J. Supp. **188**, 358–383 (2010)
36. Nonaka, A., Aspden, A.J., Zingale, M., Almgren, A.S., Bell, J.B., Woosley, S.E.: High-resolution simulations of convection preceding ignition in type ia supernovae using adaptive mesh refinement. Astrophys. J. **745**, 73 (2012). doi:10.1088/0004-637X/745/1/73
37. Ogura, Y., Phillips, N.A.: Scale analysis of deep and shallow convection in the atmosphere. J. Atmos. Sci. **19**, 173–179 (1962)
38. O'Neill, W., Klein, R.: A moist pseudo-incompressible model. Atmos. Res. (2013)
39. Plewa, T., Calder, A.C., Lamb, D.Q.: Type ia supernova explosion: gravitationally confined detonation. Astrophys. J. **612**, L37–L40 (2004)
40. Rehm, R.G., Baum, H.R.: The equations of motion for thermally driven buoyant flows. J. Res. Natl. Bur. Stan. **83**, 297–308 (1978)
41. Tapp, M., White, P.: A non-hydrostatic mesoscale model. Q. J. Roy. Meteor. Soc. **102**(432), 277–296 (1976)
42. Timmes, F.X., Brown, E.F., Truran, J.W.: On variations in the peak luminosity of type Ia supernovae. Astrophys. J. **590**, L83–L86 (2003). doi:10.1086/376721
43. Vasil, G.M., Lecoanet, D., Brown, B.P., Wood, T.S., Zweibel, E.G.: Energy conservation and gravity waves in sound-proof treatments of stellar interiors. ii. lagrangian constrained analysis. Astrophys. J. **773**, 169 (2013)
44. Wilhelmson, R., Ogura, Y.: The pressure perturbation and the numerical modeling of a cloud. J. Atmos. Sci. **29**, 1295–1307 (1972)
45. Woosley, S.E., Wunsch, S., Kuhlen, M.: Carbon ignition in type ia supernovae: an analytic model. Astrophys. J. **607**, 921–930 (2004)
46. Zingale, M., Almgren, A.S., Bell, J.B., Nonaka, A., Woosley, S.E.: Low mach number modeling of type ia supernovae. IV. white dwarf convection. Astrophys. J. **704**, 196–210 (2009)
47. Zingale, M., Nonaka, A., Almgren, A.S., Bell, J.B., Malone, C.M., Orvedahl, R.J.: Low mach number modeling of convection in helium shells on sub-chandrasekhar white dwarfs. I. Methodology. Astrophys. J. **764**, 97 (2013). doi:10.1088/0004-637X/764/1/97
48. Zingale, M., Nonaka, A., Almgren, A.S., Bell, J.B., Malone, C.M., Woosley, S.E.: The convective phase preceding type ia supernovae. Astrophys. J. **740**, 8 (2011)

Entropy Method and Asymptotic Behaviours of Finite Volume Schemes

Claire Chainais-Hillairet

Abstract When deriving a numerical scheme for a system of PDEs coming for instance from physics or engineering, it is crucial to propose a scheme which preserves the asymptotic behaviour of the continuous system, with respect to time as with respect to some parameters. In this paper, we want to show how the entropy method can be applied to some finite volume schemes and permits to show that some schemes are asymptotic preserving. We focus on two problems: the nonlinear diffusion equation (long time behaviour) and the drift-diffusion system (long time behaviour and quasi-neutral limit). Some results have been obtained in collaboration with Jüngel and Schuchnigg [10] and the others with Bessemoulin-Chatard and Vignal [4].

1 Introduction

1.1 Entropy Method and Long Time Behaviour

The entropy method is initially devoted to the study of the convergence to equilibrium of systems composed of a large number of particules. Roughly speaking, the trend to equilibrium is governed by a thermodynamical principle: a given functional, called physical entropy, increases when the time increases and the equilibrium is defined as the maximum of the entropy. The entropy method has been widely studied and applied since the beginning of the 90s: see [1] and all the references therein. As

C. Chainais-Hillairet (✉)
Laboratoire P. Painlevé, UMR CNRS 8524, Université Lille 1,
59655 Villeneuve d'Ascq Cedex, France
e-mail: Claire.Chainais@math.univ-lille1.fr

Team MEPHYSTO, INRIA Lille Nord Europe, 40 av. Halley,
59650 Villeneuve d'Ascq Cedex, France

J. Fuhrmann et al. (eds.), *Finite Volumes for Complex Applications VII - Methods and Theoretical Aspects*, Springer Proceedings in Mathematics & Statistics 77,
DOI: 10.1007/978-3-319-05684-5_2,

written in this paper, it appears that "entropy methods have proved over the last years to be an efficient tool for the understanding of the qualitative properties of physically sound models, for accurate numerics and for a more mathematical understanding of nonlinear PDEs".

In order to explain the principles of the entropy method, let us consider a system of partial differential equations written basically under the form:

$$\begin{aligned} &\partial_t f + Af = 0, \quad t \geq 0, \\ &f(0) = f_0, \end{aligned}$$

where A is a partial differential operator containing also the boundary conditions. A stationary state is defined by $Af_\infty = 0$. The question worthy of interest concerns the convergence of $f(t)$ towards f_∞ when t tends to $+\infty$. The strategy consists in proving the convergence in relative entropy: considering an entropy (a convex nonnegative Lyapunov functional), the idea is to prove that $E(f) \to E(f_\infty)$ or equivalently $E(f|f_\infty) = E(f) - E(f_\infty) \to 0$ when $t \to +\infty$. The result is based on the relation

$$\frac{d}{dt}(E(f(t)|f_\infty) + D(f(t))) = 0, \text{ with } D(f) = \langle Af, E'(f)\rangle.$$

The term $D(f)$ is the entropy dissipation. It must be nonnegative so that the entropy is nonincreasing (the mathematical entropy is the opposite of the physical entropy). Moreover, if the dissipation is related to the entropy thanks to some relation like $D(f) \geq \lambda E(f|f_\infty)$ (respectively $D(f) \geq K E(f|f_\infty)^{1+\gamma}$), an exponential (respectively polynomial) convergence of the relative entropy towards the equilibrium can be obtained.

This technique has been widely used for many systems of PDEs coming from the physics in many different areas of applications. We can refer to the survey paper [1] and the references therein. The entropy method has been applied for instance for electro-reaction-diffusion systems [22], thin-film type equations [5], reaction-diffusion equations [13], coagulation-fragmentation models [6].

In the sequel of the paper, we will consider two different problems: the nonlinear diffusion equation (porous medium/fast diffusion equation) and the drift-diffusion system coming from the modelling of semiconductor devices.

The Nonlinear Diffusion Equation

Let Ω be an open bounded domain of $\mathbb{R}^d$ such that $\mathrm{m}(\Omega) = 1$ and $\beta > 0$. We consider the following nonlinear diffusion equation supplemented with initial and homogeneous Neumann boundary conditions:

$$\partial_t u - \Delta(u^\beta) = 0, \text{ in } \Omega, t > 0 \text{ with } u(\cdot, 0) = u_0, \text{ in } \Omega, \tag{1a}$$

$$\nabla(u^\beta)\cdot\nu = 0,\ \text{on } \partial\Omega, t > 0. \tag{1b}$$

When $\beta > 1$, it is called the porous-medium equation, describing the flow of an isentropic gas through a porous medium. When $\beta < 1$, it is referred as the fast-diffusion equation. In [9], the entropy-entropy dissipation method was applied to (1a) in the whole space to prove the decay of the solutions to the asymptotic self-similar profile. The convergence towards the constant steady-state on the one-dimensional torus was proved in [7].

We note that the solution to (1a), (1b) satisfies $\int_\Omega u(x,t)dx = \int_\Omega u_0(x)dx$ for all $t \geq 0$. Therefore, the stationary state is constant and equal to $u_\infty = \int_\Omega u_0(x)dx$. In order to study the convergence towards the stationary state, we introduce the following family of zeroth-order relative entropies:

$$E_\alpha(u) = \frac{1}{\alpha+1}\left(\int_\Omega u^{\alpha+1}dx - \left(\int_\Omega u dx\right)^{\alpha+1}\right),\quad \alpha > 0. \tag{2}$$

In [10], we study, among other things, the algebraic and the exponential decay of these entropies. The functional inequalities relating entropy and dissipation are obtained from generalized Beckner inequalities.

The Drift-Diffusion System

The drift-diffusion-Poisson system has been introduced by van Roosbroeck [27] for the modelling of semiconductor devices. Let Ω be an open bounded set of $\mathbb{R}^d$ describing the geometry of the semiconductor device, the system writes:

$$\partial_t N + \operatorname{div}(\mu_N(-\nabla N + N\nabla\Psi)) = -R(N,P),\ \text{in } \Omega,\ t > 0, \tag{3a}$$

$$\partial_t P + \operatorname{div}(\mu_P(-\nabla P - P\nabla\Psi)) = -R(N,P),\ \text{in } \Omega,\ t > 0, \tag{3b}$$

$$-\lambda^2\Delta\Psi = P - N + C,\ \text{in } \Omega,\ t > 0. \tag{3c}$$

where the given function $C(x)$ is the doping profile and $R(N,P)$ the recombination-generation rate. The dimensionless physical parameters μ_N, μ_P and λ are the rescaled mobilities of electrons and holes and the rescaled Debye length. This system is generally supplemented with Dirichlet-Neumann boundary conditions ($\partial\Omega = \Gamma^D \cup \Gamma^N$):

$$N = N^D,\ P = P^D,\ \Psi = \Psi^D \text{ on } \Gamma^D \times (0,T), \tag{4a}$$

$$\nabla N\cdot\nu = 0,\ \nabla P\cdot\nu = 0,\ \nabla\Psi\cdot\nu = 0,\ \text{on } \Gamma^N \times (0,T), \tag{4b}$$

and with initial conditions N_0, P_0.

The stationary state for the drift-diffusion model is referred as the thermal equilibrium (N^*, P^*, Ψ^*). It is defined under some compatibility assumptions on the boundary data. The convergence of the solution to (3a)–(4b) towards the thermal equilibrium has been established by Jüngel in [24] (including the case of nonlinear diffusion) and Gajewski and Gärtner in [16] (for the linear system with magnetic field). Both proofs are based on an entropy method. In this case, the relative entropy is defined by:

$$\mathbb{E}(t) = \int_\Omega \Big(H(N) - H(N^*) - \log(N^*)(N - N^*) \\ + H(P) - H(P^*) - \log(P^*)(P - P^*) + \frac{\lambda^2}{2}|\nabla\Psi - \nabla\Psi^*|^2 \Big) dx,$$

with $H(x) = x \log x - x + 1$.

1.2 Entropy Method and Quasi-Neutral Limit

In the drift-diffusion model (3a)–(4b), the quasi-neutral limit consists in letting the scaled Debye length λ tend to 0. From a physical point of view, this means that only the large scale structures with respect to the Debye length are taken into account. For the sake of simplicity, we will now assume that $\mu_N = \mu_P = 1$, $R(N, P) = 0$ and that the doping profile vanishes. Under these hypotheses, the system (3a)–(4b) will be denoted $(\mathscr{P}_\lambda)$. The quasi-neutral limit is formally obtained by setting $\lambda = 0$ in $(\mathscr{P}_\lambda)$. It implies that the Poisson equation reduces to an algebraic equation on N and P. The system $(\mathscr{P}_0)$ rewrites:

$$\partial_t N - \Delta N = 0, \tag{5a}$$

$$\operatorname{div}(N\nabla\Psi) = 0, \tag{5b}$$

$$P = N. \tag{5c}$$

Jüngel and Peng [25] performed rigorously the quasi-neutral limit for the drift-diffusion system with a zero doping profile and mixed Dirichlet and homogeneous Neumann boundary conditions. Under quasi-neutrality assumptions on the initial and boundary conditions ($N_0 - P_0 = 0$ and $N^D - P^D = 0$), they prove that a weak solution to $(\mathscr{P}_\lambda)$, denoted by $(N^\lambda, P^\lambda, \Psi^\lambda)$, converges, when $\lambda \to 0$, to (N^0, P^0, Ψ^0) solution to $(\mathscr{P}_0)$ in the following sense:

$$N^\lambda \to N^0,\ P^\lambda \to P^0 \text{ in } L^p(\Omega \times (0,T)) \text{ strongly, for all } p \in [1, +\infty),$$
$$N^\lambda \rightharpoonup N^0,\ P^\lambda \rightharpoonup P^0,\ \Psi^\lambda \rightharpoonup \Psi^0 \text{ in } L^2(0, T, H^1(\Omega)) \text{ weakly.}$$

The same kind of result is established for the drift-diffusion system with homogeneous Neumann boundary conditions by Gasser in [17] for a zero doping profile and by Gasser et al. in [18] for a regular doping profile. In all these papers, the rigorous proof of the quasi-neutral limit is based on an entropy method.

In the case of Dirichlet-Neumann boundary conditions, we will consider that the boundary data N^D, P^D, Ψ^D are defined on the whole domain Ω and verify N^D, $P^D \in L^\infty \cap H^1(\Omega)$, $\Psi^D \in H^1(\Omega)$. Then, the entropy functional, which has the physical meaning of a free energy, is defined (see [25]) by

$$\mathbb{E}(t) = \int_\Omega \Big(H(N) - H(N^D) - \log(N^D)(N - N^D) \\ + H(P) - H(P^D) - \log(P^D)(P - P^D) + \frac{\lambda^2}{2} |\nabla\Psi - \nabla\Psi^D|^2 \Big) dx$$

and the entropy dissipation functional is defined by

$$\mathbb{D}(t) = \int_\Omega \left(N\, |\nabla(\log N - \Psi)|^2 + P\, |\nabla(\log P + \Psi)|^2 \right) dxdt.$$

The entropy and the entropy dissipation satisfy the following relation:

$$\frac{d\mathbb{E}}{dt}(t) + \frac{1}{2}\mathbb{D}(t) \leq K_D \quad \forall t \geq 0, \tag{6}$$

where K_D is a constant depending only on data. This inequality is crucial in order to perform rigorously the quasi-neutral limit. Indeed, if $\mathbb{E}(0)$ is uniformly bounded in λ, (6) provides a uniform bound on $\int_0^T \mathbb{D}(s)ds$. It implies a priori uniform bounds on $(N^\lambda, P^\lambda, \Psi^\lambda)$ solution to $(\mathscr{P}_\lambda)$ and therefore compactness of a sequence of solutions.

1.3 Aim of the Paper

The preservation of the structure of the equations (or system of equations) is a very important property of a numerical scheme. Positivity, maximum principle, appropriate a priori estimates are the bases for the proof of convergence of finite volume schemes for instance. The properties of entropy consistency or entropy dissipation by numerical schemes are also crucial and have been investigated in different frameworks, see for instance [8, 14, 19–21, 23].

In this paper, we want to present some recent results obtained with Jüngel and Schuchnigg for the nonlinear diffusion equation [10] and with Bessemoulin-Chatard and Vignal for the drift-diffusion system [4]. In both cases, we study the asymptotic behaviour of some finite volume schemes using a discrete entropy method.

Section 2 is devoted to the presentation of the notations. In Sect. 3, we are interested in the long time behaviour of some numerical schemes. We first present results

obtained in [10] for the nonlinear diffusion equation. In this case, thanks to discrete functional inequalities, we can establish polynomial or exponential decay of a family of discrete relative entropies. We will also mention some known results for the numerical approximation of the drift-diffusion system.

In Sect. 4, we consider a Euler implicit in time and finite volume in space scheme for the drift-diffusion system. With the choice of Scharfetter-Gummel approximation for the convection-diffusion fluxes [26], we can derive a discrete counterpart of (6). We then prove that the scheme is asymptotic preserving at the quasi-neutral limit : it converges for all $\lambda \geq 0$ and the corresponding limit (N, P, Ψ) is a solution to $(\mathscr{P}_\lambda)$, for $\lambda > 0$ as for $\lambda = 0$.

2 Notations

In order to define the numerical schemes under consideration in this paper, we need to introduce the discretization settings and some notations. We restrict the presentation to a two-dimensional case but generalization to higher dimension is straightforward. We consider that Ω is an open bounded polygonal subset of $\mathbb{R}^2$.

The mesh $\mathscr{M} = (\mathscr{T}, \mathscr{E}, \mathscr{P})$ is given by $\mathscr{T}$, a family of open polygonal control volumes, $\mathscr{E}$, a family of edges and $\mathscr{P} = (x_K)_{K\in\mathscr{T}}$ a family of points. As it is classical in the finite volume discretization of elliptic or parabolic equations with a two-points flux approximations, we assume that the mesh is admissible in the sense of [15] (Definition 9.1).

We distinguish in $\mathscr{E}$ the interior edges, $\sigma = K|L$, from the exterior edges, $\sigma \subset \partial\Omega$. Therefore $\mathscr{E}$ is split into $\mathscr{E} = \mathscr{E}_{int} \cup \mathscr{E}_{ext}$. Within the exterior edges, we distinguish (if necessary) the edges included in Γ^D from the edges included in Γ^N: $\mathscr{E}_{ext} = \mathscr{E}_{ext}^D \cup \mathscr{E}_{ext}^N$. For a given control volume $K \in \mathscr{T}$, we define $\mathscr{E}_K$ the set of its edges, which is also split into $\mathscr{E}_K = \mathscr{E}_{K,int} \cup \mathscr{E}_{K,ext}^D \cup \mathscr{E}_{K,ext}^N$. For each edge $\sigma \in \mathscr{E}$, there exists at least one cell $K \in \mathscr{T}$ such that $\sigma \in \mathscr{E}_K$. Then, we can denote this cell K_σ. In the case where σ is an interior edge ($\sigma = K|L$), K_σ can be either equal to K or to L.

For all edges $\sigma \in \mathscr{E}$, we define $\mathrm{d}_\sigma = \mathrm{d}(x_K, x_L)$ if $\sigma = K|L \in \mathscr{E}_{int}$ and $\mathrm{d}_\sigma = \mathrm{d}(x_K, \sigma)$ if $\sigma \in \mathscr{E}_{ext}$ with $\sigma \in \mathscr{E}_K$. Then, the transmissibility coefficient is defined by $\tau_\sigma = \mathrm{m}(\sigma)/\mathrm{d}_\sigma$, for all $\sigma \in \mathscr{E}$. We assume that the mesh satisfies the following regularity constraint:

$$\exists \xi > 0 \text{ such that } \mathrm{d}(x_K, \sigma) \geq \xi\, \mathrm{d}_\sigma, \quad \forall K \in \mathscr{T}, \forall \sigma \in \mathscr{E}_K. \tag{7}$$

Let $T > 0$, we consider a subdivision of the interval $[0, T]$ defined by $(t^n = n\Delta t)_{0\leq n\leq N_T}$, where Δt is the time step and $N_T \Delta t = T$. A classical finite volume approximation provides an approximate solution which is constant on each cell of the mesh and on each time interval. Let $X(\mathscr{T})$ be the linear space of functions $\Omega \to \mathbb{R}$ which are constant on each cell $K \in \mathscr{T}$. To a discrete set $(u_K)_{K\in\mathscr{T}}$, we associate

$u_{\mathcal{T}} = \sum_{K\in\mathcal{T}} u_K \mathbf{1}_K \in X(\mathcal{T})$. The L^p-norm of $u_{\mathcal{T}}$ is

$$\|u_{\mathcal{T}}\|_{0,p} = \left(\sum_{K\in\mathcal{T}} \mathrm{m}(K)|u_K|^p\right)^{1/p}.$$

When there are Dirichlet boundary conditions on a part of the boundary, we need to define approximate values for u at the corresponding boundary edges: $u_{\mathcal{E}^D} = (u_\sigma)_{\sigma\in\mathcal{E}^D_{ext}} \in \mathbb{R}^{\theta^D}$ (with $\theta^D = \mathrm{Card}(\mathcal{E}^D_{ext})$). Therefore, the vector containing the approximate values in the control volumes and the approximate values at the boundary edges is denoted by $u_{\mathcal{M}} = (u_{\mathcal{T}}, u_{\mathcal{E}^D})$. For any vector $u_{\mathcal{M}} = (u_{\mathcal{T}}, u_{\mathcal{E}^D})$, we define, for all $K \in \mathcal{T}$, for all $\sigma \in \mathcal{E}_K$,

$$u_{K,\sigma} = \begin{cases} u_L, & \text{if } \sigma = K|L \in \mathcal{E}_{K,int}, \\ u_\sigma, & \text{if } \sigma \in \mathcal{E}^D_{K,ext}, \\ u_K, & \text{if } \sigma \in \mathcal{E}^N_{K,ext}, \end{cases} \tag{8a}$$

$$Du_{K,\sigma} = u_{K,\sigma} - u_K \quad \text{and} \quad D_\sigma u = \left|Du_{K,\sigma}\right|. \tag{8b}$$

It permits to define the discrete H^1-semi-norm $|\cdot|_{1,2,\mathcal{M}}$:

$$|u_{\mathcal{M}}|^2_{1,2,\mathcal{M}} = \sum_{\sigma\in\mathcal{E}} \tau_\sigma \, (D_\sigma u)^2, \quad \forall u_{\mathcal{M}} = (u_{\mathcal{T}}, u_{\mathcal{E}^D}).$$

If $\mathcal{E}^D = \emptyset$, we have $u_{\mathcal{M}} = u_{\mathcal{T}}$ and we will write $|u_{\mathcal{T}}|_{1,2,\mathcal{T}} = |u_{\mathcal{M}}|_{1,2,\mathcal{M}}$.

3 Long Time Behaviour of Some Finite Volume Schemes

3.1 First Example: Nonlinear Diffusion Equations

In this section, we consider a classical Euler implicit in time and finite volume in space discretization of the nonlinear diffusion Eq. (1a), (1b).

Theoretical Results

We assume that $u_0 \in L^\infty(\Omega)$, with $m \le u_0 \le M$ a.e. on Ω, with $m \ge 0$. For the sake of simplicity, we also assume that $\mathrm{m}(\Omega) = 1$. The scheme writes:

$$\mathrm{m}(K)\frac{u_K^{n+1}-u_K^n}{\Delta t}+\sum_{\substack{\sigma\in\mathscr{E}_{K,int},\\ \sigma=K|L}}\tau_\sigma\left((u_K^{n+1})^\beta-(u_L^{n+1})^\beta\right)=0, \tag{9a}$$

$$u_K^0=\frac{1}{\mathrm{m}(K)}\int_K u_0(x)dx. \tag{9b}$$

Existence and uniqueness of a discrete solution to (9a), (9b) is a well-known result (see [15]). Moreover, it is clear that $m\le u_K^n\le M$ for all $K\in\mathscr{T}$ and for all $0\le n\le N_T$. Due to the Neumann boundary conditions, we also have:

$$\sum_{K\in\mathscr{T}}\mathrm{m}(K)u_K^n=\|u_0\|_{L^1(\Omega)}.$$

At each time step, we can reconstruct the approximate solution $u_\mathscr{T}^n\in X(\mathscr{T})$. Our aim is to study the convergence of $(u_\mathscr{T}^n)_{n\ge0}$ when n tends to $+\infty$ towards the constant function equal to $\|u_0\|_{L^1(\Omega)}$. Therefore, we can use the relative entropies E_α defined in (2) for $\alpha>0$. Let us note that

$$E_\alpha[u_\mathscr{T}^n]=\frac{1}{\alpha+1}\left(\sum_{K\in\mathscr{T}}\mathrm{m}(K)(u_K^n)^{\alpha+1}-\left(\sum_{K\in\mathscr{T}}\mathrm{m}(K)u_K^n\right)^{\alpha+1}\right),$$

Using the convexity of the function $x\mapsto x^{\alpha+1}$ and the scheme (9a), (9b), we easily get:

$$E_\alpha[u_\mathscr{T}^{n+1}]-E_\alpha[u_\mathscr{T}^n]\le-\Delta t\sum_{\substack{\sigma\in\mathscr{E}_{int},\\ \sigma=K|L}}\tau_\sigma\left((u_K^{n+1})^\alpha-(u_L^{n+1})^\alpha\right)\left((u_K^{n+1})^\beta-(u_L^{n+1})^\beta\right).$$

Then, using the following inequality:

$$(y^\alpha-x^\alpha)(y^\beta-x^\beta)\ge\frac{4\alpha\beta}{(\alpha+\beta)^2}(y^{(\alpha+\beta)/2}-x^{(\alpha+\beta)/2})^2,\quad\forall x,y\ge0,$$

we get that

$$E_\alpha[u_\mathscr{T}^{n+1}]-E_\alpha[u_\mathscr{T}^n]\le-\frac{4\alpha\beta\Delta t}{(\alpha+\beta)^2}\left|(u_\mathscr{T}^{n+1})^{(\alpha+\beta)/2}\right|_{1,2,\mathscr{T}}^2. \tag{10}$$

With another choice of inequality:

$$(y^\beta-x^\beta)(y^\alpha-x^\alpha)\ge\frac{4\alpha\beta}{(\alpha+1)^2}\min(x^{\beta-1},y^{\beta-1})(y^{(\alpha+1)/2}-x^{(\alpha+1)/2})^2,\ \forall x,y\ge0,$$

we get:

$$E_\alpha[u_{\mathscr{T}}^{n+1}] - E_\alpha[u_{\mathscr{T}}^{n}] \le -\frac{4\alpha\beta\,\Delta t}{(\alpha+1)^2} \inf_{K\in\mathscr{T}} (u_K^{n+1})^{\beta-1} \left| (u_{\mathscr{T}}^{n+1})^{(\alpha+1)/2} \right|_{1,2,\mathscr{T}}^2 . \quad (11)$$

In both cases, the dissipation of the entropy of the approximate solution is stated in terms of the discrete H^1-semi-norm of some discrete function. In order to relate the dissipation to the entropy, we need some functional inequalities. The relation between either $|(u_{\mathscr{T}}^{n+1})^{(\alpha+\beta)/2}|_{1,2,\mathscr{T}}^2$ or $|(u_{\mathscr{T}}^{n+1})^{(\alpha+1)/2}|_{1,2,\mathscr{T}}^2$, to $E_\alpha[u_{\mathscr{T}}^{n+1}]$ will be done through discrete generalized Beckner inequalities, established in [10].

Lemma 1

- *Let* $0 < q < 2$, $pq > 1$ *or* $q = 2$ *and* $0 < p \le 1$, *and* $f_{\mathscr{T}} \in X(\mathscr{T})$. *Then*

$$\int_\Omega |f_{\mathscr{T}}|^q dx - \left(\int_\Omega |f_{\mathscr{T}}|^{1/p} dx \right)^{pq} \le \frac{C_b(p,q)}{\xi^{q/2}} |f_{\mathscr{T}}|_{1,2,\mathscr{T}}^q \quad (12)$$

  holds, where $C_b(p,q)$ *only depends on* p, q, Ω *and with* ξ *defined in* (7).
- *Let* $0 < q < 2$, $pq \ge 1$, *and* $f_{\mathscr{T}} \in X(\mathscr{T})$. *Then*

$$\|f_{\mathscr{T}}\|_{0,q,\mathscr{T}}^{2-q} \left(\int_\Omega |f_{\mathscr{T}}|^q dx - \left(\int_\Omega |f_{\mathscr{T}}|^{1/p} dx \right)^{pq} \right) \le \frac{C_b'(p,q)}{\xi} |f_{\mathscr{T}}|_{1,2,\mathscr{T}}^2 \quad (13)$$

  holds, where $C_b'(p,q)$ *only depends on* p, q, Ω *and with* ξ *defined in* (7).

Applying (12) with $p = (\alpha+\beta)/2$, $q = 2(\alpha+1)/(\alpha+\beta)$ and $f_{\mathscr{T}} = (u_{\mathscr{T}}^{n+1})^{(\alpha+\beta)/2}$, we deduce from (10):

$$E_\alpha[u_{\mathscr{T}}^{n+1}] - E_\alpha[u_{\mathscr{T}}^{n}] \le -K\,\Delta t\, E_\alpha[u_{\mathscr{T}}^{n+1}]^{(\alpha+\beta)/(\alpha+1)},$$

with K depending on α, β, Ω and ξ. Then, a discrete nonlinear Gronwall lemma (see [10]) leads to the polynomial decay of the discrete entropy.

Theorem 1 (Polynomial decay) *Let* $\alpha > 0$ *and* $\beta > 1$. *Let* $(u_{\mathscr{T}}^n)_{n\ge 0}$ *be the solution to the finite-volume scheme* (9a), (9b) *with* $\inf_{K\in\mathscr{T}} u_K^0 \ge 0$. *Then*

$$E_\alpha[u_{\mathscr{T}}^{n}] \le \frac{1}{(c_1 t^n + c_2)^{(\alpha+1)/(\beta-1)}}, \quad \forall n \ge 0,$$

where c_1 *depends on* α, β, Ω, ξ *and* Δt *(but stays bounded when* Δt *tends to 0) and* $c_2 = E_\alpha[u_{\mathscr{T}}^0]^{-(\beta-1)/(\alpha+1)}$.

Applying (13) with $p = (\alpha+1)/2$, $q = 2$ and $f_{\mathscr{T}} = (u_{\mathscr{T}}^{n+1})^{(\alpha+1)/2}$, we deduce from (11):

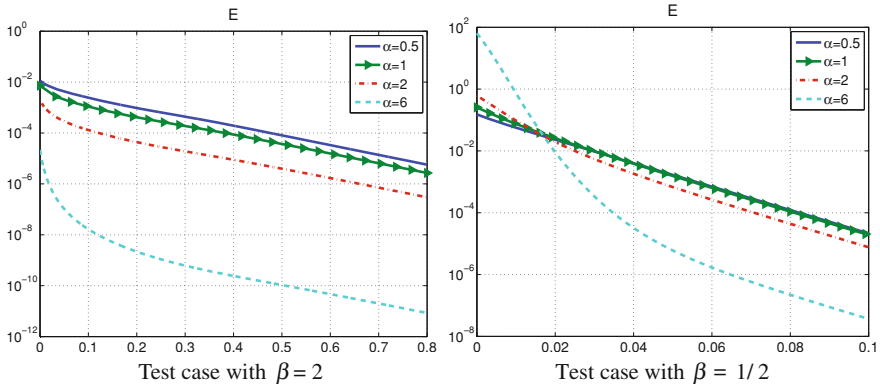


Fig. 1 Evolution of the discrete entropies with respect to time for different values of α

$$E_\alpha[u^{n+1}_{\mathcal{T}}] - E_\alpha[u^n_{\mathcal{T}}] \leq -K'\Delta t \inf_{K\in\mathcal{T}} (u^0_K)^{\beta-1} E_\alpha[u^{n+1}_{\mathcal{T}}],$$

with K' depending on α, β, Ω and ξ. Then, we can conclude to the exponential decay of the discrete entropy.

Theorem 2 (Exponential decay) *Let* $0 < \alpha \leq 1$ *and* $\beta > 0$. *Let* $(u^n_{\mathcal{T}})_{n\geq 0}$ *be the solution to the finite-volume scheme* (9a), (9b) *with* $\inf_{K\in\mathcal{T}} u^0_K \geq 0$. *Then*

$$E_\alpha[u^n_{\mathcal{T}}] \leq E_\alpha[u^0_{\mathcal{T}}] e^{-\lambda t^n}, \quad \forall n \geq 0,$$

with λ *depending on* α, β, Ω, ξ *and* $\inf_{K\in\mathcal{T}} (u^0_K)^{\beta-1}$.

Numerical Experiments

We illustrate on Fig. 1 the time decay of the solutions to the discretized porous-medium equation ($\beta = 2$) and to the fast-diffusion equation ($\beta = 1/2$). Both test cases are two-dimensional, with $\Omega = (0,1) \times (0,1)$. When $\beta = 2$, we choose a Barenblatt profile as initial condition. We observe that the decay of the discrete entropies seems to be exponential for large times, even for values of α not covered by Theorem 2. When $\beta = 1/2$, we choose $u_0(x) = C((R^2 - |x - x_0|^2)^+)^2$ with $x_0 = (0.5, 0.5)$, $R = 0.2$, $C = 3000$ as initial condition. We observe similarly an exponential decay of the discrete entropies for large times.

3.2 Second Example: Drift-Diffusion System

As recalled in the introduction, the drift-diffusion system (3a), (3b), (3c) is made of two convection-diffusion-reaction equations on the densities coupled with a Poisson equation on the electric potential. Writing a two-points finite volume scheme for this system is not difficult. However, the choice of the time discretization and of the numerical approximation of the convection-diffusion fluxes will be crucial for the preservation of the asymptotic behaviours.

When writing the scheme, we have to define for instance $\mathscr{F}_{K,\sigma}$ the numerical approximation of $\int_\sigma(-\nabla N + N\nabla\Psi)\cdot\nu_{K,\sigma}$. Scharfetter and Gummel [26] have proposed to discretize simultaneously the convection and diffusion terms. It leads to the following numerical fluxes:

$$\mathscr{F}_{K,\sigma} = \tau_\sigma\left(B(-D\Psi_{K,\sigma})N_K - B(D\Psi_{K,\sigma})N_{K,\sigma}\right)$$

where B is the Bernoulli function defined by:

$$B(0) = 1 \text{ and } B(x) = \frac{x}{\exp(x)-1}\ \forall x \neq 0. \tag{14}$$

Gajewski and Gärtner [16] have shown that the Euler implicit in time and finite volume in space scheme, with a Scharfetter-Gummel approximation of the convection-diffusion fluxes, is entropy dissipative (the scheme is detailed in the next section). Later, Chatard [12] has also obtained a discrete counterpart of the entropy method for this scheme (with a different way of proof). The numerical experiments in [12] show the exponential decay in time of the discrete entropy for the Scharfetter-Gummel scheme. They also show that this property is no more satisfied by the scheme proposed in [11], where the diffusion terms are discretized classically and the convection terms are discretized with upwind fluxes.

4 Finite Volume Scheme at the Quasi-Neutral Limit

In this Section, we study a numerical scheme for the simplified drift-diffusion system $(\mathscr{P}_\lambda)$, similar to the schemes studied in [16] or in [12]. We will use the entropy method in order to show that the scheme is asymptotic preserving at the quasi-neutral limit $\lambda \to 0$. More precisely, we will establish that the a priori estimates needed for the proof of convergence hold for all $\lambda \geq 0$.

We make the following assumptions on the data:

$$N_0,\, P_0 \in L^\infty(\Omega), \tag{15a}$$

$$N^D,\, P^D \in L^\infty \cap H^1(\Omega),\ \Psi^D \in H^1(\Omega), \tag{15b}$$

$$\exists m > 0, M > 0 \text{ such that } m \leq N_0,\, P_0,\, N^D,\, P^D \leq M \text{ a.e. on } \Omega. \tag{15c}$$

4.1 Presentation of the Scheme

For $u = N, P, \Psi$, the approximate solution is defined by $u^n_{\mathscr{T}}$ and the approximate values at the boundary are $u^n_{\mathscr{E}^D} = (u^n_\sigma)_{\sigma\in\mathscr{E}^D_{ext}}$, at each time step, $0 \le n \le N_T$. Let us first discretize the initial and the boundary conditions. We set:

$$u^0_K = \frac{1}{\mathrm{m}(K)}\int_K u_0(x)\,dx, \quad \forall K \in \mathscr{T}, \text{ for } u = N, P, \tag{16}$$

$$u^D_\sigma = \frac{1}{\mathrm{m}(\sigma)}\int_\sigma u(\gamma)d\gamma, \quad \forall \sigma \in \mathscr{E}^D_{ext}, \text{ for } u = N, P, \Psi.$$

Moreover, we define

$$u^n_\sigma = u^D_\sigma, \quad \forall \sigma \in \mathscr{E}^D_{ext}, \forall n \ge 0, \text{ for } u = N, P, \Psi. \tag{17}$$

We consider a Euler implicit in time and finite volume in space discretization. The scheme writes:

$$\mathrm{m}(K)\frac{N^{n+1}_K - N^n_K}{\Delta t} + \sum_{\sigma\in\mathscr{E}_K}\mathscr{F}^{n+1}_{K,\sigma} = 0, \quad \forall K \in \mathscr{T}, \forall n \ge 0, \tag{18a}$$

$$\mathrm{m}(K)\frac{P^{n+1}_K - P^n_K}{\Delta t} + \sum_{\sigma\in\mathscr{E}_K}\mathscr{G}^{n+1}_{K,\sigma} = 0, \quad \forall K \in \mathscr{T}, \forall n \ge 0, \tag{18b}$$

$$-\lambda^2\sum_{\sigma\in\mathscr{E}_K}\tau_\sigma D\Psi^n_{K,\sigma} = \mathrm{m}(K)(P^n_K - N^n_K), \quad \forall K \in \mathscr{T}, \forall n \ge 0. \tag{18c}$$

We choose a Scharfetter-Gummel approximation for the convection-diffusion fluxes:

$$\mathscr{F}^{n+1}_{K,\sigma} = \tau_\sigma\left(B(-D\Psi^{n+1}_{K,\sigma})N^{n+1}_K - B(D\Psi^{n+1}_{K,\sigma})N^{n+1}_{K,\sigma}\right), \quad \forall K \in \mathscr{T}, \forall \sigma \in \mathscr{E}_K, \tag{19a}$$

$$\mathscr{G}^{n+1}_{K,\sigma} = \tau_\sigma\left(B(D\Psi^{n+1}_{K,\sigma})P^{n+1}_K - B(-D\Psi^{n+1}_{K,\sigma})P^{n+1}_{K,\sigma}\right), \quad \forall K \in \mathscr{T}, \forall \sigma \in \mathscr{E}_K, \tag{19b}$$

where B is the Bernoulli function defined by (14).

In the sequel, we denote by $(\mathscr{S}_\lambda)$ the scheme (16)–(19b). It is a fully implicit in time scheme: the numerical solution $(N^{n+1}_K, P^{n+1}_K, \Psi^{n+1}_K)_{K\in\mathscr{T}}$ at each time step is defined as a solution of the nonlinear system of Eqs. (18a)–(19b). When choosing $D\Psi^n_{K,\sigma}$ instead of $D\Psi^{n+1}_{K,\sigma}$ in the definition of the fluxes (19a), (19b), we would get a decoupled scheme whose solution is obtained by solving successively three linear systems of equations for N, P and Ψ. However, this other choice of time

discretization induces a stability condition of the form $\Delta t \leq C\lambda^2$ (see for instance [2]). Therefore, it cannot be used in practice for small values of λ and it does not preserve the quasi-neutral limit.

Setting $\lambda = 0$ in the scheme $(\mathscr{S}_\lambda)$ leads to the scheme $(\mathscr{S}_0)$ defined hereafter. The scheme for the Poisson Eq. (18c) becomes $P_K^n - N_K^n = 0$ for all $K \in \mathscr{T}, n \in \mathbb{N}$. In order to avoid any incompatibility condition at $n = 0$ (which would correspond to an initial layer), we assume that the initial conditions N_0 and P_0 satisfy the quasi-neutrality assumption:

$$P_0 - N_0 = 0. \tag{20}$$

Adding and subtracting (18a) and (18b), and using $P_K^n = N_K^n$ for all $K \in \mathscr{T}$ and $n \in \mathbb{N}$, we get

$$\mathrm{m}(K)\frac{N_K^{n+1} - N_K^n}{\Delta t} + \frac{1}{2}\sum_{\sigma\in\mathscr{E}_K}\left(\mathscr{F}_{K,\sigma}^{n+1} + \mathscr{G}_{K,\sigma}^{n+1}\right) = 0, \forall K \in \mathscr{T}, \forall n \geq 0,$$
$$\text{and} \quad \sum_{\sigma\in\mathscr{E}_K}\left(\mathscr{F}_{K,\sigma}^{n+1} - \mathscr{G}_{K,\sigma}^{n+1}\right) = 0, \forall K \in \mathscr{T}, \forall n \geq 0.$$

But, using the following property of the Bernoulli function $B(x) - B(-x) = -x$, $\forall x \in \mathbb{R}$, we have, $\forall K \in \mathscr{T}, \forall \sigma \in \mathscr{E}_{K,int} \cup \mathscr{E}_{K,ext}^N$:

$$\mathscr{F}_{K,\sigma}^{n+1} - \mathscr{G}_{K,\sigma}^{n+1} = \tau_\sigma D\Psi_{K,\sigma}^{n+1}(N_K^{n+1} + N_{K,\sigma}^{n+1}),$$
$$\text{and} \quad \mathscr{F}_{K,\sigma}^{n+1} + \mathscr{G}_{K,\sigma}^{n+1} = -\tau_\sigma\left(B(D\Psi_{K,\sigma}^{n+1}) + B(-D\Psi_{K,\sigma}^{n+1})\right)DN_{K,\sigma}^{n+1}.$$

Let us note that these equalities still hold for each Dirichlet boundary edge $\sigma \in \mathscr{E}_{K,ext}^D$ if $N_\sigma^D = P_\sigma^D$. In the sequel, when studying the scheme at the quasi-neutral limit $(\mathscr{S}_0)$, we assume the quasi-neutrality of the initial conditions (20) and of the boundary conditions:

$$P^D - N^D = 0. \tag{21}$$

Finally, the scheme $(\mathscr{S}_0)$ can be rewritten: $\forall K \in \mathscr{T}, \forall n \geq 0$,

$$\mathrm{m}(K)\frac{N_K^{n+1} - N_K^n}{\Delta t} - \sum_{\sigma\in\mathscr{E}_K}\tau_\sigma\frac{B(D\Psi_{K,\sigma}^{n+1}) + B(-D\Psi_{K,\sigma}^{n+1})}{2}DN_{K,\sigma}^{n+1} = 0, \tag{22a}$$

$$-\sum_{\sigma\in\mathscr{E}_K}\tau_\sigma D\Psi_{K,\sigma}^{n+1}(N_K^{n+1} + N_{K,\sigma}^{n+1}) = 0, \tag{22b}$$

$$P_K^n - N_K^n = 0, \tag{22c}$$

with the initial conditions (15a), (15b), (15c) and the boundary conditions (17).

Existence of a solution to the scheme $(\mathscr{S}_\lambda)$ is proved in [4] without any condition on Δt and for all $\lambda \geq 0$. Moreover, under the hypotheses (15a), (15b), (15c), we have:

$$m \leq N_K^n, P_K^n \leq M, \forall n \in \mathbb{N}.$$

It means in particular that all the densities are positive.

4.2 Entropy-Dissipation Estimate

In this Section, we establish the discrete counterpart of the entropy-dissipation inequality (6). As N^D, P^D and Ψ^D are defined on the whole domain, we can set:

$$u_K^D = \frac{1}{\mathrm{m}(K)} \int_K u(x)dx, \ \forall K \in \mathscr{T}, \ \text{for } u = N, P, \Psi.$$

Then, for all $n \in \mathbb{N}$, the discrete entropy is defined by:

$$\begin{aligned}\mathbb{E}^n = &\sum_{K\in\mathscr{T}} \mathrm{m}(K) \left(H(N_K^n) - H(N_K^D) - \log(N_K^D)\left(N_K^n - N_K^D\right)\right) \\ &+ \sum_{K\in\mathscr{T}} \mathrm{m}(K) \left(H(P_K^n) - H(P_K^D) - \log(P_K^D)(P_K^n - P_K^D)\right) \\ &+ \frac{\lambda^2}{2} \left|\Psi_{\mathscr{M}}^n - \Psi_{\mathscr{M}}^D\right|_{1,2,\mathscr{M}}^2,\end{aligned}$$

and the discrete entropy dissipation by:

$$\begin{aligned}\mathbb{D}^n = \sum_{\substack{\sigma\in\mathscr{E},\\(K=K_\sigma)}} \tau_\sigma \Big[&\min\left(N_K^n, N_{K,\sigma}^n\right)\left(D_\sigma\left(\log N^n - \Psi^n\right)\right)^2 \\ &+ \min\left(P_K^n, P_{K,\sigma}^n\right)\left(D_\sigma\left(\log P^n + \Psi^n\right)\right)^2\Big],\end{aligned}$$

where the notation $\sum_{\substack{\sigma\in\mathscr{E},\\(K=K_\sigma)}}$ means a sum over all the edges $\sigma \in \mathscr{E}$, with $K = K_\sigma$ (and therefore σ is an edge of the cell K) in the term inside the sum.

Theorem 3 (Discrete entropy-dissipation inequality) *Let assume* (15a), (15b), (15c) *and let* $\mathscr{T}$ *be an admissible mesh of* Ω *satisfying* (7) *and* $\Delta t > 0$. *Then, there exists* K_E, *not depending on* λ, Δt *and* size$(\mathscr{T})$, *such that, for all* $\lambda \geq 0$, *a solution to the scheme* $(\mathscr{S}_\lambda)$, $(N_\mathscr{T}^n, P_\mathscr{T}^n, \Psi_\mathscr{T}^n)_{0\leq n\leq N_T}$, *satisfies the following inequality:*

$$\frac{\mathbb{E}^{n+1}-\mathbb{E}^{n}}{\Delta t}+\frac{1}{2}\,\mathbb{D}^{n+1}\leq K_E,\quad \forall n\geq 0. \tag{23a}$$

Furthermore, if N^0 and P^0 satisfy the quasi-neutrality assumption (20), we have

$$\sum_{n=0}^{N_T-1}\Delta t\,\mathbb{D}^{n+1}\leq K_E(1+\lambda^2). \tag{23b}$$

Sketch of the proof We mimic the proof at the continuous level. The scheme on N (18a) is multiplied by $\Delta t\,(\log(N_K^{n+1}-\log(N_K^D))$ and a sum over the control volumes $K\in\mathcal{T}$ is achieved. A similar procedure is applied to the scheme on P (18b). Both terms are summed up and the sums are rearranged in order to use the scheme on Ψ (18c). In order to let the discrete entropy dissipation appear $\mathbb{D}^{n+1}$, we crucially use the discretization by the Scharfetter-Gummel fluxes. In practice, the result is based on the following properties satisfied by the Scharfetter-Gummel fluxes:

$$\begin{aligned}\mathscr{F}_{K,\sigma}^{n+1}\,D(\log N-\Psi)_{K,\sigma}^{n+1}\;&\leq-\tau_\sigma\,\min(N_K^{n+1},N_{K,\sigma}^{n+1})\left(D_\sigma(\log N-\Psi)^{n+1}\right)^2,\\ \mathscr{G}_{K,\sigma}^{n+1}\,D(\log P+\Psi)_{K,\sigma}^{n+1}\;&\leq-\tau_\sigma\,\min(P_K^{n+1},P_{K,\sigma}^{n+1})\left(D_\sigma(\log P+\Psi)^{n+1}\right)^2.\end{aligned}$$

Moreover, if $\min(N_K^{n+1},N_{K,\sigma}^{n+1})\geq 0$ and $\min(P_K^{n+1},P_{K,\sigma}^{n+1})\geq 0$, we also have

$$\begin{aligned}\left|\mathscr{F}_{K,\sigma}^{n+1}\right|&\leq\tau_\sigma\,\max(N_K^{n+1},N_{K,\sigma}^{n+1})\left|D_\sigma\,(\log N-\Psi)^{n+1}\right|,\\ \left|\mathscr{G}_{K,\sigma}^{n+1}\right|&\leq\tau_\sigma\,\max(P_K^{n+1},P_{K,\sigma}^{n+1})\left|D_\sigma\,(\log P+\Psi)^{n+1}\right|.\end{aligned}$$

4.3 New a Priori Estimates in Order to Get the Compactness

As it is classical in the finite volume framework and especially for elliptic and parabolic equations, we want to prove some a priori estimates satisfied by the discrete solution. In our case, it is crucial to establish a priori estimates which remain satisfied when $\lambda\to 0$. They will be deduced from the bound on the entropy dissipation (23b).

Theorem 4 (A priori estimates satisfied by the approximate solution) *Let assume (15a), (15b), (15c) and let $\mathcal{T}$ be an admissible mesh of Ω satisfying (7) and $\Delta t>0$. We also assume that the initial and boundary conditions satisfy the quasi-neutrality relations (20) and (21). Then, there exists a constant K_F not depending on λ, Δt and* size$(\mathcal{T})$*, such that, for all $\lambda\geq 0$, a solution to the scheme $(\mathscr{S}_\lambda)$, $(N_{\mathcal{T}}^n,P_{\mathcal{T}}^n,\Psi_{\mathcal{T}}^n)_{0\leq n\leq N_T}$, satisfies the following inequalities:*

$$\sum_{n=0}^{N_T-1} \Delta t \sum_{\sigma\in\mathscr{E}} \tau_\sigma D_\sigma \Psi^{n+1}\Big((D_\sigma P^{n+1})^2 + (D_\sigma N^{n+1})^2\Big) \leq K_F(1+\lambda^2), \quad (24a)$$

$$\sum_{n=0}^{N_T-1} \Delta t \sum_{\sigma\in\mathscr{E}} \tau_\sigma (D_\sigma N^{n+1})^2 + \sum_{n=0}^{N_T-1} \Delta t \sum_{\sigma\in\mathscr{E}} \tau_\sigma (D_\sigma P^{n+1})^2 \leq K_F(1+\lambda^2), \quad (24b)$$

$$\sum_{n=0}^{N_T-1} \Delta t \sum_{\sigma\in\mathscr{E}} \tau_\sigma (D_\sigma \Psi^{n+1})^2 \leq K_F(1+\lambda^2). \quad (24c)$$

We refer to [4] for the proof of this Theorem. The estimates are obtained successively: first, we establish the weak-BV inequality on N and P (24a); then, we deduce the $L^2(0,T,H^1)$ estimate on N and P and finally we conclude with the $L^2(0,T,H^1)$ estimate on Ψ.

The $L^2(0,T,H^1(\Omega))$-estimates on N, P (24b) and Ψ (24c) lead to compactness in space of the sequences of approximate solutions. The compactness in time is deduced from estimates on the time translates obtained by reusing the scheme. To prove the convergence of the numerical method, it remains to pass to the limit in the scheme and by this way prove that the limit of the sequence of approximate solutions is a weak solution to $(\mathscr{P}_\lambda)$. It can be done as in [11], but dealing with the Scharfetter-Gummel fluxes as in [3].

4.4 Some Numerical Experiments

We illustrate now the stability of the fully implicit Scharfetter-Gummel scheme for all nonnegative values of the rescaled Debye length λ. Therefore, we consider a one dimensional test case on $\Omega = (0,1)$. Initial data are constant $N_0(x) = P_0(x) = 0.5$, $\forall x \in (0,1)$. We consider quasi-neutral Dirichlet boundary conditions $N^D(0) = P^D(0) = 0.1$, $\Psi^D(0) = 0$ and $N^D(1) = P^D(1) = 0.9$, $\Psi^D(1) = 4$.

Since the exact solution to the problem $(\mathscr{P}_\lambda)$ is not available, we compute a reference solution on a uniform mesh made of $10240 = 20\times 2^9$ cells, with time step $\Delta t = 10^{-6}$, for different values of λ^2 in $[0,1]$. This reference solution is then used to compute the L^1 error for the variables N, P and Ψ. In order to prove the asymptotic preserving behaviour of the scheme, we compute L^1 errors at time $T = 0.1$ for different numbers of cells $\theta = 20\times 2^i$, $i \in \{0,...,8\}$, with different time steps Δt in $[10^{-5}, 10^{-2}]$ and various rescaled Debye length λ^2 in $[0,1]$. Figure 2 presents the L^1 error on the electron density and on the electrostatic potential as functions of Δt for different values of λ^2. It clearly shows the uniform behaviour in the limit $\lambda \to 0$ since the convergence rate is of order 1 for all variables even for small values of λ^2, including zero.

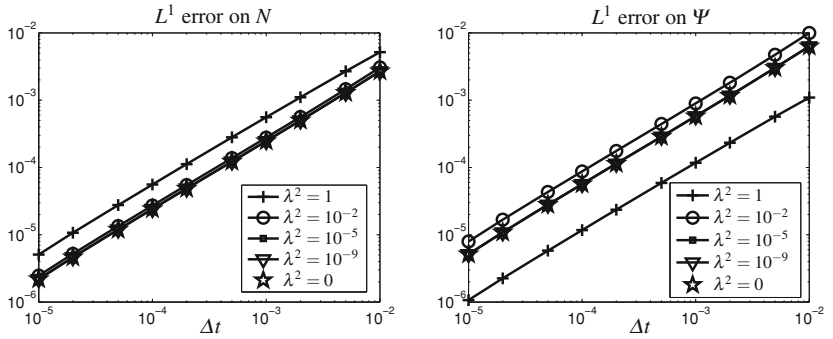


Fig. 2 Errors in L^1 norm as functions of Δt, for different values of λ^2

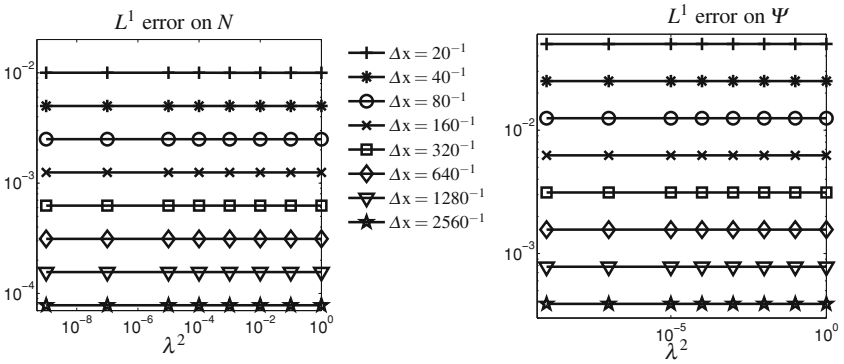


Fig. 3 Errors in L^1 norm as functions of λ^2, for different values of Δx

On Fig. 3, we plot the L^1 errors as functions of λ^2 for different values of the space step. We still observe the asymptotic preserving property of the scheme in the limit $\lambda \to 0$. Moreover, the errors are independent of λ^2.

Further numerical experiments (in 2D, with a non-vanishing doping profile,...) can be found in [4].

5 Conclusion

In this paper, we have first presented some results obtained with Jüngel and Schuchnigg on the long time behaviour of a classical scheme for nonlinear diffusion equations. We have obtained the exponential/polynomial decay of discrete zeroth-order relative entropies. The proof is based on an entropy method and on discrete functional inequalities. We refer to [10] for further results on first-order entropies (like the Fisher information).

On the drift-diffusion system, we have explained how the discrete counterpart of the entropy method can take part in the proof of convergence of a particular (but widely used) finite volume scheme. Particularly, it permits to establish that the considered scheme is asymptotic preserving at the quasi-neutral limit. We refer to the joint work with Bessemoulin-Chatard and Vignal [4] for the details of the proofs.

References

1. Arnold, A., Carrillo, J.A., Desvillettes, L., Dolbeault, J., Jüngel, A., Lederman, C., Markowich, P.A., Toscani, G., Villani, C.: Entropies and equilibria of many-particle systems: an essay on recent research. Monatsh. Math. **142**(1–2), 35–43 (2004)
2. Bataillon, C., Bouchon, F., Chainais-Hillairet, C., Fuhrmann, J., Hoarau, E., Touzani, R.: Numerical methods for the simulation of a corrosion model with moving oxide layer. J. Comput. Phys. **231**(18), 6213–6231 (2012)
3. Bessemoulin-Chatard, M.: A finite volume scheme for convection-diffusion equations with nonlinear diffusion derived from the scharfetter-gummel scheme. Numer. Math. **121**(4), 637–670 (2012)
4. Bessemoulin-Chatard, M., Chainais-Hillairet, C., Vignal, M.H.: Study of a finite volume scheme for the drift-diffusion system. asymptotic behavior in the quasi-neutral limit. submitted for publication. http://hal.archives-ouvertes.fr/hal-00801912
5. Carlen, E.A., Ulusoy, S.: An entropy dissipation-entropy estimate for a thin film type equation. Commun. Math. Sci. **3**(2), 171–178 (2005)
6. Carrillo, J.A., Desvillettes, L., Fellner, K.: Exponential decay towards equilibrium for the inhomogeneous aizenman-bak model. Comm. Math. Phys. **278**(2), 433–451 (2008)
7. Carrillo, J.A., Dolbeault, J., Gentil, I., Jüngel, A.: Entropy-energy inequalities and improved convergence rates for nonlinear parabolic equations. Discrete Contin. Dyn. Syst. Ser. B **6**(5), 1027–1050 (2006)
8. Carrillo, J.A., Jüngel, A., Tang, S.: Positive entropic schemes for a nonlinear fourth-order parabolic equation. Discrete Contin. Dyn. Syst. Ser. B **3**(1), 1–20 (2003)
9. Carrillo, J.A., Toscani, G.: Asymptotic l^1-decay of solutions of the porous medium equation to self-similarity. Indiana Univ. Math. J. **49**(1), 113–142 (2000)
10. Chainais-Hillairet, C., Jüngel, A., Schuchnigg, S.: Entropy-dissipative discretization of nonlinear diffusion equations and discrete beckner inequalities. submitted for publication. http://hal.archives-ouvertes.fr/hal-00924282
11. Chainais-Hillairet, C., Liu, J.G., Peng, Y.J.: Finite volume scheme for multi-dimensional drift-diffusion equations and convergence analysis. m2an. Math. Model. Numer. Anal. **37**(2), 319–338 (2003)
12. Chatard, M.: Asymptotic behavior of the Scharfetter-Gummel scheme for the drift-diffusion model. In: Finite volumes for complex applications. VI. Problems and perspectives. Volume 1, 2, Springer Proceedings Mathematics, vol. 4, pp. 235–243. Springer, Heidelberg (2011)
13. Desvillettes, L., Fellner, K.: Exponential decay toward equilibrium via entropy methods for reaction-diffusion equations. J. Math. Anal. Appl. **319**(1), 157–176 (2006)
14. Eymard, R., Gallouët, T., Ghilani, M., Herbin, R.: Error estimates for the approximate solutions of a nonlinear hyperbolic equation given by finite volume schemes. IMA J. Numer. Anal. **18**(4), 563–594 (1998)
15. Eymard, R., Gallouët, T., Herbin, R.: Finite volume methods. Handb. Numer. Anal. **7**, 713–1018 (2000)
16. Gajewski, H., Gärtner, K.: On the discretization of van roosbroeck's equations with magnetic field. Z. Angew. Math. Mech. **76**(5), 247–264 (1996)

17. Gasser, I.: The initial time layer problem and the quasineutral limit in a nonlinear drift diffusion model for semiconductors. NoDEA Nonlinear Diff. Equat. Appl. **8**(3), 237–249 (2001)
18. Gasser, I., Levermore, C.D., Markowich, P.A., Schmeiser, C.: The initial time layer problem and the quasineutral limit in the semiconductor drift-diffusion model. Eur. J. Appl. Math. **12**(4), 497–512 (2001)
19. Glitzky, A.: Exponential decay of the free energy for discretized electro-reaction-diffusion systems. Nonlinearity **21**(9), 1989–2009 (2008)
20. Glitzky, A.: Uniform exponential decay of the free energy for voronoi finite volume discretized reaction-diffusion systems. Math. Nachr. **284**(17–18), 2159–2174 (2011)
21. Glitzky, A., Gärtner, K.: Energy estimates for continuous and discretized electro-reaction-diffusion systems. Nonlinear Anal. **70**(2), 788–805 (2009)
22. Glitzky, A., Hünlich, R.: Global estimates and asymptotics for electro-reaction-diffusion systems in heterostructures. Appl. Anal. **66**(3–4), 205–226 (1997)
23. Grün, G., Rumpf, M.: Nonnegativity preserving convergent schemes for the thin film equation. Numer. Math. **87**(1), 113–152 (2000)
24. Jüngel, A.: Qualitative behavior of solutions of a degenerate nonlinear drift-diffusion model for semiconductors. Math. Models Methods Appl. Sci. **5**(4), 497–518 (1995)
25. Jüngel, A., Peng, Y.J.: A hierarchy of hydrodynamic models for plasmas. quasi-neutral limits in the drift-diffusion equations. Asymptot. Anal. **28**(1), 49–73 (2001)
26. Scharfetter, D., Gummel, H.: Large-signal analysis of a silicon read diode oscillator. Electron Devices, IEEE Transactions on **16**(1), 64–77 (1969)
27. Van Roosbroeck, W.: Theory of the flow of electrons and holes in germanium and other semiconductors. Bell Syst. Tech. J. **29**, 560–607 (1950)

Interpolated Pressure Laws in Two-Fluid Simulations and Hyperbolicity

Philippe Helluy and Jonathan Jung

Abstract We consider a two-fluid compressible flow. Each fluid obeys a stiffened gas pressure law. The continuous model is well defined without considering mixture regions. However, for numerical applications it is often necessary to consider artificial mixtures, because the two-fluid interface is diffused by the numerical scheme. We show that classic pressure law interpolations lead to a non-convex hyperbolicity domain and failure of well-known numerical schemes. We propose a physically relevant pressure law interpolation construction and show that it leads to a necessary modification of the pure phase pressure laws. We also propose a numerical scheme that permits to approximate the stiffened gas model without artificial mixture.

1 Introduction

The numerical simulation of compressible two-fluid flows with Eulerian finite volume approximation has been widely studied. We refer for instance to [17, 20, 22] and included references. The Eulerian approach is very appealing compared to Lagrangian front tracking methods because it generally leads to much simpler algorithms. However, one has to circumvent the pressure oscillations that appear at the two-fluid interface with standard conservative schemes. The lack of accuracy of the Godunov scheme at contact waves is a well-known issue. It is not only observed in two-fluid simulations but also in one-fluid simulations when the single fluid satisfies a complex pressure law [8].

P. Helluy (✉) · J. Jung
IRMA, University of Strasbourg and Inria Tonus, Strasbourg, France
e-mail: helluy@math.unistra.fr

J. Jung
e-mail: jonathan.jung@unistra.fr

J. Fuhrmann et al. (eds.), *Finite Volumes for Complex Applications VII - Methods and Theoretical Aspects,* Springer Proceedings in Mathematics & Statistics 77,
DOI: 10.1007/978-3-319-05684-5_3,

Another less known aspect of classic two-fluid models is that their hyperbolicity domain is generally not convex. Thus in some cases, the Godunov scheme is unstable and fails after only one time iteration.

In this paper we consider the flow of a gas and a liquid modeled by two stiffened gas equations of state. The pressure law is initially only defined in the pure phases: the mass fraction of gas φ can take only two values $\varphi = 0$ or $\varphi = 1$.

For numerical reasons, the pressure law is often interpolated in an artificial mixture region $0 < \varphi < 1$. We show that a naive numerical interpolation always leads to the non-convexity of the hyperbolicity domain and thus to the instability of the Godunov scheme.

We then propose two alternative cures to this issue:

1. The construction of a mixture pressure law based on physical and thermodynamical arguments. We are then able to recover a convex hyperbolicity domain. But we also prove that it is necessary to modify the pressure law of the liquid phase, which cannot be a simple stiffened gas anymore.
2. For keeping the simplicity of the stiffened gas model, we propose another scheme, the Random Interface Solver (RIS) [16]. This scheme does not diffuse the mass fraction profile and allows stable computations of the two-fluid model.

Finally we present some two-dimensional numerical results obtained with the RIS scheme.

2 Two-Fluid Flows and Godunov Scheme

We consider a two-fluid model (air and liquid water) written as a first order system of conservation laws

$$\partial_t W + \partial_x F(W) = 0. \tag{1}$$

The space variable is x. The time variable is t. We use the notations $\partial_t = \partial/\partial t$, $\partial_x = \partial/\partial x$. The conservative unknowns $(x, t) \to W(x, t)$ are

$$W = (\rho, \rho u, \rho e, \rho\varphi)^T, \tag{2}$$

with the density ρ, velocity u, total energy e and mass fraction of gas φ. The total energy e is related to the internal energy ε by

$$e = \varepsilon + \frac{1}{2}u^2. \tag{3}$$

The flux of the conservative system is given by

$$F(W) = (\rho u, \rho u^2 + p, (\rho e + p)u, \rho\varphi u)^T. \tag{4}$$

The pressure law is of the form

$$p = P(\rho, \varepsilon, \varphi). \tag{5}$$

If at the initial time $t = 0$ the mass fraction φ takes only two values 0 (pure liquid phase) and 1 (pure gas phase) then it is also true at any later time. This property implies that theoretically it is only necessary to provide the pressure laws $P(\rho, \varepsilon, 0)$ for the liquid and $P(\rho, \varepsilon, 1)$ for the gas. A classic choice is the stiffened gas pressure law

$$P(\rho, \varepsilon, \varphi) = (\gamma(\varphi) - 1)\rho\varepsilon - \gamma(\varphi)\pi(\varphi), \tag{6}$$

$$\gamma(1) = \gamma_1 > 1, \quad \pi(1) = \pi_1 = 0, \quad \gamma(0) = \gamma_2 > 1, \quad \pi(0) = \pi_2 > 0. \tag{7}$$

The constants γ_1, γ_2 and π_2 are obtained from physical measurements. For instance, for air and water, we can take [1]

$$\gamma_1 = 1.4, \quad \gamma_2 = 3, \quad \pi_2 = 8533 \times 10^5 \text{ Pa}.$$

In the following, we consider the properties of system (1)–(6). In short, we call it the two-fluid model.

In practice, it is difficult to impose $\varphi = 0$ or $\varphi = 1$ in the numerical approximation. A widely used possibility (see for instance [2, 11, 17, 20–22]) is to interpolate the pressure laws parameter $\gamma(\varphi)$ and $\pi(\varphi)$ for $\varphi \in]0, 1[$.

For the stiffened gas pressure law, the sound speed is given by

$$c = \sqrt{\frac{\gamma(\varphi)(p + \pi(\varphi))}{\rho}},$$

we thus obtain that the system (1)–(6) is hyperbolic if W is in the hyperbolicity domain

$$\Omega = \{(\rho, \rho u, \rho e, \rho\varphi), \rho \geq 0, \varphi \in [0, 1], p + \pi(\varphi) \geq 0\}.$$

The Riemann problem for the two-fluid system consists in solving the following initial value problem

$$\partial_t W + \partial_x F(W) = 0,$$

$$W(x, 0) = \begin{cases} W_L & \text{if } x < 0, \\ W_R & \text{otherwise.} \end{cases}$$

The left and right constant states W_L, W_R are taken into the hyperbolicity domain Ω. The solution is self-similar, denoted by

$$R(W_L, W_R, \frac{x}{t}) = W(x, t).$$

It is made of shock waves, rarefaction waves and contact waves, separated by constant states. It is well known that the solution of the Riemann problem is generally not unique. In order to reduce the set of solutions, we can for instance consider only shock waves that satisfy the Lax characteristic criterion. Below, we will also discuss the Lax entropy criterion.

An essential feature of the two-fluid system is that the Riemann problem admits a unique global solution that satisfies the Lax characteristic criterion whenever the left and right initial states are in the hyperbolicity domain Ω. This global solution is constructed from standard wave parameterization [18]. It is sometimes necessary to introduce vacuum state in the gas phase (see [2] for details). The solution is not only theoretical. It can also be computed almost analytically in an efficient way.

The Riemann problem being uniquely solvable, it is then tempting to apply the Godunov scheme to the two-fluid model with arbitrary initial data.

We consider a sequence of time t_n, $n \in \mathbb{N}$ such that the time step $\tau_n = t_{n+1} - t_n > 0$. We consider also a space step h. We define the cell centers by $x_i = ih$. The cell C_i is the interval $]x_{i-1/2}, x_{i+1/2}[$. We consider an approximation

$$W_i^n \simeq \frac{1}{h} \int_{x \in C_i} W(x, t_n) dx.$$

A time step of the Godunov scheme is made of two stages:

- Step1: Exact resolution starting from approximated cell averages

$$\partial_t V + \partial_x F(V) = 0,$$

$$V(x, 0) = W_i^n, \quad x \in C_i.$$

- Step 2: Averaging of the exact solution

$$W_i^{n+1} = \frac{1}{h} \int_{C_i} V(x, \tau_n) dx. \tag{8}$$

The time marching scheme also admits a finite volume formulation

$$W_i^{n+1} = W_i^n - \frac{h}{\tau_n} \left(F_{i+1/2}^n - F_{i-1/2}^n \right).$$

The numerical fluxes are computed from exact solutions of the Riemann problem

$$F_{i+1/2}^n = F(R(W_i^n, W_{i+1}^n, 0)).$$

The time step satisfies a CFL condition

$$\tau_n \leq \frac{h}{2\lambda_n^{\max}},$$

where $\lambda_n^{\max}$ is an upper bound of all the wave speeds in the solutions of the interface Riemann problems at time t_n.

In our application, Step 1 of the Godunov scheme is not a problem if all the W_i^n are in Ω because the Riemann problem admits a unique physically relevant solution $W(x, t) \in \Omega$. Surprisingly, Step 2 is much more problematic when $\pi_1 \neq \pi_2$, because of the following result [16]:

Theorem 1 *Consider the two-fluid system (1)–(6) and suppose a continuous interpolation of the pressure law parameters $\varphi \to \gamma(\varphi)$, $\varphi \to \pi(\varphi)$ for $\varphi \in [0, 1]$ satisfying (7). Then, the hyperbolicity set Ω is never convex.*

The non-convexity of Ω is a big issue, because even if we have $V(x, \tau_n) \in \Omega$ in the averaging formula (8), we cannot conclude that $W_i^{n+1} \in \Omega$. In practice it is possible to construct initial data for which the Godunov scheme fails after only one iteration [16, 21]. In conclusion, the Godunov scheme applied to the two-fluid model with $\pi_1 \neq \pi_2$ is generally unstable.

Remark 1 The numerical resolution of the two-fluid model with stiffened gas pressure law has been extensively studied. When the two fluids are perfect gases (when $\pi_1 = \pi_2 = 0$) the hyperbolicity set Ω is convex. In this case, the Godunov scheme is stable. But it often gives very inaccurate results in contact waves. In the literature, this behavior is called the pressure oscillation phenomenon. Let us emphasize that the instability of the Godunov scheme for $\pi_2 \neq \pi_1$ and the pressure oscillations in contact waves are two different and independent shortcomings of the Godunov scheme applied to two-fluid flows.

Remark 2 A very popular method for avoiding pressure oscillations has been proposed by Abgrall and Saurel in [22]. The method relies on a non-conservative numerical resolution of the transport equation

$$\partial_t \varphi + u \partial_x \varphi = 0,$$

and a special interpolation of the pressure law coefficients $\gamma(\varphi)$, $\pi(\varphi)$ that ensures that the pressure and the velocity remain numerically constant in contact waves. Even if this trick improves the Godunov scheme accuracy, it does not improve the stability. Indeed, it is possible to show that the hyperbolicity set, expressed in the non conservative variables

$$\Omega' = \{(\rho, \rho u, \rho e, \varphi), \rho \geq 0, \varphi \in [0, 1], p + \pi(\varphi) \geq 0\}$$

is also generally not convex. We can also exhibit physical initial state that leads to the failure of the Abgrall-Saurel scheme after only one iteration [16].

Remark 3 A Lax entropy $W \to U(W) \in \mathbb{R} \cup \{+\infty\}$ is a strictly convex function of W associated to an entropy flux $W \to G(W)$ such that the smooth solutions of the two-fluid model satisfy the additional conservation law

$$\partial_t U(W) + \partial_x G(W) = 0.$$

If a system of conservation laws admits a Lax entropy, then from Mock's theorem [19], we know that the system is hyperbolic on the domain of U and thus

$$\Omega = \mathrm{Dom}U = \left\{ W \in \mathbb{R}^4, U(W) < +\infty \right\}.$$

Because the domain of a convex function is a convex set, we deduce that the two-fluid model cannot possess a global Lax entropy.

In conclusion of this section, we have two alternatives for approximating the two-fluid model in a robust and precise way:

1. We can abandon the stiffened gas pressure law (6) and construct another pressure law that ensures the convexity of the hyperbolicity set.
2. If we keep the stiffened gas pressure law we have to construct a scheme that is stable with respect to non-convex hyperbolicity set.

In Sect. 3, we investigate the first possibility, while in Sect. 4 we consider the second.

3 Convex Mixture

In this section, we consider a mixture of a perfect gas and a liquid satisfying the stiffened gas pressure law. From physical entropy arguments, we construct a mixture pressure law. This pressure law is naturally associated to a convex Lax entropy of the two-fluid system. Mock's theorem then ensures the convexity of the hyperbolicity set. The construction is split into several steps.

3.1 Construction of an Extensive Mixture Entropy

The pressure law is constructed as follows. First, we consider two fluids indexed by $i = 1$ and $i = 2$ of mass M_i, energy E_i, occupying a volume V_i. We introduce the specific heat χ_i of fluid i. Then, for $M_i > 0$, $V_i > 0$ and $E_i > \pi_i V_i$, the entropy function of fluid i is

$$S_i(V_i, E_i, M_i) = -\chi_i \gamma_i M_i \ln M_i + \chi_i(\gamma_i - 1) M_i \ln V_i + \chi_i M_i \ln(E_i - \pi_i V_i). \quad (9)$$

For completely rigorous proofs, we have to define the entropies for all $(V_i, E_i, M_i) \in \mathbb{R}^3$. We thus also set

$$S_i(V_i, E_i, 0) = 0, \quad V_i \geq 0, \quad E_i \geq \pi_i V_i, \quad (10)$$

and

$$S_i(V_i, E_i, 0) = -\infty \tag{11}$$

in all the other cases. With this definition, S_i are concave upper semicontinous (in short: usc) functions and

$$\begin{aligned} \mathrm{Dom}S_i = &\{(V_i, E_i, M_i), V_i > 0, E_i > \pi_i V_i, M_i > 0\} \\ &\cup \{(V_i, E_i, 0), V_i \geq 0, E_i \geq \pi_i V_i\}. \end{aligned} \tag{12}$$

$\mathrm{Dom}S_i$ are convex cones. In addition, the entropies S_i are Positively Homogeneous function of degree 1 (in short PH1)

$$\forall \lambda \geq 0, \forall W \in \mathbb{R}^3, S(\lambda W) = \lambda S(W).$$

We then define the entropy of the immiscible mixture by

$$S(V, E, M, M_1) = \sup_{V_1, E_1} \left(S_1(V_1, E_1, M_1) + S_2(V - V_1, E - E_1, M - M_1)\right). \tag{13}$$

This formula is physically justified by the fact that the entropy is an additive quantity and by the second principle of thermodynamics: the mixture of the two fluids evolves until it reaches a maximum of entropy. For more details, we refer to [1, 14–16] and included references. Let us also observe that we do not optimize the mixture entropy with respect to M_1 because we do not consider phase transition between the liquid and the gas.

From standard convex analysis, it is possible to prove the following result [16]:

Theorem 2 *Let S be defined by (13), where S_1 and S_2 satisfy (9)–(11). Then S is a PH1 concave and usc function. Its domain is a convex cone given by*

$$\begin{aligned} Dom S = &\{(V, E, M, M_1), V > 0, E > 0, M \geq M_1 \geq 0\} \\ &\cup \{(V, E, 0, 0), V \geq 0, E \geq 0\}. \end{aligned} \tag{14}$$

3.2 Intensive Mixture Entropy and Pressure law

From the extensive PH1 entropy, we can go back to intensive variables. We have the following relations

$$\rho = \frac{M}{V}, \tau = \frac{1}{\rho} = \frac{V}{M}, \varepsilon = \frac{E}{M}, \varphi = \frac{M_1}{M}, s = \frac{S}{M}, \sigma = \frac{S}{V}.$$

We then define the intensive specific entropy

$$s(\tau, \varepsilon, \varphi) = S(\tau, \varepsilon, 1, \varphi).$$

From Theorem 2, the specific entropy is a concave function. We define the pressure p, temperature T and chemical potential λ of the mixture by

$$T = \frac{1}{\partial_\varepsilon s}, \quad p = T\partial_\tau s, \quad \lambda = T\partial_\varphi s, \tag{15}$$

in such a way that

$$Tds = d\varepsilon + pd\tau + \lambda d\varphi.$$

We can also consider the volumic entropy

$$\sigma(\rho, \rho\varepsilon, \rho\varphi) = S(1, \rho\varepsilon, \rho, \rho\varphi) = \rho s(\frac{1}{\rho}, \frac{\rho\varepsilon}{\rho}, \frac{\rho\varphi}{\rho}).$$

In the same way, the volumic entropy is a concave function of $(\rho, \rho\varepsilon, \rho\varphi)$.

It is then standard [10, 12, 16] to deduce that the quantity

$$U(W) = -\sigma(\rho, \rho\varepsilon, \rho\varphi), \quad \text{with } W = (\rho, \rho u, \rho\varepsilon + 1/2\rho u^2, \rho\varphi)^T,$$

is a convex Lax entropy associated to the entropy flux

$$G(W) = uU(W).$$

Our whole construction ensures that the two-fluid system with the pressure law given by (15) is necessarily hyperbolic on a convex domain and that this convex domain is simply the domain of the Lax entropy U.

3.3 Explicit Pressure Law

It is interesting to perform the full computations in order to see how the resulting pressure $P(\rho, \varepsilon, \varphi)$ is different from the interpolated pressure law (6). The computations are not very difficult but a little bit lengthy. They are rigorously detailed in [16].

We give the final result. Let us just mention that the same formula can be obtained by formally assuming pressure and temperature equilibrium between the two phases

$$p_1 = p_2, \quad T_1 = T_2.$$

The temperatures T_1 and T_2 of the two stiffened gases are given by the relation

$$\chi_i T_i = 1/\partial_{\varepsilon_i} s_i(\tau_i, \varepsilon_i) = \varepsilon_i - \pi_i \tau_i.$$

Of course, such equilibrium assumption has no meaning when $\varphi = 0$ or $\varphi = 1$ because in this case only one phase is present in the mixture. The entropy optimization procedure is more rigorous and allows handling the cases $\varphi = 0$ or $\varphi = 1$.

We take a density $\rho > 0$, an internal energy $\varepsilon > 0$. We define the heat capacity of the mixture by

$$\chi = \chi(\varphi) = \chi_1 \varphi + (1 - \varphi)\chi_2.$$

We also define the energy fraction of the mixture by

$$\zeta = \zeta(\varphi) = \frac{\chi_1 \varphi}{\chi}.$$

The mixture polytropic parameter is then

$$\gamma = \gamma(\varphi) = \zeta \gamma_1 + (1 - \zeta)\gamma_2.$$

We also consider the following quantities

$$\delta = -\gamma_2 \pi_2, \quad r = (\delta + (\gamma - 1)\rho e)^2 - 4\delta(\gamma_1 - 1)\zeta \rho e > 0.$$

The volume fraction of gas is then given by

$$\alpha = \alpha(\varphi) = \frac{\delta + (\gamma - 1)\rho e - \sqrt{r}}{2\delta}.$$

1. If $0 < \varphi < 1$ then

$$P(\rho, \varepsilon, \varphi) = (\gamma - 1)\rho\varepsilon - \gamma(1 - \alpha)\pi_2.$$

2. If $\varphi = 1$ then

$$P(\rho, \varepsilon, \varphi) = (\gamma_1 - 1)\rho\varepsilon.$$

3. If $\varphi = 0$ then

$$P(\rho, \varepsilon, 0) = \max(0, (\gamma_2 - 1)\rho\varepsilon - \gamma_2 \pi_2). \tag{16}$$

The main result of this analysis is that even if the mass of gas vanishes ($\varphi = 0$) the remaining liquid does not always obey a pure stiffened gas law. When the energy is small enough, the liquid pressure vanishes. Intuitively, this means that the liquid undergoes a cavitation phenomenon. The liquid pressure cannot be negative anymore, while it was possible in the pressure law (6).

The pressure law obtained with the entropy optimization procedure ensures a convex hyperbolicity domain. It thus ensures the stability of the Godunov scheme. However, the pressure law is more complex than the stiffened gas law. For instance,

in a pure liquid region, the two-fluid model can degenerate to a pressureless Euler system. This system is known to lead to theoretical and numerical difficulties. In addition, we have verified in numerical experiments that the pressure oscillation phenomenon is still present at contact waves. Therefore, we will also propose in the next section a practical numerical method for solving directly the two-fluid model with a non-convex hyperbolicity domain.

4 A New Random Interface Solver

In this section, we return to the simple stiffened gas pressure law (6). We define the hyperbolicity sets in the pure phases

$$\Omega_0 = \Omega \cap \{\varphi = 0\}, \quad \Omega_1 = \Omega \cap \{\varphi = 1\}.$$

These two sets are convex. We consider a numerical initial condition in the pure phases

$$\forall i, \quad W_i^0 \in \Omega_0 \cup \Omega_1.$$

We show how to construct a scheme, the Random Interface Solver (RIS) that satisfies the stability condition

$$\forall i, \quad \forall n \geq 0, \quad W_i^n \in \Omega_0 \cup \Omega_1. \tag{17}$$

From the literature and the considerations above we know two things:

1. The new scheme cannot be exactly conservative at each time step;
2. If we average the mass fraction on the cells of the initial mesh, we will introduce a numerical diffusion and certainly pressure oscillations at the interface.

In order to avoid these two pitfalls we will

1. use a random sampling strategy at the interface. It allows avoiding the diffusion of the mass fraction profile. It is not perfectly conservative, but we can prove that it is statistically conservative on long times;
2. before the random sampling, we use a Lagrangian conservative finite volume scheme at the interface. In this Lagrangian step, the mass fraction is not diffused either.

We now enter into the details of the RIS scheme. Each time step of the scheme is made of two stages: an Arbitrary Lagrangian Eulerian (ALE) step and a Projection step. The idea to combine the Glimm's scheme approach [9] and a Lagrangian scheme approach was first proposed by Chalons and Goatin in [5]. See also [3, 4, 13].

4.1 ALE Stage

In the first stage, we allow the cell boundary $x_{i+1/2}$ to move at velocity $\xi^n_{i+1/2}$. At the end of the first stage, the cell boundary is

$$x^{n+1,-}_{i+1/2} = x_{i+1/2} + \tau_n \xi^n_{i+1/2}.$$

Integrating the conservation law (1) on the moving cells, we obtain the following finite volume approximation

$$h^{n+1,-}_i W^{n+1,-}_i - h W^n_i + \tau_n (F^n_{i+1/2} - F^n_{i-1/2}) = 0.$$

The new size of cell i is given by

$$h^{n+1,-}_i = x^{n+1,-}_{i+1/2} - x^{n+1,-}_{i-1/2} = h + \tau_n (\xi^n_{i+1/2} - \xi^n_{i-1/2}).$$

The numerical flux is now of the ALE form

$$F^n_{i+1/2} = F(W^n_{i+1/2}) - \xi^n_{i+1/2} W^n_{i+1/2}.$$

The intermediate state $W^n_{i+1/2}$ is obtained by the resolution of a Riemann problem

$$W^n_{i+1/2} = R(W^n_i, W^n_{i+1}, \xi^n_{i+1/2}).$$

In practice, R can also be an approximate Riemann solver [16].

Finally, the interface velocity is defined by

$$\xi^n_{i+1/2} = \begin{cases} u^n_{i+1/2} & \text{if} \quad (\varphi^n_i - 1/2)(\varphi^n_{i+1} - 1/2) < 0, \\ 0 & \text{else.} \end{cases} \tag{18}$$

The numerical flux is thus a classic Godunov flux in the pure fluid. It is a Lagrangian numerical flux at the two-fluid interface.

4.2 Projection Step

The second stage of the time step is needed for returning to the initial mesh. We have to compute on the cells C_i of the initial mesh the averages of $W^{n+1,-}_i$, defined on the moved cells $C^{n+1,-}_i =]x^{n+1,-}_{i-1/2}, x^{n+1,-}_{i+1/2}[$. Instead of a standard integral averaging method, we rather consider a random sampling averaging process. We consider a pseudo random sequence $\omega_n \in [0, 1[$ and we perform the following sampling

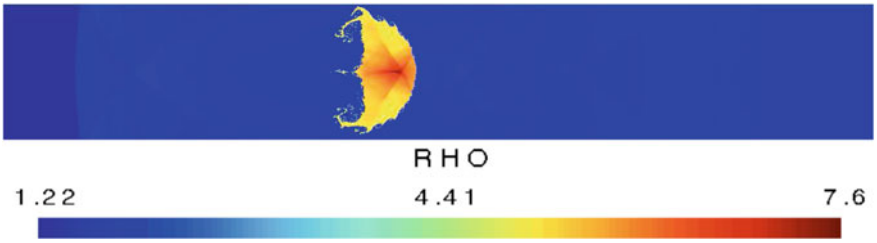


Fig. 1 Shock-droplet simulation. Density plot. Full view of the computational domain

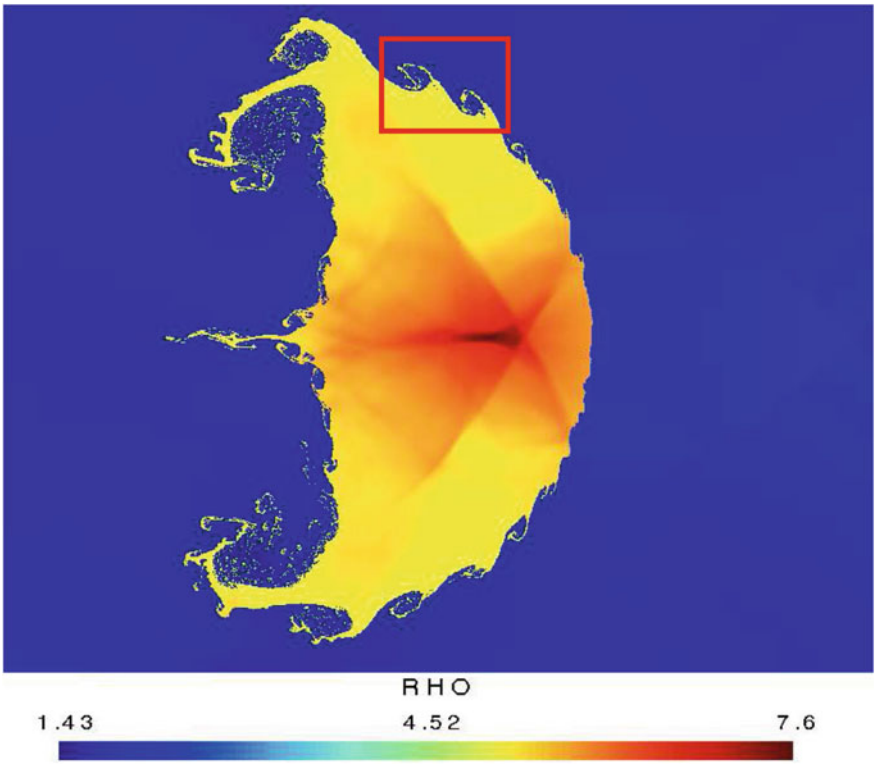


Fig. 2 Shock-droplet simulation. Density plot. Zoom on the droplet

$$W_i^{n+1} = \begin{cases} W_{i-1}^{n+1,-}, & \text{if } \omega_n < \frac{\xi_{i-1/2}^n \tau_n}{h}, \\ W_i^{n+1,-}, & \text{if } \frac{\xi_{i-1/2}^n \tau_n}{h} \le \omega_n \le 1 + \frac{\xi_{i+1/2}^n \tau_n}{h}, \\ W_{i+1}^{n+1,-}, & \text{if } \omega_n > 1 + \frac{\xi_{i+1/2}^n \tau_n}{h}. \end{cases} \tag{19}$$

A good choice for the pseudo-random sequence ω_n is the (k_1, k_2) van der Corput sequence. In practice, we consider the $(5, 3)$ van der Corput sequence.

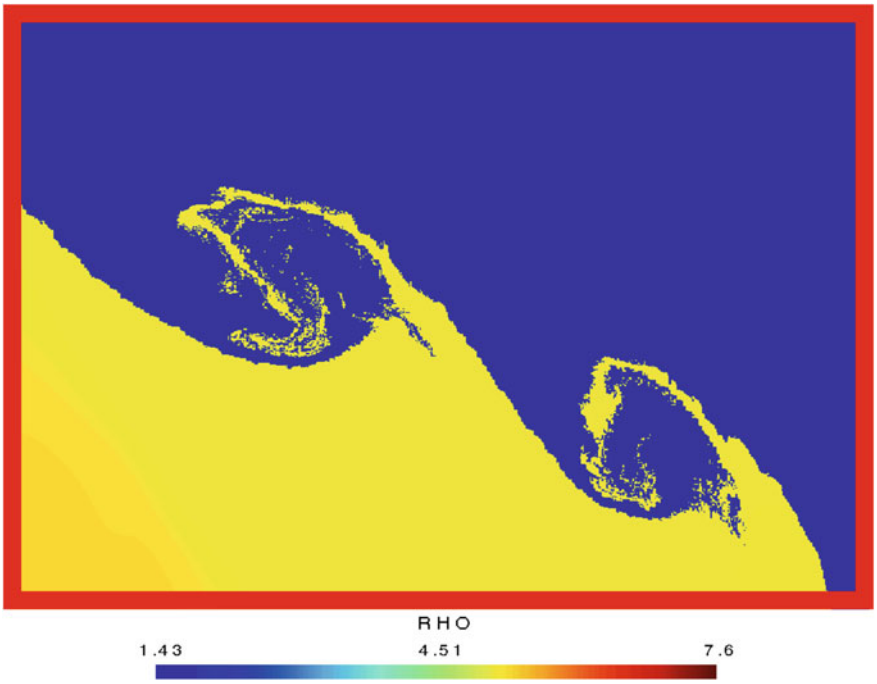


Fig. 3 Shock-droplet simulation. Density plot. Second zoom on the droplet

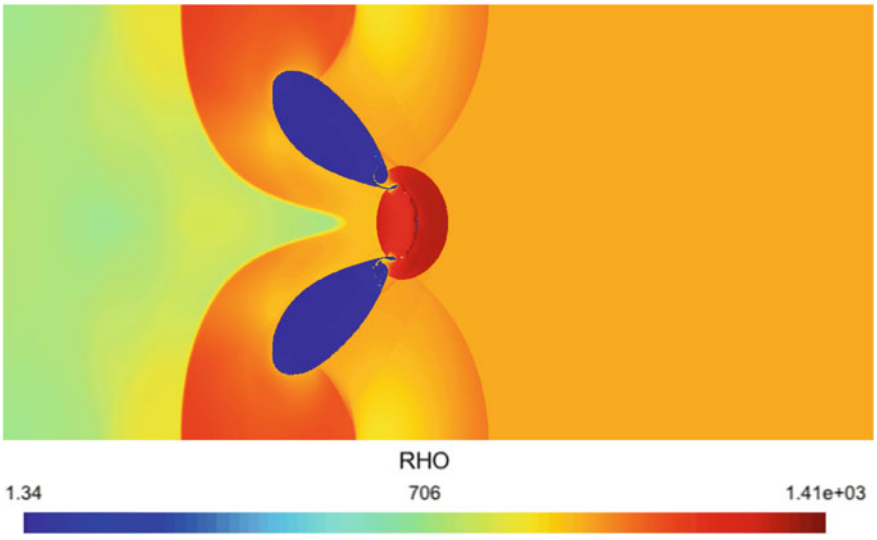


Fig. 4 Shock-bubble simulation. Density plot

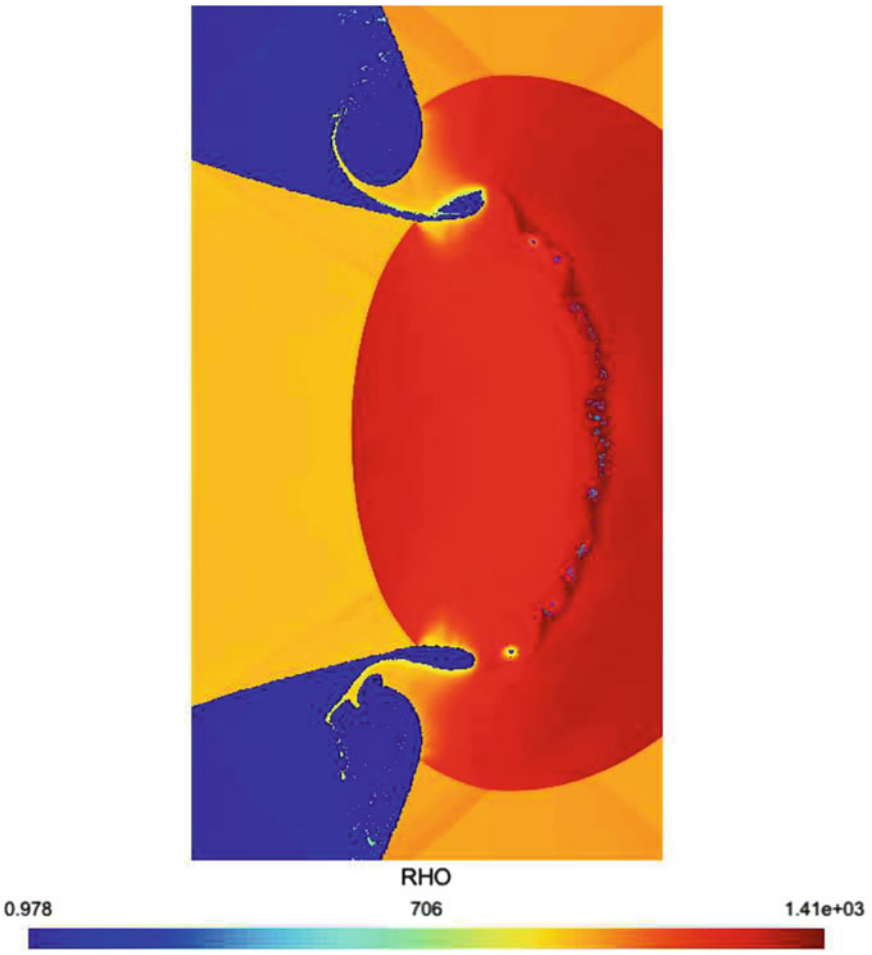


Fig. 5 Shock-bubble simulation. First zoom. Density plot

The resulting scheme has the following properties [16]:

- it is stable in the sense of (17);
- it is conservative in a statistical sense;
- it is entropy dissipative in a statistical sense;
- it does not produce spurious oscillations at the two-fluid interface.

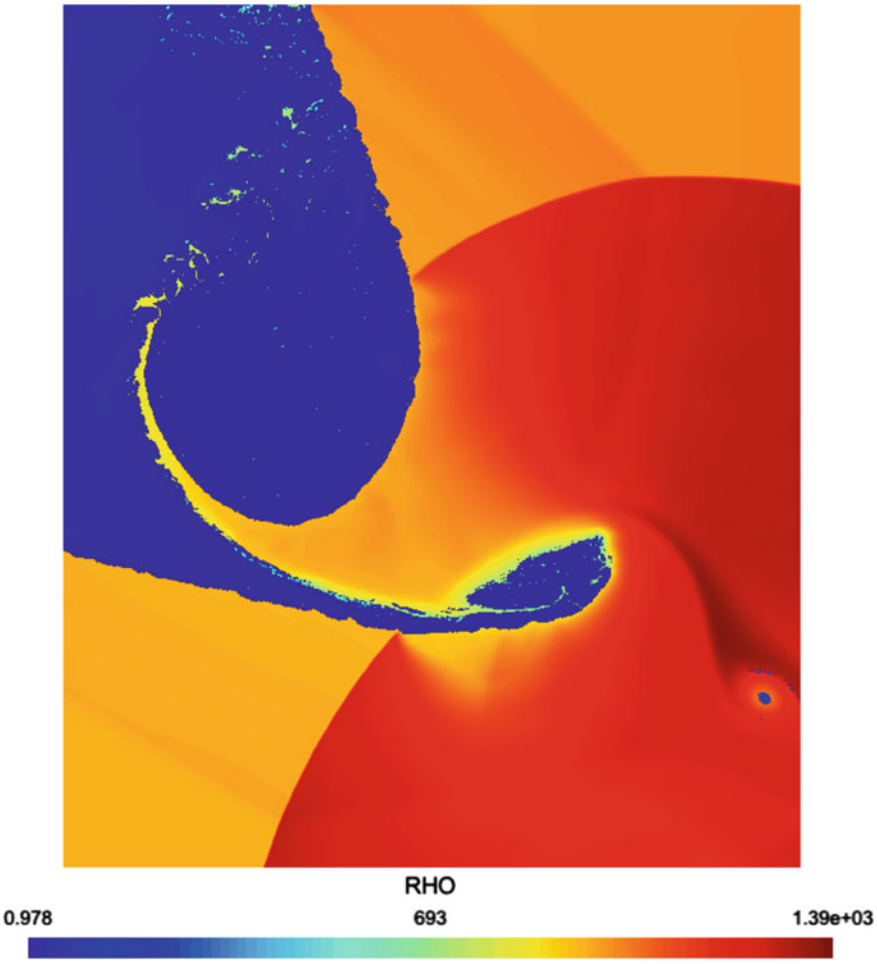


Fig. 6 Shock-bubble simulation. Second zoom. Density plot

5 Numerical Results

We can extend the scheme to higher dimensions with dimensional splitting (for more details we refer to [13]). It is remarkable that the same random number can be used for one time step in the x and y directions. It is also remarkable that despite dimensional splitting, the two-dimensional scheme converges towards the right solution without oscillation. Indeed, since the work of Colella [6], it was generally admitted that applying the dimensional splitting procedure to the Glimm's scheme leads to poor numerical results.

We present in Fig. 1 the results of a two-dimensional shock-droplet simulation. The initial droplet is a disk. A shock-wave coming from the right of the computational

domain interacts with the droplet. The computations have been realized thanks to an OpenCL/MPI implementation of the two-dimensional RIS scheme. For this test case, we use a cluster of four AMD Radeon HD 7970 GPU. The detailed description of the test case is given in [13]. We display the droplet after the interaction. We observe that we are able to capture a sharp interface. The numerical noise is moderate, despite the random nature of the scheme. Because we use a very fine mesh with $20,000 \times 5,000$ cells, we are able to zoom on small Kelvin-Helmholtz vortices (see Figs. 2 and 3).

We present in Fig. 4 the results of a two-dimensional shock-bubble simulation. The initial bubble is a disk. A shock wave coming from the left of the computational domain interacts with the bubble. We display the bubble after that it has been split by the shock wave. The computations have been realized thanks to an OpenCL/MPI implementation of the two-dimensional RIS scheme. For this test case, we use a cluster of ten NVIDIA K20 GPU. The detailed description of the test case is given in [13]. We use a very fine mesh with $40,000 \times 20,000$ cells. As in the previous test case, we can zoom in order to observe small details of the split region (see Figs. 5 and 6).

6 Conclusion

We have shown that a widely used two-fluid liquid-gas model has a non-convex hyperbolicity domain. This non-convexity can lead to the failure of the Godunov scheme after only one time step. This is true even if the continuous model admits a perfectly well defined solution that satisfies the Lax characteristic criterion.

We have proposed a modified equation of state for recovering a convex hyperbolicity domain. The resulting pressure law is more complicated and is not a stiffened gas equation of state anymore in the pure liquid, when the gas mass fraction $\varphi = 0$.

For keeping the simplicity of the stiffened gas equation, we have thus constructed a finite volume scheme, the RIS scheme, which avoids the numerical diffusion of the mass fraction. The RIS scheme is based on a Lagrangian finite volume approach coupled with a random sampling at the interface.

Let us conclude that the construction of a conservative finite volume scheme that would give accurate results at contact waves for one-fluid or two-fluid general pressure laws is still an open question.

References

1. Barberon, T., Helluy, P.: Finite volume simulation of cavitating flows. Computers & Fluids **34**, 832–858 (2005) NULL
2. Barberon, T., Helluy, P., Rouy, S.: Practical computation of axisymmetrical multifluid flows. Int. J. Finite **1**(1), 1–34 (2004)
3. C. Chalons, F. Coquel. Capturing infinitely sharp discrete shock profiles with the Godunov scheme. Hyperbolic problems: theory, numerics, applications, 363–370, Springer, Berlin, (2008).

4. C. Chalons, F. Coquel. Computing material fronts with a Lagrange-Projection approach. http://arxiv.org/abs/1012.4561 (2010)
5. Chalons, C., Goatin, P.: Transport-equilibrium schemes for computing contact discontinuities in traffic flow modeling. Commun. Math. Sci. **5**(3), 533–551 (2007)
6. P. Colella. Glimm's Method For Gas Dynamics. SIAM, J. Sci. Stat. Comput., 3(1) (1982).
7. J.-P. Croisille. Contribution à l'étude théorique et à l'approximation par éléments finis du système hyperbolique de la dynamique des gaz multidimensionnelle et multiespèces. PhD thesis, Université Paris VI (1990). NULL.
8. Gallouët, T., Hérard, J.-M., Seguin, N.: A hybrid scheme to compute contact discontinuities in one-dimensional euler systems. m2an. Math. Model. Numer. Anal. **36**(6), 1133–1159 (2003)
9. Glimm, J.: Solutions in the large for nonlinear hyperbolic systems of equations. Comm. Pure Appl. Math. **18**, 697–715 (1965)
10. E. Godlewski, P.-A. Raviart. Numerical approximation of hyperbolic systems of conservation laws. Applied Mathematical Sciences, 118, Springer-Verlag, New York (1996).
11. Golay, F., Helluy, P.: Numerical schemes for low mach wave breaking. Int. J. Comput. Fluid Dyn. **21**(2), 69–86 (2007)
12. Harten, A., Lax, P.D., Levermore, C.D., Morokoff, W.J.: Convex entropies and hyperbolicity for general euler equations. SIAM J. Numer. Anal. **35**(6), 2117–2127 (1998)
13. Helluy, P., Jung, J.: Opencl simulations of two-fluid compressible flows with a random choice method. Int. J. Finite Volumes **10**, 1–38 (2013)
14. Helluy, P., Mathis, H.: Pressure laws and fast legendre transform. Math. Models Methods Appl. Sci. **21**(4), 745–775 (2011)
15. Helluy, P., Seguin, N.: Relaxation models of phase transition flows. m2an. Math. Model. Numer. Anal. **40**(2), 331–352 (2006)
16. J. Jung. Schémas numériques adaptés aux accélérateurs multicoeurs pour les écoulements bifluides. PhD thesis. University of Strasbourg (2013).
17. Karni, S.: Multicomponent flow calculations by a consistent primitive algorithm. Journal of Computational Physics **47**, 1115–1145 (1994)
18. Lax, P.D.: Hyperbolic systems of conservation laws. II. Comm. Pure Appl. Math. **10**, 537–566 (1957)
19. Mock, M.S.: Systems of conservation laws of mixed type. J. Diff. Eq. **7**, 70–88 (1980)
20. Mulder, W., Osher, S., Sethian, J.A.: Computing interface motion in compressible gas dynamics. J. Comput. Phys. **100**(2), 209–228 (1992)
21. Müller, S., Helluy, P., Ballmann, J.: Numerical simulation of a single bubble by compressible two-phase fluids. Internat. J. Numer. Methods Fluids **62**(6), 591–631 (2010)
22. Saurel, R., Abgrall, R.: A simple method for compressible multifluid flows. SIAM J. Sci. Comput. **21**(3), 1115–1145 (1999)

Part II
Theoretical Aspects

An ALE Formulation for Explicit Runge-Kutta Residual Distribution

Remi Abgrall, Luca Arpaia and Mario Ricchiuto

Abstract We consider the solution of hyperbolic conservation laws on moving meshes by means of an Arbitrary Lagrangian Eulerian (ALE) formulation of the Runge-Kutta RD schemes of Ricchiuto and Abgrall (*J.Comput.Phys* 229, 2010). Up to the authors knowledge, the problem of recasting RD schemes into ALE framework has been solved with first order explicit schemes and with second order implicit schemes. Our resulting scheme is explicit and second order accurate when computing discontinuous solutions.

1 Conservation Laws in Arbitrary Lagrangian Eulerian Form

We start by recalling the ALE formulation of conservation laws, which dates back to the early eighties due to the contribution of Donea [10].

Assuming that we are given a domain Ω and a field of displacements that brings every point of the domain from the reference position $\mathbf{X}$ to the actual one $\mathbf{x}(t)$ and that this field is governed by an arbitrary given motion law

$$\frac{d\mathbf{x}(t)}{dt} = \boldsymbol{\sigma}(\mathbf{x}, t), \tag{1}$$

R. Abgrall
Institut für Mathematik Universität Zürich, Winterthurerstrasse 190, 8057 Zürich, Switzerland
e-mail: remi.abgrall@math.uzh.ch

L. Arpaia (✉)
Politecnico di Milano, via La Masa 34, 20156 Milano, Italy
e-mail: luca.arpaia@hotmail.it

M. Ricchiuto
INRIA Université Bordeaux I, 200 Rue Vieille Tour, 33405 Talence, France
e-mail: mario.ricchiuto@inria.fr

J. Fuhrmann et al. (eds.), *Finite Volumes for Complex Applications VII - Methods and Theoretical Aspects*, Springer Proceedings in Mathematics & Statistics 77,
DOI: 10.1007/978-3-319-05684-5_4,

Solving Eq. 1 gives back $\forall t > 0$ the actual configuration through the mapping

$$A(t) : \Omega_X \to \Omega_x(t), \qquad \mathbf{x} = A(\mathbf{X}, t) \tag{2}$$

with the condition $A(\mathbf{X}, 0) = \mathbf{X}$. Let the Jacobian matrix of the mapping be $\mathscr{J}_A = \frac{\partial \mathbf{x}}{\partial \mathbf{X}}$ and assume that $J_A = \det \mathscr{J}_A \neq 0$, i.e. the mapping A is assumed to be invertible.

The conservation of the scalar u can be stated within a control volume which is moving following the domain arbitrary mapping of Eq. 2. The differential form of conservation law in ALE formulation reads in actual coordinates

$$\left.\frac{\partial (J_A u)}{\partial t}\right|_X + J_A \nabla \cdot (\mathbf{f} - u\boldsymbol{\sigma}) = 0 \tag{3}$$

with $\mathbf{f}$ the flux of u through the borders of the volume. Simple relations can be used to prove the so called Geometric Conservation Law (GCL)

$$\left.\frac{\partial J_A}{\partial t}\right|_X = J_A \nabla \cdot \boldsymbol{\sigma} \tag{4}$$

Last equation is a constraint the points of the domain have to satisfy during their arbitrary motion. Using Eq. 4 into Eq. 3 it is possible to obtain a mixed formulation where the ALE part of the flux is in a quasilinear form

$$\left.\frac{\partial u}{\partial t}\right|_X + \nabla \cdot \mathbf{f} - \boldsymbol{\sigma} \cdot \nabla u = 0 \tag{5}$$

2 Residual Distribution for 2D Steady Conservation Laws

The foundations of Residual Distribution (RD) can be traced to the work of [6, 16, 19] on residual based schemes, and to the fluctuation splitting approach of Roe and coworkers [20, 21]. Consider the *steady limit* of the conservation law

$$\frac{\partial u}{\partial t} + \nabla \cdot \mathbf{f}(u) = 0 \tag{6}$$

Discretize the spatial domain by a triangulation $\mathscr{T}_h$, and consider the standard P^1 continuous approximation $u^h(\mathbf{x}, t) = \sum_{j=1}^{N+1} \varphi_j(\mathbf{x}) u_j(t)$ with φ_j the continuous piecewise linear Lagrange basis functions. The RD approximation of Eq. 6 is obtained as

1. On each element $K \in \mathscr{T}_h$ compute the residual

$$\phi^K = \int_K \nabla \cdot \mathbf{f}(u^h)\, d\mathbf{x} = \int_{\partial K} \mathbf{f}(u^h) \cdot \mathbf{n}\, ds \tag{7}$$

2. Distribute the residuals to the nodes of the element $i, j, k \in K$

$$\phi_j^K = \beta_j^K \phi^K \,, \quad \sum_{j \in K} \phi_j^K = \phi^K \tag{8}$$

3. Assemble elemental contributions:

$$|S_i| \frac{du_i}{dt} + \sum_{K \in \mathscr{D}_i} \phi_i^K = 0, \quad \forall i \in \mathscr{T}_h \tag{9}$$

with $\mathscr{D}_i$ the set of elements sharing node i, an S_i the standard median dual cell. Marching Eq. 9 to steady state one obtains a discrete solution which can be shown to be an approximation of the weak solution of Eq. 6, see [5].

In practice residual in Eq. 7 can be computed either by contour integration [7, 18], or by introducing an exact Jacobian mean value linearization so that

$$\phi^K = \sum_{j \in K} k_j u_j, \qquad k_i = \frac{1}{2} \bar{\mathbf{a}} \cdot \mathbf{n}_i \tag{10}$$

with $\mathbf{n}_i$ the inward normal to the edge facing node i, scaled by the edge length and the elemental average of the advective speed $\bar{\mathbf{a}} = \frac{1}{|K|} \int_K \frac{\partial \mathbf{f}(u^h)}{\partial u} \, d\mathbf{x}$.

3 Genuinely Explicit RK-RD Time Marching Procedure

In the time dependent case we use the Petrov-Galerkin form of RD [3, 12, 14, 17]

$$\sum_{K \in \mathscr{D}_i} \sum_{j \in K} m_{ij}^K \frac{du_j}{dt} + \sum_{K \in \mathscr{D}_i} \beta_i^K \phi^K = 0 \tag{11}$$

with the mass-matrix $m_{ij}^K = \int_K \varphi_j w_i \, d\mathbf{x}$, and with $w_i = \varphi_i + \gamma_i$ the RD Petrov-Galerkin test function. Considering a time step Δt subjected to CFL condition, the second order explicit RK2-RD schemes of [17] is obtained as:

1. *First RK step:* $\dfrac{\Delta u_1}{\Delta t} + e_1 = 0$, *with* $e_1 = e(u^n)$. We use the Petrov-Galerkin RD statement and mass lumping, leading to $(\Delta u_1 = u_1 - u^n)$

$$|S_i| \frac{\Delta u_1}{\Delta t} + \sum_{K \in \mathscr{D}_i} \phi_i^K (u^n) = 0 \tag{12}$$

2. *Second RK step*: $\frac{\Delta u_2}{\Delta t} + e_2 = 0$, *with* $e_2 = (e(u^n) + e(u_1)/2)$. We use the Petrov-Galerkin RD statement, however two different approximations of the equation are used in the Galerkin part and in the stabilization, namely ($\Delta u_2 = u^{n+1} - u^n$)

$$\int_\Omega \varphi_i \left(\frac{\Delta u_2^h}{\Delta t} + \nabla \cdot \mathbf{f}_2(u^h) \right) d\mathbf{x} + \sum_{K \in \mathscr{D}_i} \int_K \gamma_i \left(\frac{\Delta u_1^h}{\Delta t} + \nabla \cdot \mathbf{f}_2(u^h) \right) d\mathbf{x} = 0 \tag{13}$$

3. Mass lumping is applied to the Galerkin integrals in Eq. 13. This leads to

$$|S_i| \left\{ \left(\frac{\Delta u_2}{\Delta t}\right)_i - \left(\frac{\Delta u_1}{\Delta t}\right)_i \right\} = - \sum_{K \in \mathscr{D}_i} \Phi_i^{RK(2)} \tag{14}$$

where

$$\Phi_i^{RK(2)} = \sum_{j \in K} m_{ij}^K \left(\frac{\Delta u_1}{\Delta t}\right)_j + \frac{1}{2}\phi_i^K(u_1) + \frac{1}{2}\phi_i^K(u^n)$$

4 Residual Distribution Schemes for Moving Grids

In this section we recast the scheme of Eq. 14 in ALE form. ALE formulations of RD have been proposed in the work of Michler and Deconinck [15], who achieved first order with an Explicit Euler time integrator, and later Dobes and Deconinck (see e.g. [8]) who moved to high order time approximation (BDF, Crank Nicholson), thus obtaining second order of accuracy. The aim of this work is to obtain a numerical solution with second order of accuracy using a faster explicit Runge Kutta time integrator.

4.1 Explicit Euler Time Stepping

We start from the stabilized Finite Element approximation of ALE Eq. 3, discretized in time with Explicit Euler (EE):

$$\frac{\Delta}{\Delta t} \int_{\Omega(t)} w_i u^h \, d\mathbf{x} + \int_{\Omega(t^*)} w_i \nabla \cdot \left(\mathbf{f}(u_h^n) - \boldsymbol{\sigma}_h^* u_h^n \right) d\mathbf{x} = 0 \tag{15}$$

Imposing a uniform flow, the discrete counterpart of Eq. 4 arises from the above approximation and it is referred to as Discrete Geometric Conservation Law (DGCL). The satisfaction of the DGCL is very important when numerically solving PDEs in ALE form [11, 13]. In [8] the problem is closed substituting directly the GCL

condition Eq. 4, discretized with the same time integrator used for the PDEs (in this case EE), into the ALE flux part of Eq. 15. In such a way, when a uniform flow is imposed, the volume variation within the time step is exactly balanced by an equal term. This tecnique could be employed for every time approximation, even the more complex, provided that the *Geometric Source Term* arising from the above substitution is discretized with the same time scheme which we use for the PDEs. Here instead we follow Farhat [11, 13] which shows that by choosing $\boldsymbol{\sigma}_j^* = (\mathbf{x}_j^{n+1} - \mathbf{x}_j^n)/\Delta t$, and by setting $t^* = t^{n+1/2}$, most single step time discretizations satisfy naturally the DGCL condition. In our case if a uniform flow is imposed one gets

$$\int_{\Omega_h^{n+1}} w_i \, d\mathbf{x} - \int_{\Omega_h^n} w_i \, d\mathbf{x} = \Delta t \int_{\Omega_h^{n+1/2}} w_i \nabla \cdot \boldsymbol{\sigma}_h^* \, d\mathbf{x} \tag{16}$$

which is in fact an identity for P^1 interpolation. This identity, and the properties of the P^1 basis functions, can be used to prove

$$\begin{aligned} &\int_{\Omega^{n+1}} w_i u^h \, d\mathbf{x} - \int_{\Omega^n} w_i u^h \, d\mathbf{x} = \\ &= \int_{\Omega^{n+1/2}} w_i \left(u_h^{n+1} - u_h^n \right) d\mathbf{x} + \Delta t \int_{\Omega^{n+1/2}} w_i \frac{\left(u_h^{n+1} + u_h^n \right)}{2} \nabla \cdot \boldsymbol{\sigma}_h^* \, d\mathbf{x} \end{aligned} \tag{17}$$

Substituting this expression in Eq. 15 we end with

$$\begin{aligned} &\int_{\Omega^{n+1/2}} \left(1 + \frac{\Delta t}{2} \nabla \cdot \boldsymbol{\sigma}_h^* \right) w_i \left(u_h^{n+1} - u_h^n \right) d\mathbf{x} + \\ &\quad + \Delta t \int_{\Omega^{n+1/2}} w_i \left(\nabla \cdot \mathbf{f}(u_h^n) - \boldsymbol{\sigma}_h^* \cdot \nabla u_h^n \right) d\mathbf{x} = 0 \end{aligned} \tag{18}$$

Lumping the mass matrix, recalling that $\nabla \cdot \boldsymbol{\sigma}_h^*\big|_K$ is constant in the P^1 case, and using the analogy with Residual Distribution method on the right-handside, we get

$$\sum_{K \in \mathscr{D}_i} \left(1 + \frac{\Delta t}{2} \nabla \cdot \boldsymbol{\sigma}_h^* \right) \frac{|K^{n+1/2}|}{3} \left(u_i^{n+1} - u_i^n \right) = -\Delta t \sum_{K \in \mathscr{D}_i} \beta_i^K \phi^K \left(u_h^n \right)$$

The final algorithm reads

$$\frac{|\tilde{S}_i^{n+1/2}|}{\Delta t} \left(u_i^{n+1} - u_i^n \right) = - \sum_{K \in \mathscr{D}_i} \beta_i^K \phi^K \left(u_h^n \right) \tag{19}$$

where the median dual cell area that appears in Eq. 9 is evaluated at the midpoint configuration, and modified to take into account the grid distortion as follows

$$|\tilde{S}_i^{n+1/2}| = \sum_{K\in\mathscr{D}_i}\left(1+\frac{\Delta t}{2}\nabla\cdot\boldsymbol{\sigma}_h^*\right)\frac{|K^{n+1/2}|}{3} \tag{20}$$

The DGCL is satisfied by construction. Note that, due to the extra term $\boldsymbol{\sigma}_h^*\cdot\nabla u_h$ in Eq. 18, the k_i parameter used to evaluate ϕ^K (cf. Eq. 10) is modified as

$$k_i = \frac{1}{2}(\bar{\mathbf{a}}-\bar{\boldsymbol{\sigma}})\cdot\mathbf{n}_i \tag{21}$$

4.2 Two-Stage RK-RD Time Stepping

The extension of scheme of Eq. 14 has to be done carefully to preserve the DGCL. The problem is related to the balance of the different time increments used in the stabilization and Galerkin parts. To handle this, we use for the stabilization term the nonconservative ALE form Eq. 5. Proceeding as in Sects. 3 and 4.1 we have:

1. *First RK step*: It is the EE of Eq. 19 with linearized residuals and geometry computed at midpoint configuration, and k_i modified according to Eq. 21.
2. *Second RK step*: The discretization of the Galerkin part writes

$$\frac{\Delta}{\Delta t}\int_{\Omega(t)}\varphi_i u_2^h\,d\mathbf{x}+\int_{\Omega^{n+1/2}}\varphi_i\nabla\cdot\left(\mathbf{f}(u^h)-\boldsymbol{\sigma}_h^* u^h\right)_2 d\mathbf{x} \tag{22}$$

   For the stabilization term instead we use

$$\sum_{K\in\mathscr{D}_i}\int_{K^{n+1/2}}\gamma_i\frac{\Delta u_1^h}{\Delta t}\,d\mathbf{x}+\sum_{K\in\mathscr{D}_i}\int_{K^{n+1/2}}\gamma_i\left(\nabla\cdot\mathbf{f}(u^h)^n-\boldsymbol{\sigma}_h^*\cdot\nabla u^h\right)_2 d\mathbf{x} \tag{23}$$

   both the parts satisfy the DGCL condition by construction.
3. We use Eq. 17, mass lump the Galerkin integrals, and sum up the two terms to get

$$|\tilde{S}_i^{n+1/2}|\left\{\left(\frac{\Delta u_2}{\Delta t}\right)_i-\left(\frac{\Delta u_1}{\Delta t}\right)_i\right\} = -\sum_{K\in\mathscr{D}_i}\Phi_i^{RK(2)} \tag{24}$$

   with

$$\Phi_i^{RK(2)} = \sum_{j\in K} m_{ij}^K\left(\frac{\Delta u_1}{\Delta t}\right)_j+\frac{1}{2}\phi_i^K(u_1)+\frac{1}{2}\phi_i^K(u^2)$$

Besides the modified definition of the k_i parameters and of the median dual area, the final scheme is formally identical to the original one.

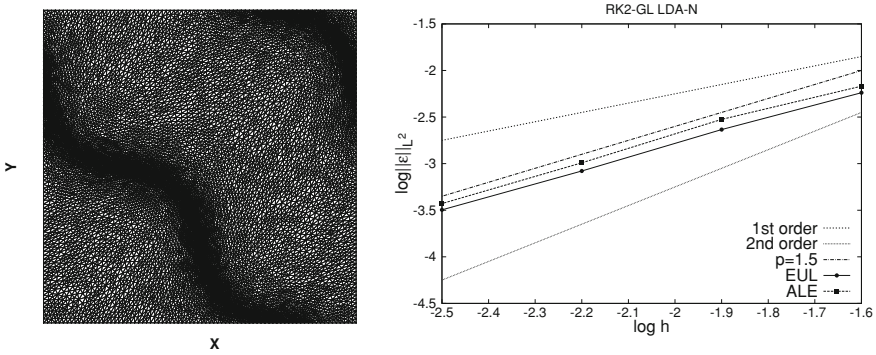


Fig. 1 Vortex advection. Mesh deformation (*left*) and grid convergence (*right*)

5 Application to the Perfect Gas Euler Equations

We tets the proposed ALE formulation on the perfect gas Euler equations. We have used the non-linear LDA-N distribution scheme. We refer to [1, 2, 4, 17] for details concerning this scheme, and for the implementation of RD for systems.

5.1 Advection of a Vortex

We measure the accuracy of the scheme on the advection of a constant density vortex (see [9] for details). The mapping of Eq. 2 is defined according to (cf. Fig. 1)

$$\mathbf{x} = \mathbf{X} + \sin(a\pi X)\sin(b\pi Y)(c\sin(d\pi t/t_{\max},\ e\sin(f\pi t/t_{\max})))$$

We can see from the right picture on Fig. 1 that the expected order of accuracy is achieved both in the fixed mesh and ALE framework.

5.2 Wind Tunnel with Wall Deflection

We consider a simple application involving moving boundaries. We have a 2D channel $[2 \times 1]$ with an hinge on the lower surface placed at $x = 0.25$. This hinge allows a rigid deflection of the lower wall of an angle α governed by the motion law

$$\begin{cases} \alpha(t) = \alpha_{max}\left(1 - e^{-t/\tau}\right) & t \le t_{switch} \\ \alpha(t) = \alpha_{max} - 2\alpha_{max}\left(1 - e^{-(t-t_{switch})/\tau}\right) & t > t_{switch} \end{cases} \tag{25}$$

The Mach number at the inlet is $M = 3$.

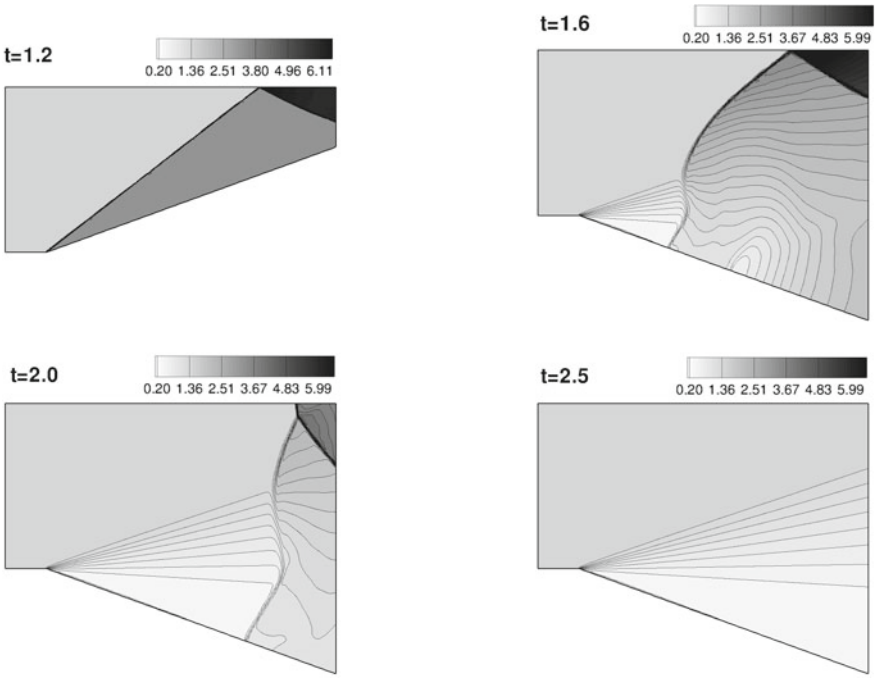


Fig. 2 Mach 3 wind tunnel with a deflecting wall. Density isolines

The flow shows two stable configurations. The first is a regular shock reflection on the upper wall ($t \approx 1.2$), the second is a supersonic Prandtl-Mayer expansion ($t \approx 2.5$). In between these two states, the flow shows a transient with the formation and shedding of complex interacting shocks, which are sharply and monotonically captured by our scheme (Fig. 2).

References

1. Abgrall, R.: Toward the ultimate conservative scheme: following the quest. J. Comput. Phys **167**(2), 277–315 (2001)
2. Abgrall, R.: Essentially non oscillatory residual distribution schemes for hyperbolic problems. J. Comput. Phys **214**(2), 773–808 (2006)
3. Abgrall, R., Mezine, M.: Construction of second-order accurate monotone and stable residual distribution schemes for unsteady flow problems. J. Comput. Phys. **188**, 16–55 (2003)
4. Abgrall, R., Mezine, M.: Construction of second-order accurate monotone and stable residual distribution schemes for steady flow problems. J. Comput. Phys. **195**, 474–507 (2004)
5. Abgrall, R., Roe, P.: High-order fluctuation schemes on triangular meshes. J. Sci. Comput. **19**(3), 3–36 (2003)
6. Brooks, A., Hughes, T.J.R.: Streamline upwind petrov-galerkin formulation for convection dominated flows with particular emphasis on the incompressible navier-stokes equations. Comp. Meth. Mech. Eng. **32**, 199–259 (1982)

7. Csik, A., Ricchiuto, M., Deconinck, H.: A conservative formulation of the multidimensional upwind residual distribution schemes for general nonlinear conservation laws. J. Comput. Phys. **179**(2), 286–312 (2002)
8. Dobes, J.: Numerical algorithms for the computation of steady and unsteady compressible flow over moving geometries—application to fluid-structure interaction. Ph.D. thesis, Von Karman Institute (2007)
9. Dobes, J., Deconinck, H.: Second order blended multidimensional upwind residual distribution scheme for steady and unsteady computations. J. Comput. Appl. Math. **215**(1), 378–389 (2006)
10. Donea, J.: Computational methods for transient analysis. Chap. 10. Arbitrary Lagrangian Eulerian Finite Element Methods. Elsevier Science Publisher, Amsterdam (1983)
11. Farhat, C., Geuzaine, P., Grandmont, C.: The discrete geometric conservation law and the nonlinaer stability of ale schemes for the solution of flow problems on moving grids. J. Comput. Phys. **174**(2), 669–694 (2000)
12. Ferrante, A., Deconinck, H.: Solution of the unsteady Euler equations using residual distribution and flux corrected transport. Technical Report VKI-PR 97–08, von Karman Institute for Fluid Dynamics, Belgium (1997)
13. Guillard, H., Farhat, C.: On the significance of the geometric conservation law for flow computations on moving meshes. Comput. Methods Appl. Mech. Eng. **190**(34), 1467–1482 (2000)
14. Maerz, J., Degrez, G.: Improving time accuracy of residual distribution schemes. Technical Report VKI-PR 96–17. von Karman Institute for Fluid Dynamics, Belgium (1996)
15. Michler, C., Deconinck, H.: An arbitrary lagrangian eulerian formulation for residual distribution schemes on moving grids. Comput. Fluids **32**(1), 59–71 (2001)
16. Ni, R.: A multiple grid scheme for solving the euler equation. AIAA J. **20**, 1565–1571 (1981)
17. Ricchiuto, M., Abgrall, R.: Explicit runge-kutta residual distribution schemes for time dependent problems: second order case. J. Comput. Phys. **229**(16), 5653–5691 (2010)
18. Ricchiuto, M., Csik, A.: H. deconinck: residual distribution for general time-dependent conservation laws. J. Comput. Phys **209**(1), 249–289 (2005)
19. Rice, J., Schnipke, R.: A monotone streamline upwind method for convection dominated problems. Comput. Methods Appl. Mech. Eng. **48**, 313–327 (1985)
20. Roe, P.: Fluctuations and signals—a framework for numerical evolution problems. In: Morton, K., Baines, M., (eds.) Numerical Methods for Fluids Dynamics, pp. 219–257. Academic Press, New York (1982)
21. Roe, P.L.: Linear advection schemes on triangular meshes. Technical Report CoA 8720, Cranfield Institute of Technology, Bedford (1987)

Gradient Schemes for an Obstacle Problem

Yahya Alnashri and Jerome Droniou

Abstract The aim of this work is to adapt the gradient schemes, discretisations of weak variational formulations using independent approximations of functions and gradients, to obstacle problems modelled by linear and non-linear elliptic variational inequalities. It is highlighted in this paper that four properties which are coercivity, consistency, limit conformity and compactness are adequate to ensure the convergence of this scheme. Under some suitable assumptions, the error estimate for linear equations is also investigated.

1 Introduction

We are interested in obstacle problems formulated as linear and non-linear elliptic variational inequalities and their approximate solutions obtained by gradient schemes. In what follows, Ω is an open bounded subset of $\mathbb{R}^d$. The problem we consider is

$$(-\mathrm{div}(\Lambda(x,\bar{u})\nabla\bar{u}) - f(x))(g(x) - \bar{u}(x)) = 0, \quad x \in \Omega, \tag{1a}$$

$$\bar{u}(x) \leq g(x), \quad x \in \Omega, \tag{1b}$$

$$\mathrm{div}(\Lambda(x,\bar{u})\nabla\bar{u}) + f(x) \geq 0, \quad x \in \Omega, \tag{1c}$$

$$\bar{u}(x) = 0, \quad x \in \partial\Omega, \tag{1d}$$

Y. Alnashri (✉) · J. Droniou
Monash University, Melbourne, Australia
e-mail: yahya.alnashri@monash.edu

J. Droniou
e-mail: jerome.droniou@monash.edu

Y. Alnashri
Umm–Alqura University, Mecca, Saudi Arabia

J. Fuhrmann et al. (eds.), *Finite Volumes for Complex Applications VII - Methods and Theoretical Aspects*, Springer Proceedings in Mathematics & Statistics 77,
DOI: 10.1007/978-3-319-05684-5_5,

under the following assumptions:

$$\begin{gathered}\Lambda \text{is a Caratheodory function from } \Omega \times \mathbb{R} \text{ to } S_d(\mathbb{R}) \\ \text{(the set of } d \times d \text{ symmetric matrices) such that,} \qquad (2a) \\ \text{for a.e. } x \in \Omega \text{ and all } s \in \mathbb{R},\ \Lambda(x,s) \text{ has eigenvalues in } (\underline{\lambda}, \overline{\lambda}) \subset (0, +\infty), \\ f \in L^2(\Omega),\ \ g \in H^1(\Omega) \text{ and } \gamma(g) \geq 0 \text{ on } \partial\Omega. \qquad (2b)\end{gathered}$$

Under these assumptions, the weak formulation of Problem (1a–1d) is written

$$\begin{gathered}\text{Find } \bar{u} \in \mathrm{K} = \{v \in H_0^1(\Omega) \ : \ v \leq g \text{ in } \Omega\} \text{ such that, } \forall v \in \mathrm{K}, \\ \int_\Omega \Lambda(x,\bar{u}) \nabla \bar{u}(x) \cdot \nabla(\bar{u}(x) - v(x)) \mathrm{d}x \ \leq \ \int_\Omega f(x)(\bar{u}(x) - v(x)) \mathrm{d}x. \end{gathered} \qquad (3)$$

Note that K is a non-empty set since $v = \min(0, g) \in K$.

Variational inequalities with different boundary conditions have been employed to model several physical problems, such as lubrication phenomena and seepage of liquid in porous media (see [7] and references therein). Mathematical theories associated to existence, uniqueness and stability of the solution of obstacle problems have been extensively developed (see [4, 9], for example). From the numerical perspective, Herbin and Marchand [8] showed that if $\Lambda \equiv I_d$ the solution of the 2-points finite volume scheme converges in $L^2(\Omega)$ to the unique solution as the size mesh tends to zero. Under H^2 regularity conditions on the exact solution they provide $\mathcal{O}(h)$ error estimate. This 2-points finite volume method, however, requires grids to satisfy a strong orthogonality assumption. Under a number of assumptions, Falk [6] underlines that the convergence estimate of finite elements method is of order h. Both schemes are only applicable for $\Lambda \equiv I_d$ in Problem (1a–1d).

Our goal in this paper is to use gradient schemes to construct a general formulation of several discretisations of Problem (3). The gradient scheme has been developed to analyse the convergence of numerical methods for diffusion equations (see [3, 5]). Furthermore, Droniou et al. [3] noticed that this framework contains various methods such as Galerkin and some MPFA schemes.

This paper is arranged as follows. In Sect. 2, we present the definitions of some concepts, which are necessary to construct gradient schemes and to prove their convergence. In Sect. 3, we give an error estimate and a convergence proof in the linear case. Since we deal here with nonconforming schemes, the technique used in [6] is not useful to obtain error estimates. Although we use a similar technique as in [5], dealing with variational inequalities in this nonconforming setting requires us to establish new preliminary estimates, which modify the final error estimate. Finally, Sect. 4 is devoted to prove a convergence result for non-linear equations. Numerical experiments will be the purpose of a future work.

2 Gradient Discretisation and Gradient Schemes

Gradient schemes are based on gradient discretisations, which consist of discrete spaces and mappings, and provide a general formulation of different numerical methods. Except for the definition of consistency, the definitions presented here are the same as in [3].

Definition 1 A gradient discretisation $\mathscr{D}$ for homogeneous Dirichlet boundary conditions is defined by a triplet $\mathscr{D} = (X_{\mathscr{D},0}, \Pi_{\mathscr{D}}, \nabla_{\mathscr{D}})$, where

1. the set of discrete unknowns $X_{\mathscr{D},0}$ is a finite dimensional vector space of $\mathbb{R}$,
2. the linear mapping $\Pi_{\mathscr{D}} : X_{\mathscr{D},0} \longrightarrow L^2(\Omega)$ gives the reconstructed function,
3. the linear mapping $\nabla_{\mathscr{D}} : X_{\mathscr{D},0} \longrightarrow L^2(\Omega)^d$ gives a reconstructed discrete gradient, which must be defined such that $\|\cdot\|_{\mathscr{D}} := \|\nabla_{\mathscr{D}} \cdot \|_{L^2(\Omega)^d}$ is a norm on $X_{\mathscr{D},0}$.

Throughout this paper, $\mathscr{D}$ is a gradient discretisation in the sense of Definition 1. The gradient scheme associated to $\mathscr{D}$ for Problem (3) is given by

$$\begin{aligned}&[c]\text{Find } u \in \mathrm{K}_{\mathscr{D}} = \{v \in X_{\mathscr{D},0} \ : \ \Pi_{\mathscr{D}} v \le g \text{ in } \Omega\} \text{ such that}, \forall v \in \mathrm{K}_{\mathscr{D}},\\ &\int_\Omega \Lambda(x, \Pi_{\mathscr{D}} u(x)) \nabla_{\mathscr{D}} u(x) \cdot \nabla_{\mathscr{D}}(u - v)(x) \mathrm{d}x \ \le \ \int_\Omega f(x) \Pi_{\mathscr{D}}(u - v)(x) \mathrm{d}x.\end{aligned} \tag{4}$$

Definition 2 (*Coercivity, consistency, limit-conformity and compactness*) Let $C_{\mathscr{D}}$ be the norm of linear mapping $\Pi_{\mathscr{D}}$, defined by

$$C_{\mathscr{D}} = \max_{v \in X_{\mathscr{D},0} \setminus \{0\}} \frac{\|\Pi_{\mathscr{D}} v\|_{L^2(\Omega)}}{\|\nabla_{\mathscr{D}} v\|_{L^2(\Omega)^d}}. \tag{5}$$

A sequence $(\mathscr{D}_m)_{m \in \mathbb{N}}$ is called *coercive* if there exits $C_P \in \mathbb{R}_+$ such that $C_{\mathscr{D}_m} \le C_P$ for all $m \in \mathbb{N}$.

We say that a sequence $(\mathscr{D}_m)_{m \in \mathbb{N}}$ is *consistent* if, for all $\varphi \in K$, $\lim_{m \to \infty} S_{\mathscr{D}_m}(\varphi) = 0$, where $S_{\mathscr{D}} : K \longrightarrow [0, +\infty)$ is defined by

$$\forall \varphi \in K, \ S_{\mathscr{D}}(\varphi) = \min_{v \in K_{\mathscr{D}}} (\|\Pi_{\mathscr{D}} v - \varphi\|_{L^2(\Omega)} + \|\nabla_{\mathscr{D}} v - \nabla \varphi\|_{L^2(\Omega)^d}). \tag{6}$$

A sequence $(\mathscr{D}_m)_{m \in \mathbb{N}}$ is called *limit-conforming* if $\lim_{m \to \infty} W_{\mathscr{D}_m}(\varphi) = 0$ for all $\varphi \in H_{\mathrm{div}}(\Omega)$, where $W_{\mathscr{D}} : H_{\mathrm{div}}(\Omega) \longrightarrow [0, +\infty)$ is defined by

$$\forall \varphi \in H_{\mathrm{div}}(\Omega), \ W(\varphi) = \sup_{v \in X_{\mathscr{D},0} \setminus \{0\}} \frac{\left| \int_\Omega (\nabla_{\mathscr{D}} v \cdot \varphi + \Pi_{\mathscr{D}} v \cdot \mathrm{div}(\varphi)) \mathrm{d}x \right|}{\|\nabla_{\mathscr{D}} v\|_{L^2(\Omega)^d}}. \tag{7}$$

A sequence $(\mathscr{D}_m)_{m\in\mathbb{N}}$ is called *compact* if, for any sequence $(u_m)_{m\in\mathbb{N}}$ with $u_m \in \mathrm{K}_{\mathscr{D}_m}$ and such that $(\|u_m\|_{\mathscr{D}_m})_{m\in\mathbb{N}}$ is bounded, the sequence $(\|\Pi_{\mathscr{D}_m} u_m\|_{L^2(\Omega)})_{m\in\mathbb{N}}$ is relatively compact in $L^2(\Omega)$.

3 Convergence and Error Estimate in the Linear Case

We consider here $\Lambda(x,u) = \Lambda(x)$. Based on the previous properties, we give an error estimate that requires $\mathrm{div}(\Lambda\nabla\bar{u}) \in L^2(\Omega)$. We note that Brezis and Stampacchia [1] establish an H^2 regularity result on $\bar{u}$ under proper assumption on the data. If we further assume that Λ is Lipschitz-continuous, then $\mathrm{div}(\Lambda\nabla\bar{u}) \in L^2(\Omega)$.
In what follows, we define the interpolant $P_{\mathscr{D}} : K \longrightarrow K_{\mathscr{D}}$ as follows

$$P_{\mathscr{D}}\varphi = \arg\min_{v\in K_{\mathscr{D}}} (\|\Pi_{\mathscr{D}} v - \varphi\|_{L^2(\Omega)} + \|\nabla_{\mathscr{D}} v - \nabla\varphi\|_{L^2(\Omega)^d}). \tag{8}$$

Theorem 1 (Error estimate) *Under Assumptions* (2a), (2b), *let* $\bar{u} \in K$ *be the solution to Problem* (1a–1d) *and let* $D = \{x \in \Omega : \bar{u}(x) = g(x)\}$. *If we assume that* $\mathscr{D}$ *is a gradient discretisation and* $K_{\mathscr{D}}$ *is a non-empty set, then there exists a unique solution* $u \in K_{\mathscr{D}}$ *to the gradient scheme* (4). *Moreover, if* $\mathrm{div}(\Lambda\nabla\bar{u}) \in L^2(\Omega)$ *then this solution satisfies the following inequalities:*

$$\|\nabla_{\mathscr{D}} u - \nabla\bar{u}\|_{L^2(\Omega)^d} \le \sqrt{\frac{2}{\underline{\lambda}} E_{\mathscr{D}}(\bar{u}) + \frac{1}{\underline{\lambda}^2}[W_{\mathscr{D}}(\Lambda\nabla\bar{u}) + \overline{\lambda} S_{\mathscr{D}}(\bar{u})]^2} + S_{\mathscr{D}}(\bar{u}), \tag{9}$$

$$\|\Pi_{\mathscr{D}} u - \bar{u}\|_{L^2(\Omega)} \le C_{\mathscr{D}} \sqrt{\frac{2}{\underline{\lambda}} E_{\mathscr{D}}(\bar{u}) + \frac{1}{\underline{\lambda}^2}[W_{\mathscr{D}}(\Lambda\nabla\bar{u}) + \overline{\lambda} S_{\mathscr{D}}(\bar{u})]^2} + S_{\mathscr{D}}(\bar{u}), \tag{10}$$

in which $E_{\mathscr{D}}(\bar{u}) = \displaystyle\int_D (\mathrm{div}(\Lambda\nabla\bar{u}) + f)(\bar{u} - \Pi_{\mathscr{D}}(P_{\mathscr{D}}\bar{u}))\mathrm{d}x.$

Remark 1 Note that $|E_{\mathscr{D}}(\bar{u})| \le \|\mathrm{div}(\Lambda\nabla\bar{u}) + f\|_{L^2(\Omega)} \|\bar{u} - \Pi_{\mathscr{D}}(P_{\mathscr{D}}\bar{u})\|_{L^2(\Omega)}$.

Proof The techniques used in [5] and [7] will be followed in this proof.

Since $\mathrm{K}_{\mathscr{D}}$ is a closed convex set, we can apply Stampacchia's theorem which states that there exists a unique solution to Problem (4).

Under the assumption that $\mathrm{div}(\Lambda\nabla\bar{u}) \in L^2(\Omega)$, we note that $\Lambda\nabla\bar{u} \in H_{\mathrm{div}}(\Omega)$. For any $v \in X_{\mathscr{D},0}$, replacing φ with $\Lambda\nabla\bar{u}$ in the definition of limit conformity (7) therefore implies

$$\int_\Omega \nabla_{\mathscr{D}} v \cdot \Lambda\nabla\bar{u}\mathrm{d}x + \int_\Omega \Pi_{\mathscr{D}} v \cdot \mathrm{div}(\Lambda\nabla\bar{u})\mathrm{d}x \le \|\nabla_{\mathscr{D}} v\|_{L^2(\Omega)^d} W_{\mathscr{D}}(\Lambda\nabla\bar{u}). \tag{11}$$

It is obvious that

$$\begin{aligned}\int_\Omega \Pi_{\mathscr{D}}(u - P_{\mathscr{D}}\bar{u})\mathrm{div}(\Lambda\nabla\bar{u})\mathrm{d}x = &\int_\Omega (\Pi_{\mathscr{D}}u - g)(\mathrm{div}(\Lambda\nabla\bar{u}) + f)\mathrm{d}x\\ &+ \int_\Omega (g - \Pi_{\mathscr{D}}(P_{\mathscr{D}}\bar{u}))(\mathrm{div}(\Lambda\nabla\bar{u}) + f)\mathrm{d}x\\ &- \int_\Omega (\Pi_{\mathscr{D}}u - \Pi_{\mathscr{D}}(P_{\mathscr{D}}\bar{u}))f\mathrm{d}x.\end{aligned}$$

Using (1a–1d) and $u \in K_{\mathscr{D}}$, we obtain $\int_\Omega (\Pi_{\mathscr{D}}u - g)(\mathrm{div}(\Lambda\nabla\bar{u}) + f)\mathrm{d}x \le 0$, so that

$$\begin{aligned}\int_\Omega \Pi_{\mathscr{D}}(u - P_{\mathscr{D}}\bar{u})\mathrm{div}(\Lambda\nabla\bar{u})\mathrm{d}x \le &\int_\Omega (g - \Pi_{\mathscr{D}}(P_{\mathscr{D}}\bar{u}))(\mathrm{div}(\Lambda\nabla\bar{u}) + f)\mathrm{d}x\\ &- \int_\Omega (\Pi_{\mathscr{D}}u - \Pi_{\mathscr{D}}(P_{\mathscr{D}}\bar{u}))f\mathrm{d}x\\ = &\int_\Omega (g - \bar{u})(\mathrm{div}(\Lambda\nabla\bar{u}) + f)\mathrm{d}x\\ &+ \int_\Omega (\bar{u} - \Pi_{\mathscr{D}}(P_{\mathscr{D}}\bar{u}))(\mathrm{div}(\Lambda\nabla\bar{u}) + f)\mathrm{d}x\\ &- \int_\Omega (\Pi_{\mathscr{D}}u - \Pi_{\mathscr{D}}(P_{\mathscr{D}}\bar{u}))f\mathrm{d}x.\end{aligned}$$

It follows, since $\bar{u}$ is the solution to Problem (1a–1d),

$$\begin{aligned}\int_\Omega \Pi_{\mathscr{D}}(u - P_{\mathscr{D}}\bar{u})\mathrm{div}(\Lambda\nabla\bar{u})\mathrm{d}x \le &\int_\Omega (\bar{u} - \Pi_{\mathscr{D}}(P_{\mathscr{D}}\bar{u}))(\mathrm{div}(\Lambda\nabla\bar{u}) + f)\mathrm{d}x\\ &- \int_\Omega (\Pi_{\mathscr{D}}u - \Pi_{\mathscr{D}}(P_{\mathscr{D}}\bar{u}))f\mathrm{d}x.\end{aligned}$$

Because $\mathrm{div}(\Lambda\nabla\bar{u}) + f = 0$ in $\Omega \setminus D$, the above inequality becomes

$$\begin{aligned}\int_\Omega \Pi_{\mathscr{D}}(u - P_{\mathscr{D}}\bar{u})\mathrm{div}(\Lambda\nabla\bar{u})\mathrm{d}x \le &\int_D (\bar{u} - \Pi_{\mathscr{D}}(P_{\mathscr{D}}\bar{u}))(\mathrm{div}(\Lambda\nabla\bar{u}) + f)\mathrm{d}x\\ &- \int_\Omega (\Pi_{\mathscr{D}}u - \Pi_{\mathscr{D}}(P_{\mathscr{D}}\bar{u}))f\mathrm{d}x.\end{aligned}$$

Using the definition of $E_{\mathscr{D}}(\bar{u})$, one has

$$\int_\Omega \Pi_{\mathscr{D}}(P_{\mathscr{D}}\bar{u} - u)\mathrm{div}(\Lambda\nabla\bar{u})\mathrm{d}x \ge -E_{\mathscr{D}}(\bar{u}) - \int_\Omega \Pi_{\mathscr{D}}(P_{\mathscr{D}}\bar{u} - u)f\mathrm{d}x.$$

From this inequality and setting $v = P_{\mathscr{D}}\bar{u} - u \in X_{\mathscr{D},0}$ in (11), we obtain

$$\int_\Omega \nabla_{\mathscr{D}}(P_{\mathscr{D}}\bar{u}-u)\cdot\Lambda\nabla\bar{u}\mathrm{d}x-\int_\Omega f\Pi_{\mathscr{D}}\,(P_{\mathscr{D}}\bar{u}-u)\mathrm{d}x\le E_{\mathscr{D}}(\bar{u})$$
$$+\|\nabla_{\mathscr{D}}(P_{\mathscr{D}}\bar{u}-u)\|_{L^2(\Omega)^d}W_{\mathscr{D}}(\Lambda\nabla\bar{u}).$$

Since u is the solution to Problem (4), we get

$$\int_\Omega \Lambda\nabla_{\mathscr{D}}(P_{\mathscr{D}}\bar{u}-u)\cdot[\nabla\bar{u}-\nabla_{\mathscr{D}}u]\mathrm{d}x\le\|\nabla_{\mathscr{D}}(P_{\mathscr{D}}\bar{u}-u)\|_{L^2(\Omega)^d}W_{\mathscr{D}}(\Lambda\nabla\bar{u})+E_{\mathscr{D}}(\bar{u})$$

and, thanks to the definition of $P_{\mathscr{D}}$, we obtain

$$\underline{\lambda}\|\nabla_{\mathscr{D}}(P_{\mathscr{D}}\bar{u})-\nabla_{\mathscr{D}}u\|^2_{L^2(\Omega)^d}$$
$$\le\|\nabla_{\mathscr{D}}(P_{\mathscr{D}}\bar{u})-\nabla_{\mathscr{D}}u\|_{L^2(\Omega)^d}[W_{\mathscr{D}}(\Lambda\nabla\bar{u})+\overline{\lambda}S_{\mathscr{D}}(\bar{u})]+E_{\mathscr{D}}(\bar{u}).$$

Applying Young's inequality leads to

$$\|\nabla_{\mathscr{D}}(P_{\mathscr{D}}\bar{u})-\nabla_{\mathscr{D}}u)\|_{L^2(\Omega)^d}\le\sqrt{\frac{2}{\underline{\lambda}}E_{\mathscr{D}}(\bar{u})+\frac{1}{\underline{\lambda}^2}[W_{\mathscr{D}}(\Lambda\nabla\bar{u})+\overline{\lambda}S_{\mathscr{D}}(\bar{u})]^2}$$

and, from $\|\nabla_{\mathscr{D}}(P_{\mathscr{D}}\bar{u})-\nabla\bar{u}\|\le S_{\mathscr{D}}(\bar{u})$, Estimate (9) is achieved. Using (5), we obtain

$$\|\Pi_{\mathscr{D}}(P_{\mathscr{D}}\bar{u}-u)\|_{L^2(\Omega)}\le C_{\mathscr{D}}\sqrt{\frac{2}{\underline{\lambda}}E_{\mathscr{D}}(\bar{u})+\frac{1}{\underline{\lambda}^2}[W_{\mathscr{D}}(\Lambda\nabla\bar{u})+\overline{\lambda}S_{\mathscr{D}}(\bar{u})]^2}$$

which shows that (10) holds, owing to $\|\Pi_{\mathscr{D}}(P_{\mathscr{D}}\bar{u})-\bar{u}\|_{L^2(\Omega)}\le S_{\mathscr{D}}(\bar{u})$. □

Remark 2 It can be seen in [5] that for most gradient schemes based on meshes, $W_{\mathscr{D}}$ and $S_{\mathscr{D}}$ are $\mathscr{O}(h)$ (where h is the mesh size) if $\bar{u}\in H^2(\Omega)\cap H^1_0(\Omega)$ and Λ is Lipschitz-continuous. In these cases, Theorem 1 gives an $\mathscr{O}(\sqrt{h})$ error estimate. Given that $\mathrm{div}(\Lambda\nabla\bar{u})+f=0$ outside D and $u=g$ on D, there is potential, if g is constant or smooth, for the interplant $P_{\mathscr{D}}$ to give a better approximation of $\bar{u}$ on D. The term $\bar{u}-\Pi_{\mathscr{D}}(P_{\mathscr{D}}\bar{u})$ therefore may be much lower on D than $S_{\mathscr{D}}(\bar{u})$. This means that $E_{\mathscr{D}}$ is expected to be lower than $\mathscr{O}(h)$ and therefore that the error estimate could be indeed better than $\mathscr{O}(\sqrt{h})$ in practice.

From the above theorem, we can obtain the following convergence of the scheme.

Corollary 1 (Convergence) *Let $(\mathscr{D}_m)_{m\in\mathbb{N}}$ be a sequence of gradient discretisation which is coercive, consistent and limit-conforming. Let $\bar{u}$ be the exact solution to Problem (3). Assume that $K_{\mathscr{D}_m}$ is a non-empty set for any $m\in\mathbb{N}$. If $u_m\in K_{\mathscr{D}_m}$ is the solution to gradient scheme (4), then $\Pi_{\mathscr{D}_m}u_m$ converges strongly to $\bar{u}$ in $L^2(\Omega)$ and $\nabla_{\mathscr{D}_m}u_m$ strongly converges in $L^2(\Omega)^d$ to $\nabla\bar{u}$.*

Remark 3 It is noted that the convergence proof and error estimate for linear equations are obtained without using compactness property.

4 Convergence in Non-Linear Case

In this section, we study the convergence of non-linear case written as Problem (1a–1d). Such this non-linear equation can be seen in the seepage problems (see [10]).

Theorem 2 (Convergence) *Under Hypotheses* (2a), (2b), *let* $(\mathscr{D}_m)_{m\in\mathbb{N}}$ *be a sequence of gradient discretisations, which is coercive, consistent, limit-conforming and compact, and such that* $K_{\mathscr{D}_m}$ *is a non-empty set for any* m. *Then, for any* $m \in \mathbb{N}$, *the gradient scheme* (4) *has at least one solution* $u_m \in K_{\mathscr{D}_m}$ *and, up to a subsequence,* $\Pi_{\mathscr{D}_m} u_m$ *converges strongly in* $L^2(\Omega)$ *to a weak solution* $\bar{u}$ *of Problem* (3), *and* $\nabla_{\mathscr{D}_m} u_m$ *strongly converges in* $L^2(\Omega)^d$ *to* $\nabla\bar{u}$.

Proof We follow here the same approach used in [3].

Define the mapping $T : v \longrightarrow u$ where for any $v \in X_{\mathscr{D},0}$, $u \in K_{\mathscr{D}}$ is defined as the solution to

$$\text{for all } w \in \mathrm{K}_{\mathscr{D}},\ \int_\Omega \Lambda(x, \Pi_{\mathscr{D}} v)\nabla_{\mathscr{D}} u \cdot \nabla_{\mathscr{D}}(u - w)\mathrm{d}x \ \le\ \int_\Omega f\,\Pi_{\mathscr{D}}(u - w)\mathrm{d}x.$$

That is u is the solution to the variational inequality with the non-linearity in Λ frozen to v. There is only one such u, so the mapping T is well defined, and it is clearly continuous from $X_{\mathscr{D},0}$ into $X_{\mathscr{D},0}$. Since it sends all of $X_{\mathscr{D},0}$ inside a fixed ball of this space (see estimate to follow), Brouwer's theorem ensures the existence of a fixed point $u = T(u)$, which is a solution to the non-linear variational inequality.

Let $\varphi \in K$. Thanks to the consistency, we can choose $v_m \in K_{\mathscr{D}_m}$ defined as $v_m = P_{\mathscr{D}_m}\varphi$ (see (8)). Setting $u := u_m$ and $v := v_m \in K_{\mathscr{D}_m}$ in (4) and applying the Cauchy-Schwarz inequality, we deduce

$$\begin{aligned}\underline{\lambda}\|\nabla_{\mathscr{D}_m} u_m\|^2_{L^2(\Omega)^d} \le \|f\|_{L^2(\Omega)}(\|\Pi_{\mathscr{D}_m} u_m\|_{L^2(\Omega)} + \|\Pi_{\mathscr{D}_m} v_m\|_{L^2(\Omega)})\\ +\overline{\lambda}\|\nabla_{\mathscr{D}_m} v_m\|_{L^2(\Omega)^d}\|\nabla_{\mathscr{D}_m} u_m\|_{L^2(\Omega)^d}.\end{aligned} \tag{12}$$

Since $\|v_m\|_{\mathscr{D}_m}$ is bounded, (12) can be written as

$$\|\nabla_{\mathscr{D}_m} u_m\|_{L^2(\Omega)^d} \le C$$

in which $C > 0$ is constant. Using Lemma 1.13 in [2] (see also the proof of Theorem 3.5 in [3]), there exists a subsequence, still denoted by $(\mathscr{D}_m)_{m\in\mathbb{N}}$, and $\bar{u} \in H_0^1(\Omega)$, such that $\Pi_{\mathscr{D}_m} u_m$ converges weakly to $\bar{u}$ in $L^2(\Omega)$ and $\nabla_{\mathscr{D}_m} u_m$ converges weakly to $\nabla\bar{u}$ in $L^2(\Omega)^d$. Since $u_m \in K_{\mathscr{D}_m}$, passing to the limit in $\Pi_{\mathscr{D}_m} u_m \le g$ shows that $\bar{u}$ is in K. Using the compactness hypothesis, we see that the convergence of $\Pi_{\mathscr{D}_m} u_m$ to $\bar{u}$ is actually strong in $L^2(\Omega)$. Up to another subsequence, we can therefore assume that this convergence is also true almost everywhere. To complete the proof, it remains to prove the strong convergence of $\nabla_{\mathscr{D}_m} u_m$ and that $\bar{u}$ is the solution to (3).

It is classical that if $U_m \to U$ in $L^2(\Omega)^d$, then $\|U\|_{L^2(\Omega)^d} \leq \liminf_{m\longrightarrow\infty} \|U_m\|_{L^2(\Omega)^d}$. Using the positiveness of Λ, the strong convergence of $\Pi_{\mathscr{D}_m} u_m$ to $\bar{u}$ and the weak convergence of $\nabla_{\mathscr{D}_m} u_m$ to $\nabla\bar{u}$, we can adapt the proof of this classical result to see that

$$\int_\Omega \Lambda(x,\bar{u})\nabla\bar{u}\cdot\nabla\bar{u}\mathrm{d}x \leq \liminf_{m\longrightarrow\infty} \int_\Omega \Lambda(x,\Pi_{\mathscr{D}_m}u_m)\nabla_{\mathscr{D}_m}u_m\cdot\nabla_{\mathscr{D}_m}u_m\mathrm{d}x. \tag{13}$$

Thanks to the consistency of the gradient discretisations, $\Pi_{\mathscr{D}_m}(P_{\mathscr{D}_m}\varphi) \longrightarrow \varphi$ strongly in $L^2(\Omega)$ and $\nabla_{\mathscr{D}_m}(P_{\mathscr{D}_m}\varphi) \longrightarrow \nabla\varphi$ strongly in $L^2(\Omega)^d$. This later convergence and the a.e. convergence of $\Lambda(.,\Pi_{\mathscr{D}_m}u_m)$ show that $\Lambda(\cdot,\Pi_{\mathscr{D}_m}u_m)\nabla_{\mathscr{D}_m}(P_{\mathscr{D}_m}\varphi)$ converges to $\Lambda(\cdot,\bar{u})\nabla\varphi$ in $L^2(\Omega)$. Using (13) and the fact that u_m is a solution to (4), we get

$$\begin{aligned}\int_\Omega \Lambda(x,\bar{u})\nabla\bar{u}\cdot\nabla\bar{u}\mathrm{d}x &\leq \liminf_{m\longrightarrow\infty}\Big[\int_\Omega f\Pi_{\mathscr{D}_m}(u_m - P_{\mathscr{D}_m}\varphi)\mathrm{d}x\\ &\qquad + \int_\Omega \Lambda(x,\Pi_{\mathscr{D}_m}u_m)\nabla_{\mathscr{D}_m}u_m\cdot\nabla_{\mathscr{D}_m}(P_{\mathscr{D}_m}\varphi)\mathrm{d}x\Big]\\ &= \int_\Omega f(\bar{u}(x)-\varphi(x)) + \int_\Omega \Lambda(x,\bar{u})\nabla\bar{u}(x)\cdot\nabla\varphi(x)\mathrm{d}x.\end{aligned}$$

This shows that $\bar{u}$ is a weak solution to (3). Now, we prove the strong convergence of the discrete gradients. For a given $v_m \in K_{\mathscr{D}_m}$, we have

$$\begin{aligned}0 &\leq \limsup_{m\longrightarrow\infty} \underline{\lambda}\|\nabla_{\mathscr{D}_m}u_m - \nabla\bar{u}\|^2_{L^2(\Omega)^d}\\ &\leq \limsup_{m\longrightarrow\infty} \int_\Omega \Lambda(x,\Pi_{\mathscr{D}_m}u_m)(\nabla_{\mathscr{D}_m}u_m - \nabla\bar{u})(\nabla_{\mathscr{D}_m}u_m - \nabla\bar{u})\mathrm{d}x\\ &\leq \limsup_{m\longrightarrow\infty}\Big[\int_\Omega f\Pi_{\mathscr{D}_m}(u_m - v_m)\mathrm{d}x + \int_\Omega \Lambda(x,\Pi_{\mathscr{D}_m}u_m)\nabla\bar{u}\cdot\nabla\bar{u}\mathrm{d}x\\ &\quad -2\int_\Omega \Lambda(x,\Pi_{\mathscr{D}_m}u_m)\nabla_{\mathscr{D}_m}u_m\cdot\nabla\bar{u}\mathrm{d}x + \int_\Omega \Lambda(x,\Pi_{\mathscr{D}_m}u_m)\nabla_{\mathscr{D}_m}u_m\cdot\nabla_{\mathscr{D}_m}v_m\mathrm{d}x\Big]\end{aligned}$$

since u_m is a solution to (4). Choosing $v_m = P_{\mathscr{D}_m}\bar{u}$ in this inequality and passing to the limit leads to $\limsup_{m\longrightarrow\infty} \|\nabla_{\mathscr{D}_m}u_m - \nabla\bar{u}\|_{L^2(\Omega)^d} \leq 0$ and concludes the proof. □

References

1. Brezis, H., Stampacchia, G.: Sur la régularité de la solution d'inéquations elliptiques. Bull. Soc. Math. **96**, 153–180 (1968)
2. Droniou, J., Eymard, R., Gallouet, T., Guichard, G., Herbin, R.: Gradient schemes for elliptic and parabolic problems (2014). In preparation
3. Droniou, J., Eymard, R., Gallouët, T., Herbin, R.: Gradient schemes: A generic framework for the discretisation of linear, nonlinear and nonlocal elliptic and parabolic equations. Math. Models Methods Appl. Sci. **23**(13), 2395–2432 (2013)

4. Duvaut, G., Lions, J.: Inequalities in Mechanics and Physics. Springer, Berlin (1976)
5. 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)
6. Falk, R.S.: Error estimates for the approximation of a class of variational inequalities. Math. Comput. **28**(128), 963–971 (1974)
7. Glowinski, R., Lions, J., Tremolieres, R.: Numerical Analysis of Variational Inequalities, 8 edn. North-Holland Publishing Company, Amsterdam (1981)
8. Herbin, R., Marchand, E.: Finite volume approximation of a class of variational inequalities. IMA J. Numer. Anal. **21**(2), 553–585 (2001)
9. Kinderlehrer, D., Stampacchia, G.: An Introduction to Variational Inequalities and their Applications, 8 edn. Academic Press, New York (1980)
10. Zheng, H., Chao Dai, H., Liu, D.F.: A variational inequality formulation for unconfined seepage problems in porous media. Appl. Math. Model. **33**(1), 437–450 (2009)

The Complete Flux Scheme in Cylindrical Coordinates

M. J. H. Anthonissen and J. H. M. ten Thije Boonkkamp

Abstract We consider the complete flux (CF) scheme, a finite volume method (FVM) presented in [3]. CF is based on an integral representation for the fluxes, found by solving a local boundary value problem that includes the source term. It performs well (second order accuracy) for both diffusion and advection dominated problems. In this paper we focus on cylindrically symmetric conservation laws of advection-diffusion-reaction type.

1 Introduction

We consider a stationary conservation law of advection-diffusion-reaction type, viz.

$$\nabla \cdot (\mathbf{u}\varphi - \varepsilon \nabla \varphi) = s, \tag{1}$$

where $\mathbf{u}$ is a mass flux or (drift) velocity, $\varepsilon \geq \varepsilon_{\min} > 0$ a diffusion coefficient, and s a source term describing, e.g., chemical reactions or ionization. The unknown φ is then the mass fraction of one of the constituent species in a chemically reacting flow or a plasma [2]. The parameters ε and s are usually (complicated) functions of φ whereas the vector field $\mathbf{u}$ has to be computed from (flow) equations corresponding to (1). However, for the sake of discretization, we will consider these parameters as given functions of the spatial coordinates.

Associated with Eq. (1) we introduce the flux vector $\mathbf{f}$, defined by $\mathbf{f} := \mathbf{u}\varphi - \varepsilon \nabla \varphi$. Consequently, Eq. (1) can be concisely written as $\nabla \cdot \mathbf{f} = s$. Integrating this equation

M. J. H. Anthonissen (✉) and J. H. M. ten Thije Boonkkamp
Department of Mathematics and Computer Science, Eindhoven University of Technology,
P.O. Box 513, 5600 MBEindhoven, The Netherlands
e-mail: m.j.h.anthonissen@tue.nl

J. H. M. ten Thije Boonkkamp
e-mail: j.h.m.tenthijeboonkkamp@tue.nl

J. Fuhrmann et al. (eds.), *Finite Volumes for Complex Applications VII - Methods and Theoretical Aspects,* Springer Proceedings in Mathematics & Statistics 77,
DOI: 10.1007/978-3-319-05684-5_6,

over a fixed domain Ω and applying Gauss's theorem we obtain the integral form of the conservation law, i.e.,

$$\oint_{\Gamma} (\mathbf{f}, \mathbf{n})\, dS = \int_{\Omega} s\, dV, \tag{2}$$

where $\mathbf{n}$ is the outward unit normal on the boundary $\Gamma = \partial\Omega$. In the FVM [1] we cover the domain with a finite number of disjunct control volumes or cells and impose the integral form (2) for each of these cells.

For cylindrical coordinates (r, θ, z), we assume cylindrical symmetry, i.e., $\varphi = \varphi(r, z)$ and $\mathbf{f} = f_r(r, z)\mathbf{e}_r + f_z(r, z)\mathbf{e}_z$. Equation (1) becomes

$$\frac{1}{r}\frac{\partial}{\partial r}(rf_r) + \frac{\partial}{\partial z}(f_z) = \frac{1}{r}\frac{\partial}{\partial r}\left(r\left(u_r\varphi - \varepsilon\frac{\partial\varphi}{\partial r}\right)\right) + \frac{\partial}{\partial z}\left(u_z\varphi - \varepsilon\frac{\partial\varphi}{\partial z}\right) = s. \tag{3}$$

We choose a uniform tensor product grid with coordinates (r_i, z_j) and grid spacings Δr and Δz. A control volume is the cylindrical shell $\Omega_{i,j} = [r_{i-\frac{1}{2}}, r_{i+\frac{1}{2}}] \times [0, 2\pi) \times [z_{j-\frac{1}{2}}, z_{j+\frac{1}{2}}]$. The surface integral of the flux over the boundary $\Gamma_{i,j} = \partial\Omega_{i,j}$ contains four terms and is given by

$$\oint_{\Gamma_{i,j}} (\mathbf{f}, \mathbf{n})\, dS = \int_{r=r_{i+\frac{1}{2}}} f_r\, dS - \int_{r=r_{i-\frac{1}{2}}} f_r\, dS + \int_{z=z_{j+\frac{1}{2}}} f_z\, dS - \int_{z=z_{j-\frac{1}{2}}} f_z\, dS, \tag{4}$$

where for example $r = r_{i+\frac{1}{2}}$ denotes the interface $\{r_{i+\frac{1}{2}}\} \times \left[0, 2\pi\right) \times \left[z_{j-\frac{1}{2}}, z_{j+\frac{1}{2}}\right]$, and likewise for all other interfaces. All integrals on the right hand side are approximated by the midpoint rule. Taking (2) for $\Omega = \Omega_{i,j}$ and approximating the volume integral for s with the midpoint rule too, we obtain the discrete conservation law

$$\left(r_{i+\frac{1}{2}}F_{r,i+\frac{1}{2},j} - r_{i-\frac{1}{2}}F_{r,i-\frac{1}{2},j}\right)\Delta z + r_i\left(F_{z,i,j+\frac{1}{2}} - F_{z,i,j-\frac{1}{2}}\right)\Delta r = r_i\, s_{i,j}\, \Delta r\, \Delta z, \tag{5}$$

where $F_{r,i+\frac{1}{2},j}$ is the numerical flux approximating $f_r(r_{i+\frac{1}{2}}, z_j)$ and likewise for $F_{z,i,j+\frac{1}{2}}$. The FVM has to be completed with expressions for the numerical flux. The derivation of expressions for the numerical flux is detailed in the next sections.

2 Integral Representation of the Flux

In this section we derive the r-component of the flux in *polar* coordinates by solving a local one-dimensional problem. To determine an integral relation for the flux $f_r := u_r\varphi - \varepsilon\frac{d\varphi}{dr}$ at $r = r_{i+\frac{1}{2}}$, we consider the one-dimensional model BVP:

$$\frac{d}{dr}(rf_r) = rs, \quad r_i < r < r_{i+1}, \qquad \varphi(r_i) = \varphi_i, \quad \varphi(r_{i+1}) = \varphi_{i+1}. \tag{6}$$

We assume $u_r \neq 0$, $\varepsilon > 0$ and s to be sufficiently smooth functions of r.

We introduce the variables U, D, a, A and S by

$$U := ru_r, \quad D := \varepsilon r, \quad a := \frac{U}{D}, \quad A := \int_{r_{i+\frac{1}{2}}}^{r} a(\rho)\,d\rho, \quad S := \int_{r_{i+\frac{1}{2}}}^{r} \rho\, s(\rho)\,d\rho, \tag{7}$$

and integrate (6) from the cell boundary $r_{i+\frac{1}{2}}$ to $r \in (r_i, r_{i+1})$ to find the integral relation $rf_r - (rf_r)_{i+\frac{1}{2}} = S$. Then we rewrite the flux in terms of its integrating factor, viz. $f_r = -\varepsilon\, \mathrm{e}^{A} \frac{d}{dr}\left(\mathrm{e}^{-A}\,\varphi\right)$, substitute it in the integral relation and subsequently integrate over the interval (r_i, r_{i+1}), to arrive at the following expression

$$(rf_r)_{i+\frac{1}{2}} = \underbrace{\frac{\mathrm{e}^{-A_i}\varphi_i - \mathrm{e}^{-A_{i+1}}\varphi_{i+1}}{\langle D^{-1}, \mathrm{e}^{-A}\rangle}}_{(rf_r^{\mathrm{hom}})_{i+\frac{1}{2}}} \underbrace{- \frac{\langle D^{-1}S, \mathrm{e}^{-A}\rangle}{\langle D^{-1}, \mathrm{e}^{-A}\rangle}}_{(rf_r^{\mathrm{inh}})_{i+\frac{1}{2}}}. \tag{8}$$

Here we have used the inner product $\langle f, g\rangle := \int_{r_i}^{r_{i+1}} fg\,dr$. In (8) we have introduced the homogeneous flux f_r^{hom} and the inhomogeneous flux f_r^{inh}. Note that the fluxes correspond to the advection-diffusion operator and the source term, respectively.

Using $A' = a$, we find by straightforward evaluation $\langle a, \mathrm{e}^{-A}\rangle = \mathrm{e}^{-A_i} - \mathrm{e}^{-A_{i+1}}$ and $\langle a, 1\rangle = A_{i+1} - A_i$. Now, we can write the homogeneous flux as

$$\left(rf_r^{\mathrm{hom}}\right)_{i+\frac{1}{2}} = \frac{\langle a, \mathrm{e}^{-A}\rangle / \langle a, 1\rangle}{\langle D^{-1}, \mathrm{e}^{-A}\rangle}\Big(B\big(-\langle a, 1\rangle\big)\varphi_i - B\big(\langle a, 1\rangle\big)\varphi_{i+1}\Big), \tag{9}$$

with $B(x) := \frac{x}{\mathrm{e}^x - 1}$. Note that expression (9) for the homogeneous flux is exact. No approximations have been made so far. If both U and ε are constant, we have

$$A = \frac{U}{\varepsilon}\ln\left(\frac{r}{r_{i+\frac{1}{2}}}\right), \quad \langle a, 1\rangle = \frac{U}{\varepsilon}\ln\left(\frac{r_{i+1}}{r_i}\right), \quad \langle D^{-1}, \mathrm{e}^{-A}\rangle = \frac{1}{U}\langle a, \mathrm{e}^{-A}\rangle, \tag{10}$$

and (9) reduces to the *constant coefficient flux*

$$\left(rf_r^{\mathrm{hom}}\right)_{i+\frac{1}{2}} = \frac{\varepsilon}{\ln\left(r_{i+1}/r_i\right)}\Big(B(-P)\varphi_i - B(P)\varphi_{i+1}\Big), \quad P := \frac{U}{\varepsilon}\ln\left(\frac{r_{i+1}}{r_i}\right). \tag{11}$$

Here we have introduced the Péclet number P.

Let us now consider the inhomogeneous flux. We first consider the general case, so nonconstant U and ε. The denominator in (8) for $\left(rf^{\text{inh}}\right)_{i+\frac{1}{2}}$ can be written as

$$\langle D^{-1}, \mathrm{e}^{-A}\rangle = \langle \frac{a}{U}, \mathrm{e}^{-A}\rangle = \frac{1}{U^*}\langle a, \mathrm{e}^{-A}\rangle = -\frac{1}{U^*}\mathrm{e}^{-A_i}\left(\mathrm{e}^{-\langle a,1\rangle} - 1\right), \tag{12}$$

where $U^* = U(\xi)$ for some unknown $\xi \in (r_i, r_{i+1})$. For the numerator in (8), we substitute the expression for S and change the order of integration, viz.

$$\begin{aligned}\langle D^{-1}S, \mathrm{e}^{-A}\rangle = &-\int_{r_i}^{r_{i+\frac{1}{2}}}\int_{r_i}^{\rho} D(r)^{-1}\mathrm{e}^{-A(r)}\,dr\,\rho\,s(\rho)\,d\rho \\ &+\int_{r_{i+\frac{1}{2}}}^{r_{i+1}}\int_{\rho}^{r_{i+1}} D(r)^{-1}\mathrm{e}^{-A(r)}\,dr\,\rho\,s(\rho)\,d\rho.\end{aligned} \tag{13}$$

Carrying out the inner integrations over r, analogous to (12), we have

$$\left(rf_r^{\text{inh}}\right)_{i+\frac{1}{2}} = \frac{U^*}{U_1^*}\int_{r_i}^{r_{i+\frac{1}{2}}}\frac{\mathrm{e}^{-\int_{r_i}^{\rho} a\,dr} - 1}{\mathrm{e}^{-\langle a,1\rangle} - 1}\rho\,s(\rho)\,d\rho - \frac{U^*}{U_2^*}\int_{r_{i+\frac{1}{2}}}^{r_{i+1}}\frac{\mathrm{e}^{-\int_{\rho}^{r_{i+1}} a\,dr} - 1}{\mathrm{e}^{\langle a,1\rangle} - 1}\rho\,s(\rho)\,d\rho, \tag{14}$$

with $U_1^* = U(\xi_1)$ for some unknown $\xi_1 \in (r_i, r_{i+\frac{1}{2}})$ and $U_2^* = U(\xi_2)$ for some unknown $\xi_2 \in (r_{i+\frac{1}{2}}, r_{i+1})$. Let us consider again the constant coefficient case, so assume that both U and ε are constant. Expression (14) can be simplified by introducing the *normalized coordinate*

$$\sigma(r) := \frac{1}{\langle a, 1\rangle}\int_{r_i}^{r} a\,d\rho, \qquad r_i \le r \le r_{i+1}. \tag{15}$$

For both positive and negative values of u_r (and hence a) on (r_i, r_{i+1}), we have $d\sigma/dr > 0$. Therefore the normalized coordinate is an increasing function of r. It satisfies $0 \le \sigma \le 1$. Next we define the *Green's function for the flux* by

$$G(\sigma; P) := \begin{cases} \dfrac{\mathrm{e}^{-\sigma P} - 1}{\mathrm{e}^{-P} - 1}, & \text{for } 0 \le \sigma \le \sigma(r_{i+\frac{1}{2}}), \\ -\dfrac{\mathrm{e}^{(1-\sigma)P} - 1}{\mathrm{e}^{P} - 1}, & \text{for } \sigma(r_{i+\frac{1}{2}}) < \sigma \le 1. \end{cases} \tag{16}$$

With (15) and (16), Eq. (14) simplifies to

$$\left(r f_r^{\text{inh}}\right)_{i+\frac{1}{2}} = \int_0^1 G(\sigma;\langle a,1\rangle)\, r(\sigma)\, s(r(\sigma))\, \frac{\langle a,1\rangle}{a(r(\sigma))}\, d\sigma. \tag{17}$$

If $u_r = 0$ on (r_i, r_{i+1}), then also $a = P = 0$. In this case, we can define the normalized coefficient σ by replacing a with D^{-1} in (15). This leads to a similar expression as (17) and will give the same numerical flux.

3 Numerical Flux

For the numerical fluxes, we take U and ε constant on each control volume, viz. $\overline{U}_{i+\frac{1}{2}} := (U_i + U_{i+1})/2$ and $\overline{\varepsilon}_{i+\frac{1}{2}} := (\varepsilon_i + \varepsilon_{i+1})/2$. We use (11) to find the following *numerical homogeneous flux*

$$\left(r F_r^{\text{hom}}\right)_{i+\frac{1}{2}} = \alpha_{r,i+\frac{1}{2}}\varphi_i - \beta_{r,i+\frac{1}{2}}\varphi_{i+1}, \tag{18a}$$

with

$$\begin{aligned}
&\alpha_{r,i+\frac{1}{2}} := B\big(-P_{r,i+\frac{1}{2}}\big)\frac{\overline{\varepsilon}_{i+\frac{1}{2}}}{\ln(r_{i+1}/r_i)}, \quad \beta_{r,i+\frac{1}{2}} := B\big(P_{r,i+\frac{1}{2}}\big)\frac{\overline{\varepsilon}_{i+\frac{1}{2}}}{\ln(r_{i+1}/r_i)},\\
&P_{r,i+\frac{1}{2}} := \frac{\overline{U}_{i+\frac{1}{2}}}{\overline{\varepsilon}_{i+\frac{1}{2}}}\ln\left(\frac{r_{i+1}}{r_i}\right).
\end{aligned} \tag{18b}$$

The *numerical inhomogeneous flux* is based on (17). We make the approximation $\langle a,1\rangle/a(r(\sigma)) \doteq \Delta r$, and take $rs(r)$ equal to $r_i s_i$ on the interval $(0, \sigma(r_{i+\frac{1}{2}}))$ and equal to $r_{i+1}s_{i+1}$ on $(\sigma(r_{i+\frac{1}{2}}), 1)$. Next we integrate the Green's function to find

$$\left(r F_r^{\text{inh}}\right)_{i+\frac{1}{2}} := \gamma_{r,i+\frac{1}{2}} s_i - \delta_{r,i+\frac{1}{2}} s_{i+1}, \tag{19a}$$

with

$$\begin{aligned}
\gamma_{r,i+\frac{1}{2}} &:= r_i\,\Delta r \int_0^{\sigma_{i+\frac{1}{2}}} G(\sigma;\, P_{r,i+\frac{1}{2}})\, d\sigma = C(-P_{r,i+\frac{1}{2}};\, \sigma_{i+\frac{1}{2}})\, r_i\,\Delta r,\\
\delta_{r,i+\frac{1}{2}} &:= r_{i+1}\,\Delta r \int_{\sigma_{i+\frac{1}{2}}}^1 G(\sigma;\, P_{r,i+\frac{1}{2}})\, d\sigma = C(P_{r,i+\frac{1}{2}};\, 1-\sigma_{i+\frac{1}{2}})\, r_{i+1}\,\Delta r,\\
\sigma_{i+\frac{1}{2}} &:= \frac{\ln(r_{i+\frac{1}{2}}/r_i)}{\ln(r_{i+1}/r_i)}, \qquad C(x;\sigma) := \frac{\mathrm{e}^{\sigma x} - 1 - \sigma x}{x\,(\mathrm{e}^x - 1)}.
\end{aligned} \tag{19b}$$

It is easily verified that $\sigma_{i+\frac{1}{2}} \to \frac{1}{2}$ for $\Delta r \to 0$ when $r_i > 0$. The function C is plotted as a function of x for $\sigma = \frac{1}{2}$ in Fig. 1. Note that $C(x;\sigma) \to \sigma^2/2$ for $x \to 0$,

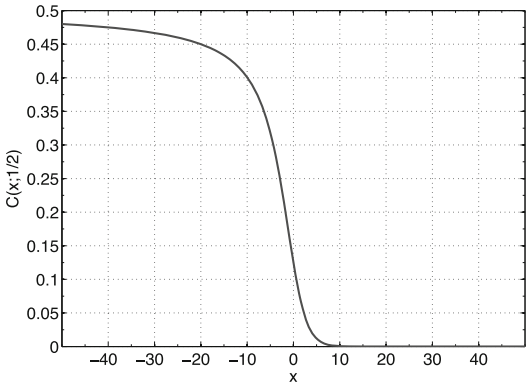


Fig. 1 Graph of the function $C(x;\sigma)$ as a function of x for $\sigma = \frac{1}{2}$. Note that $C(0;\frac{1}{2}) = \frac{1}{8}$, $C(x;\frac{1}{2}) \to 0$ for $x \to \infty$, and $C(x;\frac{1}{2}) \to \frac{1}{2}$ for $x \to -\infty$

$C(x;\sigma) \to 0$ for $x \to \infty$, and finally $C(x;\sigma) \to \sigma$ for $x \to -\infty$. This means that for small Péclet numbers, the coefficients γ and δ will be approximately equal and the inhomogeneous flux is small. For large (positive or negative) Péclet numbers, the upwind value of s has a dominant contribution to the flux. This approach is similar to the modified inhomogeneous flux scheme for Cartesian coordinates in [2].

Adding (18) and (19), we obtain the following *numerical complete flux*:

$$\left(rF_r\right)_{i+\frac{1}{2}} = \alpha_{r,i+\frac{1}{2}}\varphi_i - \beta_{r,i+\frac{1}{2}}\varphi_{i+1} + \gamma_{r,i+\frac{1}{2}}s_i - \delta_{r,i+\frac{1}{2}}s_{i+1}. \tag{20}$$

4 Extension to Two-Dimensional Conservation Laws

Up till now we have considered the numerical flux for the r-component only (radial fluxes). The derivation of the flux in z-direction is similar. It is in fact the flux in Cartesian coordinates and a detailed derivation can be found in [3]. Here we only present the numerical flux. We have (cf. (18)–(20))

$$F_{z,j+\frac{1}{2}} = \alpha_{z,j+\frac{1}{2}}\varphi_j - \beta_{z,j+\frac{1}{2}}\varphi_{j+1} + \gamma_{z,j+\frac{1}{2}}s_j - \delta_{z,j+\frac{1}{2}}s_{j+1}, \tag{21a}$$

with

$$\begin{aligned}
&\alpha_{z,j+\frac{1}{2}} := B(-\overline{P}_{z,j+\frac{1}{2}})\frac{\widetilde{P}_{z,j+\frac{1}{2}}}{\overline{P}_{z,j+\frac{1}{2}}}\frac{\widetilde{\varepsilon}_{j+\frac{1}{2}}}{\Delta z}, \quad \beta_{z,j+\frac{1}{2}} := B(\overline{P}_{z,j+\frac{1}{2}})\frac{\widetilde{P}_{z,j+\frac{1}{2}}}{\overline{P}_{z,j+\frac{1}{2}}}\frac{\widetilde{\varepsilon}_{j+\frac{1}{2}}}{\Delta z},\\
&\gamma_{z,j+\frac{1}{2}} := C(-\overline{P}_{z,j+\frac{1}{2}};\tfrac{1}{2})\,\Delta z, \quad \delta_{z,j+\frac{1}{2}} := C(\overline{P}_{z,j+\frac{1}{2}};\tfrac{1}{2})\,\Delta z, \quad P_z := \frac{u_z\,\Delta z}{\varepsilon}.
\end{aligned} \tag{21b}$$

Two averages are used in these expressions, the normal *arithmetic* mean $\overline{\varepsilon}_{j+\frac{1}{2}} := (\varepsilon_j + \varepsilon_{j+1})/2$ and a *weighted* average $\widetilde{\varepsilon}_{j+\frac{1}{2}} := W(-\overline{P}_{z,j+\frac{1}{2}})\varepsilon_j + W(\overline{P}_{z,j+\frac{1}{2}})\varepsilon_{j+1}$.

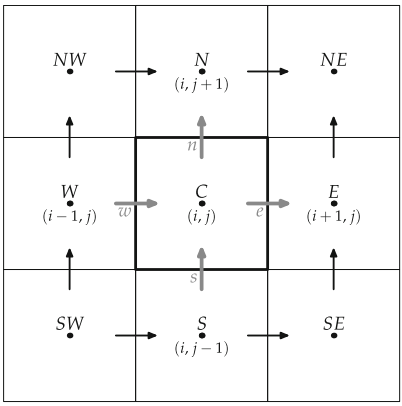


Fig. 2 Control volume Ω_C and corresponding stencil

The weight function used here is $W(x) := (e^x - 1 - x)/(x\,(e^x - 1))$. Note that the expressions for the coefficients γ and δ in (21b) are different from those in [3]. These new coefficients coincide with the old for large Péclet numbers, but are more accurate if advection is not dominant.

We will now combine the one-dimensional schemes to derive a numerical scheme for the two-dimensional equation (1). For ease of presentation, we use both index notation and compass notation; see Fig. 2. Thus, φ_C should be understood as $\varphi_{i,j}$ and $f_{r,e}$ as $f_{r,i+\frac{1}{2},j}$ etc. The key idea is to include the cross flux term $\partial f_z/\partial z$ in the evaluation of the flux in r-direction. Therefore we determine the numerical flux $F_{r,i+\frac{1}{2},j}$ from the quasi-one-dimensional boundary value problem:

$$\frac{\partial}{\partial r}\left(r\left(u_r\varphi - \varepsilon\frac{\partial\varphi}{\partial r}\right)\right) = r s_r, \qquad r_i < r < r_{i+1},\ z = z_j, \tag{22a}$$

$$\varphi(\mathbf{x}_{i,j}) = \varphi_{i,j}, \qquad \varphi(\mathbf{x}_{i+1,j}) = \varphi_{i+1,j}, \tag{22b}$$

where the modified source term s_r is defined by $s_r := s - \frac{\partial f_z}{\partial z}$. The derivation of the expression for the numerical flux is essentially the same as for (20), the main difference being the inclusion of the cross flux term $\partial f_z/\partial z$ in the source term. In the computation of s_r we replace $\partial f_z/\partial z$ by its central difference approximation and for f_z we take the homogeneous numerical flux. A similar procedure applies to the z-component of the flux. This leads to the following algorithm.

Algorithm for the computation of the numerical fluxes

1. Compute averages and Péclet numbers
   - in r-direction: $\overline{U}_e = \frac{U_C+U_E}{2}$, $\overline{\varepsilon}_e = \frac{\varepsilon_C+\varepsilon_E}{2}$, $P_{r,e} = \frac{\overline{U}_e}{\overline{\varepsilon}_e}\ln\left(\frac{r_E}{r_C}\right)$
   - in z-direction: $P_z = \frac{u_z\,\Delta z}{\varepsilon}$, $\overline{P}_{z,n} = \frac{P_{z,C}+P_{z,N}}{2}$, $\widetilde{\varepsilon}_n := W(-\overline{P}_{z,n})\varepsilon_C + W(\overline{P}_{z,n})\varepsilon_N$, $\widetilde{P}_{z,n} = W(-\overline{P}_{z,n})P_{z,C} + W(\overline{P}_{z,n})P_{z,N}$
2. Numerical homogeneous flux

- in r-direction: $\left(r F_r^{\text{hom}}\right)_e = \alpha_{r,e}\varphi_C - \beta_{r,e}\varphi_E$ with $\alpha_{r,e} = B(-P_{r,e})\frac{\overline{\varepsilon}_e}{\ln(r_E/r_C)}$, $\beta_{r,e} = B(P_{r,e})\frac{\overline{\varepsilon}_e}{\ln(r_E/r_C)}$
- in z-direction: $F_{z,n}^{\text{hom}} = \alpha_{z,n}\varphi_C - \beta_{z,n}\varphi_N$ with $\alpha_{z,n} = B(-\overline{P}_{z,n})\frac{\widetilde{P}_{z,n}}{\overline{P}_{z,n}}\frac{\widetilde{\varepsilon}_n}{\Delta z}$, $\beta_{z,n} = B(\overline{P}_{z,n})\frac{\widetilde{P}_{z,n}}{\overline{P}_{z,n}}\frac{\widetilde{\varepsilon}_n}{\Delta z}$

3. Numerical inhomogeneous flux
   - in r-direction: $\left(r F_r^{\text{inh}}\right)_e = \gamma_{r,e}s_{r,C} - \delta_{r,e}s_{r,E}$ with $\gamma_{r,e} = C\big(-P_{r,e};\sigma_e\big)\,r_C\,\Delta r$, $\delta_{r,e} = C\big(P_{r,e};1-\sigma_e\big)\,r_E\,\Delta r$, $\sigma_e = \frac{\ln(r_e/r_C)}{\ln(r_E/r_C)}$, $s_{r,C} = s_C - \frac{1}{\Delta z}\left(F_{z,n}^{\text{hom}} - F_{z,s}^{\text{hom}}\right)$
   - in z-direction: $F_{z,n}^{\text{inh}} = \gamma_{z,n}s_{z,C} - \delta_{z,n}s_{z,N}$ with $\gamma_{z,n} = C\big(-\overline{P}_{z,n};\frac{1}{2}\big)\,\Delta z$, $\delta_{z,n} = C\big(\overline{P}_{z,n};\frac{1}{2}\big)\,\Delta z$, $s_{z,C} = s_C - \frac{1}{r_C\,\Delta r}\left(\left(r F_r^{\text{hom}}\right)_e - \left(r F_r^{\text{hom}}\right)_w\right)$
4. Numerical complete flux
   - in r-direction: $\left(r F_r\right)_e = \left(r F_r^{\text{hom}}\right)_e + \left(r F_r^{\text{inh}}\right)_e$
   - in z-direction: $F_{z,n} = F_{z,n}^{\text{hom}} + F_{z,n}^{\text{inh}}$

Writing the discrete conservation law (5) in compass notation, we find

$$\Big(\left(r F_r\right)_e - \left(r F_r\right)_w\Big)\Delta z + r_C\big(F_{z,n} - F_{z,s}\big)\Delta r = r_C\,s_C\,\Delta r\,\Delta z. \tag{23}$$

Substitution of the numerical fluxes presented above leads to a 9-point stencil for the unknown φ. The complete flux scheme reduces to the homogeneous flux scheme if we set all coefficients $\gamma_{*,*}$ and $\delta_{*,*}$ to zero.

5 Numerical Experiments

We study the following model problem to test the accuracy of the new complete flux (CF) scheme and to compare it with the homogeneous flux (HF) scheme. The problem domain is given by $1 \le r \le 4$, $0 \le z \le 3$. The unknown φ satisfies the partial differential equation $\nabla\cdot(\mathbf{u}\varphi - \varepsilon\nabla\varphi) = s$ in $(1,4)\times(0,3)$. We take $\mathbf{u}(r,z) = u_r\mathbf{e}_r + u_z\mathbf{e}_z = \frac{2}{r}\mathbf{e}_r + 3\mathbf{e}_z$. We impose Dirichlet boundary conditions and choose the source term s such that the analytical solution is given by $\varphi(r,z) = r^2 + 2r + 3z^2 + 4z + 5$. We discretize the PDE on a uniform grid (r_i, z_j) with N_r grid points in r-direction and N_z points in z-direction.

Numerical results are presented in Table 1. The error provided is the infinity-norm, so $e := \max_{i,j}\left|\varphi(r_i,z_j) - \varphi_{i,j}\right|$, with $\varphi_{i,j}$ the numerical approximation computed using the CF or HF scheme. The columns labelled 'quotient' list the quotient of the errors on successive grids. Both the HF and CF schemes show second order accuracy for dominant diffusion; HF reduces to first order for dominant advection. Note that the error of the CF scheme is a factor 2 smaller for dominant diffusion, and it is orders of magnitude more accurate for dominant advection.

Table 1 Numerical results for dominant diffusion ($\varepsilon = 10^8$) and dominant advection ($\varepsilon = 10^{-8}$)

	$\varepsilon = 10^8$				$\varepsilon = 10^{-8}$			
$N_r = N_z$	Error HF	Quotient	Error CF	Quotient	Error HF	Quotient	Error CF	Quotient
6	4.6862e-03		2.0625e-03		4.5491e+00		7.6891e-01	
11	1.2442e-03	3.766	6.0227e-04	3.425	2.5988e+00	1.750	2.0921e-01	3.675
21	3.1414e-04	3.961	1.5588e-04	3.864	1.3886e+00	1.872	5.4816e-02	3.817
41	7.9087e-05	3.972	3.9466e-05	3.950	7.2017e-01	1.928	1.4213e-02	3.857
81	1.9783e-05	3.998	9.8865e-06	3.992	3.6794e-01	1.957	3.6174e-03	3.929
161	4.9464e-06	3.999	2.4729e-06	3.998	1.8647e-01	1.973	9.1393e-04	3.958
321	1.2366e-06	4.000	6.1830e-07	3.999	9.4064e-02	1.982	2.2996e-04	3.974
641	3.0916e-07	4.000	1.5459e-07	4.000	4.7314e-02	1.988	5.7721e-05	3.984

References

1. Eymard, R., Gallouët, T., Herbin, R.: Finite volume methods. In: Ciarlet, P.G., Lions, J.L. (eds.) Handbook of Numerical Analysis, vol. VII, pp. 713–1020. North-Holland, Amsterdam (2000)
2. Liu, L.: Studies on the discretization of plasma transport equations. Ph.D. thesis, Eindhoven University of Technology, Eindhoven (2013)
3. ten Thije Boonkkamp, J.H.M., anthonissen, M.J.H.: The finite volume-complete flux scheme for advection-diffusion-reaction equations. J. Sci. Comput. **46**(1), 47–70 (2011)

A Staggered Scheme with Non-conforming Refinement for the Navier-Stokes Equations

Fabrice Babik, Jean-Claude Latché, Bruno Piar and Khaled Saleh

Abstract We propose a numerical scheme for the incompressible Navier-Stokes equations. The pressure is approximated at the cell centers while the vector valued velocity degrees of freedom are localized at the faces of the cells. The scheme is able to cope with unstructured non-conforming meshes, involving hanging nodes. The discrete convection operator, of finite volume form, is built with the purpose to obtain an L^2-stability property, or, in other words, a discrete equivalent to the kinetic energy identity. The diffusion term is approximated by extending the usual Rannacher-Turek finite element to non-conforming meshes. The scheme is first order in space for energy norms, as shown by the numerical experiments.

1 Introduction

Let Ω be an open bounded connected subset of $\mathbb{R}^d$, with $d \in \{2, 3\}$, which is supposed to be polygonal if $d = 2$ and polyhedral if $d = 3$. Let $T \in \mathbb{R}^+$. We address in this paper the system of incompressible Navier-Stokes equations:

K. Saleh (✉), J.-C. Latché, B. Piar and F. Babik
Institut de radioprotection et de sûreté nucléaire (IRSN), PSN-RES, St. Paul-lez-Durance, France
e-mail: khaled.saleh@irsn.fr

J.-C. Latché
e-mail: jean-claude.latche@irsn.fr

B. Piar
e-mail: bruno.piar@irsn.fr

F. Babik
e-mail: fabrice.babik@irsn.fr

J. Fuhrmann et al. (eds.), *Finite Volumes for Complex Applications VII - Methods and Theoretical Aspects*, Springer Proceedings in Mathematics & Statistics 77,
DOI: 10.1007/978-3-319-05684-5_7,

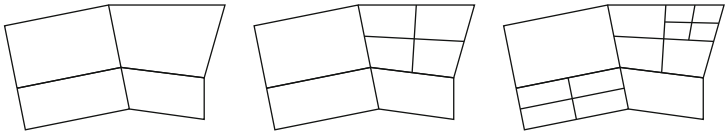


Fig. 1 An exemple of admissible mesh refinement

$$\partial_t \mathbf{u} + \mathbf{div}(\mathbf{u} \otimes \mathbf{u}) - \mu \boldsymbol{\Delta} \mathbf{u} + \boldsymbol{\nabla} p = 0, \qquad \text{on } \Omega \times (0, T), \tag{1a}$$

$$\mathrm{div}\mathbf{u} = 0, \qquad \text{on } \Omega \times (0, T), \tag{1b}$$

$$\mathbf{u}|_{\partial\Omega} = \mathbf{u}_{\partial\Omega}, \qquad \mathbf{u}|_{t=0} = \mathbf{u}_0. \tag{1c}$$

The variables $\mathbf{u} \in \mathbb{R}^d$ and $p \in \mathbb{R}$ are the velocity and the pressure in the flow, and μ is a positive constant viscosity. The initial condition $\mathbf{u}_0$ is supposed to be divergence-free, and the integral of $\mathbf{u}_{\partial\Omega} \cdot \mathbf{n}_{\partial\Omega}$ over $\partial\Omega$ vanishes, where $\mathbf{n}_{\partial\Omega}$ stands for the normal vector to $\partial\Omega$ outward Ω.

We develop in this paper a projection scheme to approximate the solution of (1), based on a staggered space discretization and able to cope with non-conforming mesh refinement. During the last years, a research program has been undertaken to develop staggered schemes satisfying a discrete kinetic energy balance [1, 5]. This point is crucial with respect to many issues: it readily provides stability estimates, a property which is a prerequisite for LES applications, and, last but not least, it is a starting point for the extension of the schemes to compressible flows (shallow water, compressible Navier-Stokes and Euler equations). The difficulty lies in the definition of the velocity convection operator, which must be in some sense consistent with the discrete mass balance; the definition of this operator is thus intricate and, to our knowledge, novel, at least for density variable flows. The objective of the present paper is to show how to extend this definition to non-conforming meshes. We first define admissible meshes (Sect. 2), then describe the scheme (Sect. 3) and finally present some numerical experiments to assess its behavior (Sect. 4).

2 Definition of the Meshes

Let $\mathcal{M}$ be a decomposition of the domain Ω either in convex quadrilaterals ($d = 2$) or hexahedra ($d = 3$). The mesh $\mathcal{M}$ is supposed to be obtained from a usual finite element regular discretization (e.g. [4]) by recursively splitting some cells in 2^d subcells obtained by joining every two opposite face centers (Fig. 1). We allow at most one hanging node at the mass center of a cell face, which means that the maximum level of refinement between two adjacent cells is one.

We denote by $\mathcal{E}(K)$ the set of the faces of an element $K \in \mathcal{M}$. We exclude the presence of a node in the interior of a face, i.e. we split an initial face in 2^{d-1} faces if one of the cells adjacent to the face is split. The number of faces, $N_K^{\mathcal{E}}$, of a cell K

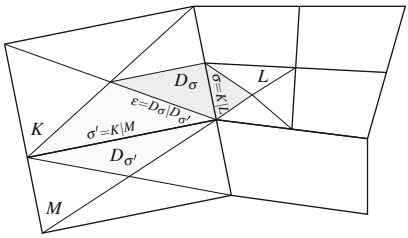


Fig. 2 Notations for control volumes and diamond cells

thus ranges between $2d$ and $2^d d$. Let $\mathscr{E} = \cup_{K\in\mathscr{M}}\mathscr{E}(K)$, $\mathscr{E}_{\rm ext} = \{\sigma \in \mathscr{E}, \sigma \subset \partial\Omega\}$ and $\mathscr{E}_{\rm int} = \mathscr{E} \setminus \mathscr{E}_{\rm ext}$. A face $\sigma \in \mathscr{E}_{\rm int}$ separating the cells K and L is denoted by $K|L$. For $\sigma \in \mathscr{E}(K)$, $\mathbf{n}_{K,\sigma}$ is the unit normal vector to σ outward K. Hereafter, $|\cdot|$ stands for the d- or $(d-1)$-dimensional measure of a subset of $\mathbb{R}^d$ or $\mathbb{R}^{d-1}$ respectively.

We define a dual mesh associated with the faces $\mathscr{E}$ as follows. When $K \in \mathscr{M}$ is a rectangle or a cuboid, for $\sigma \in \mathscr{E}(K)$, we define the half-diamond cell $D_{K,\sigma}$ as the cone with basis σ and with vertex the mass center of K (see Fig. 2). We thus obtain a partition of K in $N_K^{\mathscr{E}}$ sub-volumes, each sub-volume having a measure $|D_{K,\sigma}|$ equal to $|K|/(2d)$, when σ has not been split, or $|K|/(2^d d)$ otherwise. We extend this definition to general quadrangles and hexahedra, by supposing that we have built a partition with the same connectivities and the same ratio between the volumes of the half-diamonds and of the cell. For $\sigma \in \mathscr{E}_{\rm int}$, $\sigma = K|L$, we now define the dual (or diamond) cell D_σ associated with σ by $D_\sigma = D_{K,\sigma} \cup D_{L,\sigma}$. For $\sigma \in \mathscr{E}(K) \cap \mathscr{E}_{\rm ext}$, we define $D_\sigma = D_{K,\sigma}$. We denote by $\tilde{\mathscr{E}}(D_\sigma)$ the set of faces of D_σ, and by $\varepsilon = D_\sigma|D_{\sigma'}$ the face separating two dual cells D_σ and $D_{\sigma'}$ (see Fig. 2).

3 The Pressure Correction Scheme

3.1 General Form of the Scheme

The space discretization is staggered in the sense that the pressure and the velocity are piecewise constant functions respectively on the primal and dual mesh. The initial discrete velocity is defined on a dual cell D_σ, $\sigma \in \mathscr{E}_{\rm int}$, by the mean value $\mathbf{u}_\sigma^0$ of the function $\mathbf{u}_0$ over the face σ. The Dirichlet boundary condition is taken into account by setting $\mathbf{u}_\sigma^n$ to the mean value of $\mathbf{u}_{\partial\Omega}$ over σ, for all $\sigma \in \mathscr{E}_{\rm ext}$ and all $n \geq 0$. We consider a constant time step δt. As usual [3, 6, 9], the projection scheme is a two-step algorithm:

$$\textit{Prediction step} - \text{Find}(\mathbf{u}_\sigma^\star)_{\sigma\in\mathscr{E}_{\rm int}} \text{ such that:}$$
$$\frac{1}{\delta t}(\mathbf{u}_\sigma^\star - \mathbf{u}_\sigma^n) + \frac{1}{|D_\sigma|}\sum_{\varepsilon\in\tilde{\mathscr{E}}(D_\sigma)} F_{\sigma,\varepsilon}^n \mathbf{u}_\varepsilon^\star - \mu\,(\boldsymbol{\Delta}\mathbf{u})_\sigma^\star + (\boldsymbol{\nabla} p)_\sigma^n = 0, \qquad \sigma \in \mathscr{E}_{\rm int}, \tag{2a}$$

Correction step – Find $(\mathbf{u}_\sigma^{n+1})_{\sigma\in\mathscr{E}_{\rm int}}$ and $(p_K^{n+1})_{K\in\mathscr{M}}$ such that:

$$\frac{1}{\delta t}(\mathbf{u}_\sigma^{n+1}-\mathbf{u}_\sigma^\star)+(\nabla p)_\sigma^{n+1}-(\nabla p)_\sigma^n=0, \qquad \sigma\in\mathscr{E}_{\rm int}, \quad (2b)$$

$$\sum_{\sigma\in\mathscr{E}(K)} F_{K,\sigma}^{n+1}=0, \quad \text{with } F_{K,\sigma}^{n+1}=|\sigma|\,\mathbf{u}_\sigma^{n+1}\cdot\mathbf{n}_{K,\sigma}, \qquad K\in\mathscr{M}. \quad (2c)$$

The pressure gradient is built as the dual operator of the discrete divergence:

$$(\nabla p)_\sigma=\frac{|\sigma|}{|D_\sigma|}(p_L-p_K)\,\mathbf{n}_{K,\sigma}, \qquad \sigma=K|L.$$

The discretization of the diffusion term relies on the so-called "rotated bi-linear element" introduced by Rannacher and Turek [8]. The reference element $\widehat{K}$ is the unit d-cube $(0,1)^d$, and the discrete functional space is:

$$\tilde{Q}_1(\widehat{K})=\text{span}\left\{1,\ (\mathbf{x}_i)_{i=1,\dots,d},\ (\mathbf{x}_i^2-\mathbf{x}_{i+1}^2)_{i=1,\dots,d-1}\right\}.$$

When there is no hanging node on a face σ, we impose the jump through the face to have a zero mean value. When there is a hanging node, we only impose to zero the integral of the jump through the initial face. Hence, the set $\{\zeta_\sigma,\ \sigma\in\mathscr{E}_{\rm int}\}$ of nodal functions associated with the Rannacher-Turek element is defined as follows. When no vertices of $\sigma=K|L$ is a hanging node, we define ζ_σ such that $\text{supp}(\zeta_\sigma)\subset K\cup L$, for all $K\in\mathscr{M}$, $\zeta_\sigma|_K$ belongs to the Rannacher-Turek local discrete space of K (i.e. the image of the space $\tilde{Q}_1(\widehat{K})$ by the Q_1 mapping) and:

$$\frac{1}{|\sigma|}\int_\sigma\zeta_\sigma=1 \text{ and, for all } \sigma'\in\mathscr{E},\ \sigma'\neq\sigma,\ \int_{\sigma'}\zeta_\sigma=0. \quad (3)$$

When one of the vertices of $\sigma=K|L$ is a hanging node, it means that σ separates a cell obtained by splitting the mesh, say L, from an unsplit one, say K. The support of ζ_σ is still $K\cup L$ and on L, ζ_σ is still given by (3). Let Σ be the initial face of K including σ, and let ζ_Σ be the Rannacher-Turek usual shape function (i.e. the function satisfying an analogue of (3) on the initial mesh). Then, we define ζ_σ on K by $\zeta_\sigma(\mathbf{x})=\dfrac{|\sigma|}{|\Sigma|}\zeta_\Sigma(\mathbf{x})$.

Finally, dropping the time index, the discretization of the diffusion term reads:

$$-(\boldsymbol{\Delta}\mathbf{u})_\sigma=\frac{1}{|D_\sigma|}\sum_{K\in\mathscr{M}}\int_K\sum_{\sigma'\in\mathscr{E}(K)}\mathbf{u}_{\sigma'}(\nabla\zeta_{\sigma'}\cdot\nabla\zeta_\sigma). \quad (4)$$

In the convection term, the velocity interpolates at the internal dual faces $\mathbf{u}_\varepsilon^\star$ is chosen centered: $\mathbf{u}_\varepsilon=(\mathbf{u}_\sigma+\mathbf{u}_{\sigma'})/2$, for $\varepsilon=D_\sigma|D'_\sigma$. To make the description of

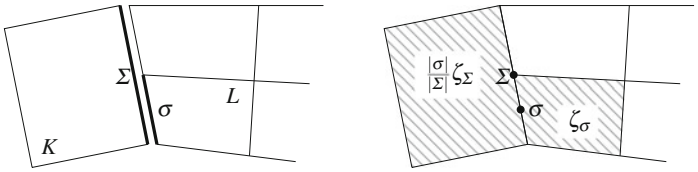


Fig. 3 Piecewise definition of ζ_σ

the scheme complete, we now only need to define the mass fluxes through the dual faces $(F^n_{\sigma,\varepsilon})_{\varepsilon\in\tilde{\mathscr{E}}}$: this is the purpose of Sect. 3.2.

3.2 Discrete Kinetic Energy and Mass Fluxes

The discrete mass fluxes through the faces of the dual mesh are built so that a finite volume discretization of the divergence constraint (1b) holds over the dual cells:

$$\sum_{\varepsilon\in\tilde{\mathscr{E}}(D_\sigma)} F^n_{\sigma,\varepsilon} = 0, \qquad \sigma \in \mathscr{E}_{\rm int}. \tag{5}$$

This is crucial in order to reproduce, at the discrete level, the derivation of a kinetic energy balance equation, thus ensuring discrete analogues of the usual $\mathrm{L}^\infty(\mathrm{L}^2)$- and $\mathrm{L}^2(\mathrm{H}^1)$- stability estimates for the velocity. It may be shown that Relation (5) holds if the dual fluxes are computed from the primal ones $(F^n_{K,\sigma})_{\sigma\in\mathscr{E}(K)}$ at the previous time-step so as to satisfy the following three constraints (see [1, 5] for details):

- *(H1)*—For all primal cell K in $\mathscr{M}$, the set $(F_{\sigma,\varepsilon})_{\varepsilon\subset K}$ of dual fluxes through faces included in K satisfies the following linear system, with $\xi^\sigma_K = |D_{K,\sigma}|/|K|$:

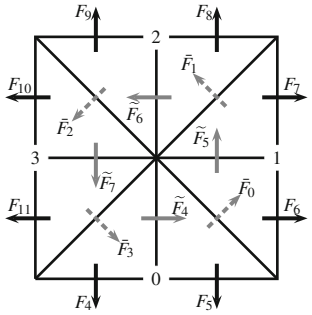
$$F_{K,\sigma} + \sum_{\varepsilon\in\tilde{\mathscr{E}}(D_\sigma),\ \varepsilon\subset K} F_{\sigma,\varepsilon} = \xi^\sigma_K \sum_{\sigma'\in\mathscr{E}(K)} F_{K,\sigma'}, \qquad \forall\sigma\in\mathscr{E}(K). \tag{6}$$

- *(H2)*—The dual fluxes are conservative: $F_{\sigma,\varepsilon} = -F_{\sigma',\varepsilon}$ for all $\varepsilon = D_\sigma|D'_\sigma$.
- *(H3)*—The dual fluxes are a bounded function of the primal ones $(F_{K,\sigma})_{\sigma\in\mathscr{E}(K)}$:

$$|F_{\sigma,\varepsilon}| \le C\max\left\{|F_{K,\sigma}|,\ \sigma\in\mathscr{E}(K)\right\}, \qquad K\in\mathscr{M},\ \sigma\in\mathscr{E}(K),\ \varepsilon\in\tilde{\mathscr{E}}(D_\sigma),\ \varepsilon\subset K.$$

3.2.1 Dual Fluxes for Non-refined Meshes

The system of equations (6) has an infinity of solutions, which makes necessary to impose in addition the constraint (H3). Since (6) is linear with respect to the $F_{\sigma,\varepsilon},\ \sigma\in\mathscr{E}(K), \varepsilon\in\tilde{\mathscr{E}}(D_\sigma),\ \varepsilon\subset K$, a solution of (6) may thus be expressed as:



$$\widetilde{F}_4 = \frac{3}{8}(F_5 + F_6 - F_{11} - F_4) + \frac{1}{8}(F_7 + F_8 - F_9 - F_{10})$$
$$\widetilde{F}_5 = \frac{3}{8}(F_7 + F_8 - F_5 - F_6) + \frac{1}{8}(F_9 + F_{10} - F_{11} - F_4)$$
$$\widetilde{F}_6 = \frac{3}{8}(F_9 + F_{10} - F_7 - F_8) + \frac{1}{8}(F_4 + F_{11} - F_5 - F_6)$$
$$\widetilde{F}_7 = \frac{3}{8}(F_4 + F_{11} - F_9 - F_{10}) + \frac{1}{8}(F_5 + F_6 - F_7 - F_8)$$

Fig. 4 Dual fluxes for the neighboring cell of refined cells (2D case)

$$F_{\sigma,\varepsilon} = \sum_{\sigma' \in \mathscr{E}(K)} (\alpha_K)_\sigma^{\sigma'} F_{K,\sigma'}, \qquad \sigma \in \mathscr{E}(K), \varepsilon \in \tilde{\mathscr{E}}(D_\sigma) \text{ and } \varepsilon \subset K,$$

and (H3) is equivalent to requiring bounded coefficients $((\alpha_K)_\sigma^{\sigma'})_{\sigma,\sigma' \in \mathscr{E}(K)}$. In addition, since $\xi_K^\sigma = 1/(2d)$ for all $K \in \mathscr{M}$ and $\sigma \in \mathscr{E}(K)$, system (6) is completely independent from the cell K under consideration. We may thus consider a particular geometry for K, let us say $K = (0, 1)^d$, and find an expression for the coefficients $((\alpha_K)_\sigma^{\sigma'})_{\sigma,\sigma' \in \mathscr{E}(K)}$ which we will apply to all the cells, thus automatically satisfying the constraint (H3). A technique for this computation is described in [1, Sect. 3.2]. The idea is to build a momentum field $\mathbf{w}$ with a constant divergence and such that $\int_\sigma \mathbf{w} \cdot \mathbf{n}_{K,\sigma} d\sigma(\mathbf{x}) = F_{K,\sigma}$, for all $\sigma \in \mathscr{E}(K)$. Then, an easy computation shows that the definition $F_{\sigma,\varepsilon} = \int_\varepsilon \mathbf{w} \cdot \mathbf{n}_{\sigma,\varepsilon} d\sigma(\mathbf{x})$ satisfies (6). The set of coefficients $((\alpha_K)_\sigma^{\sigma'})_{\sigma,\sigma' \in \mathscr{E}(K)}$ obtained for a quadrangle is given in [1, Sect. 3.2]; extension to the three-dimensional case is straightforward.

3.2.2 Dual Fluxes for 2D-Refined Meshes

Here again, we may restrict the computation to square cells. In 2D, if a primal cell is surrounded with four refined cells, the half-diamond cells are obtained by splitting the cell in four sub-squares, each one being split in two triangles. Hence, eight dual fluxes must be computed; if some of the neighboring cells are not refined, one uses a *coarsening* procedure. We begin with computing the dual fluxes across the four sub-squares faces (solid gray color in Fig. 4) so that (6) holds, with $(F_{K,\sigma})_{\sigma \in \mathscr{E}(K)}$ denoted here by F_i $(4 \le i \le 11)$ and $F_{\sigma,\varepsilon}$, $\sigma \in \mathscr{E}(K), \varepsilon \subset K$ denoted here by $\widetilde{F}_i$ $(4 \le i \le 7)$. The linear system to solve has a one dimensional kernel and a particular solution satisfying $(H3)$ is given in Fig. 4. Then, the dual fluxes across the diagonal faces $\bar{F}_i$ $(0 \le i \le 3)$ (dashed gray color in Fig. 4) are computed by isolating the sub-squares and applying the procedure described above for the non-refined case.

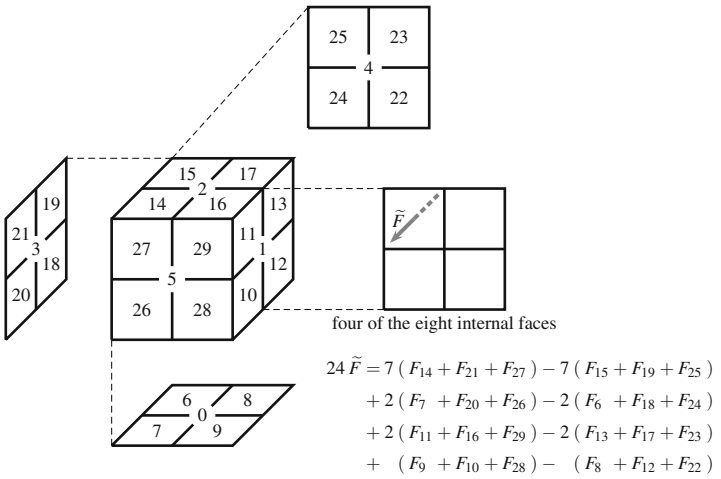


Fig. 5 Intermediate dual fluxes for the neighboring cell of refined cells (3D case)

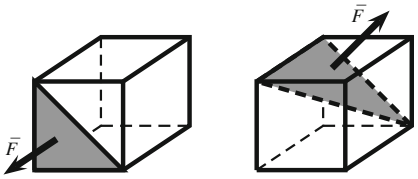


Fig. 6 Two possible types of internal half-diamond faces (3D case)

3.2.3 Dual Fluxes for 3D-Refined Meshes

The procedure is the same as in the 2D-case. The first step consists in splitting the cube in eight sub-cubes and computing the dual fluxes across the faces of these sub-cubes. The formula of one of these intermediate fluxes $\widetilde{F}$ is given in Fig. 5. The computation of the other fluxes across the faces separating two sub-cubes is deduced by permutations of the indices.

In the second step, each sub-cube is split in 3 half-diamonds of equal volumes. One obtains 24 half-diamonds and 48 internal half-diamond faces of two possible types (see Fig. 6). The dual fluxes across these faces are obtained by isolating the sub-cubes and applying the procedure described above for the non-refined case.

4 Numerical Test

We assess the behavior of the proposed numerical scheme on an exact analytical solution to the stationary Navier-Stokes equations known as the Kovasznay flow [7]. Computations are performed with the free software CALIF^{3}S developed at

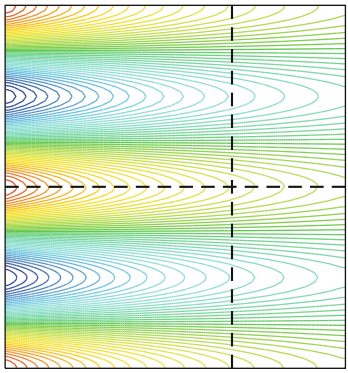


Fig. 7 Contour lines of the field $\mathbf{u}_1$. The *dashed lines* materialize the boundary of the refined area (*bottom-left* and *top-right* sub-domains)

IRSN [2]. The velocity and pressure fields are given by:

$$\mathbf{u} = \begin{bmatrix} 1 - e^{\lambda x} \cos(2\pi y) \\ \dfrac{\lambda}{2\pi} e^{\lambda x} \sin(2\pi y) \end{bmatrix}, \quad p = \frac{1}{2}(1 - e^{2\lambda x}), \quad \lambda = \frac{1}{2\mu} - \left(\frac{1}{4\mu^2} + 4\pi^2\right)^{1/2},$$

where μ stands for the viscosity of the flow, taken here as $\mu = 1/40$. The computational domain is $\Omega = (-0.5, 1) \times (-0.5, 1.5)$. The mesh is built from a regular $n \times n$ grid, where we refine the sub-domain $\Omega_f = (-0.5, 0.5) \times (-0.5, 0.5) \cup (0.5, 1) \times (0.5, 1.5)$ by splitting each (square) cell included in Ω_f in four sub-squares. The solution is computed by the projection scheme, by letting a fictitious transient tend to the desired steady state. Boundary conditions are given by the analytical solution. The obtained numerical errors for various values of n are gathered in the following table, where $\mathbf{u}_{\text{exact}}$ and p_{exact} stand for the exact velocity and pressure, respectively.

n	$\|\mathbf{u} - \mathbf{u}_{\text{exact}}\|_{\mathrm{L}^2(\Omega)}$	$\|p - p_{\text{exact}}\|_{\mathrm{L}^2(\Omega)}$
10	0.183	0.0812
20	0.0384	0.0334
40	0.00825	0.0158
80	0.00211	0.00782

The observed order of convergence in L^2-norm is approximately 2 for the velocity and 1 for the pressure. The contour lines of the first component of the velocity are drawn on Fig. 7. We may check that no spurious perturbation appears along the lines separating the refined and non-refined parts of the computational domain (in other words, the lines composed by the union of the faces including a hanging node). The theoretical study of this scheme is underway, and the error analysis confirms these experiments.

References

1. Ansanay-Alex, G., Babik, F., Latché, J.C., Vola, D.: An L^2-stable approximation of the Navier-Stokes convection operator for low-order non-conforming finite elements. Int. J. Numer. Methods Fluids **66**, 555–580 (2011)
2. CALIF^{3}S: A software components library for the computation of reactive turbulent flows. `https://www.gforge.irsn.fr/gf/project/isis`
3. Chorin, A.: Numerical solution of the Navier-Stokes equations. Math. Comput. **22**, 745–762 (1968)
4. Ciarlet, P.G.: Basic error estimates for elliptic problems. In: Ciarlet, P., Lions, J. (eds.) Handbook of Numerical Analysis, vol. II, pp. 17–351. North Holland, Amsterdam (1991)
5. Gastaldo, L., Herbin, R., Kheriji, W., Lapuerta, C., Latché, J.C.: Staggered discretizations, pressure correction schemes and all speed barotropic flows. In: Finite Volumes for Complex Applications VI—Problems and Perspectives—Prague, Czech Republic, vol. 2, pp. 39–56 (2011)
6. Guermond, J., Minev, P., Shen, J.: An overview of projection methods for incompressible flows. Comput. Methods Appl. Mech. Eng. **195**, 6011–6045 (2006)
7. Kovasznay, L.I.G.: Laminar flow behind a two-dimensional grid. Math. Proc. Cambridge Philos. Soc. **44**(58) (1948)
8. Rannacher, R., Turek, S.: Simple nonconforming quadrilateral Stokes element. Numer. Method Part. Diff. Equ. **8**, 97–111 (1992)
9. Temam, R.: Sur l'approximation de la solution des équations de Navier-Stokes par la méthode des pas fractionnaires II. Arch. Rat. Mech. Anal. **33**, 377–385 (1969)

Consistency Analysis of a 1D Finite Volume Scheme for Barotropic Euler Models

Florent Berthelin, Thierry Goudon and Sebastian Minjeaud

1 Introduction

The model. This work is concerned with the consistency study of a (staggered kinetic) Finite Volume (FV) scheme for barotropic Euler models

$$\partial_t \rho + \partial_x(\rho V) = 0, \qquad \partial_t(\rho V) + \partial_x(\rho V^2 + p(\rho)) = 0. \tag{1}$$

The unknowns are the density ρ and the velocity V. The pressure $\big(\rho \mapsto p(\rho)\big)$ is assumed to be $\mathscr{C}^2\big([0,\infty)\big)$ with $p(\rho) > 0,\ p'(\rho) > 0,\ p''(\rho) \geqslant 0,\ \forall \rho > 0$. Thus, the sound speed $c : \rho \mapsto \sqrt{p'(\rho)}$ is well defined and is an increasing function.

We consider the problem (1) on the bounded domain $(0, L) \times [0, T]$ with the boundary conditions $V(0,t) = 0 = V(L,t),\ \forall t > 0$ and the initial conditions $\rho(x,0) = \rho_0(x),\ V(x,0) = V_0(x),\ \forall x \in (0,L)$ with $\rho_0,\ V_0 \in L^\infty(0,L)$.

Let $\Phi : \rho > 0 \mapsto \Phi(\rho)$ such that $\rho\Phi'(\rho) - \Phi(\rho) = p(\rho),\ \forall \rho > 0$. The quantity $\mathscr{S} = \frac{1}{2}\rho|V|^2 + \Phi(\rho)$ is an entropy of the system: entropy solutions to (1) are required to satisfy: for any $\varphi \in \mathscr{C}_c^\infty\big((0,L)\times[0,T)\big)$ such that $\varphi \geqslant 0$,

F. Berthelin · T. Goudon
INRIA Team COFFEE & University of Nice Sophia Antipolis,
CNRS, LJAD, UMR 7351, Nice, France
e-mail: Florent.Berthelin@unice.fr

T. Goudon
e-mail: thierry.goudon@inria.fr

S. Minjeaud (✉)
INRIA Team CASTOR& University of Nice Sophia Antipolis,
CNRS, LJAD, UMR 7351, Nice, France
e-mail: minjeaud@unice.fr

J. Fuhrmann et al. (eds.), *Finite Volumes for Complex Applications VII - Methods and Theoretical Aspects*, Springer Proceedings in Mathematics & Statistics 77,
DOI: 10.1007/978-3-319-05684-5_8,

$$-\int_0^T \int_0^L \Big[\mathscr{S}\partial_t\varphi + \big(\mathscr{S}+p(\rho)\big)V\partial_x\varphi\Big](x,t)\,\mathrm{d}x\,\mathrm{d}t - \int_0^L \mathscr{S}(x,0)\varphi(x,0)\,\mathrm{d}x \leqslant 0. \tag{2}$$

Results. In [1], the authors introduced a FV scheme for (1) ensuring that discrete kinetic and internal energies evolution equations hold (see Lemma 1). As in [2], we complete here this analysis with a Lax-Wendroff-like statement: the limit of a converging (and uniformly bounded) sequence of stepwise constant functions defined from the scheme is a weak entropic-solution of the system of conservation laws.

The meshes. We consider a set of $J+1$ points $0 = x_1 < x_2 < ... < x_J < x_{J+1} = L$. The x_j are the edges of the so-called primal mesh $\mathscr{T}$. We set $\delta x_{j+1/2} = x_{j+1} - x_j$. The centers of the primal cells, $x_{j+1/2} = (x_j + x_{j+1})/2$ for $j \in \{1, .., J\}$, realize the dual mesh $\mathscr{T}^\star$. We set $\delta x_j = (\delta x_{j-1/2} + \delta x_{j+1/2})/2$ for $j \in \{2, .., J-1\}$ and $\delta x = \text{size}(\mathscr{T}) = \max_j \delta x_{j+1/2}$. The adaptive time step is δt^k and we set $\delta t = \max_k \delta t^k$.

The scheme. We analyze the scheme introduced in [1]. It works on staggered grids: the densities, $\rho_{j+1/2}$, $j \in \{1, .., J\}$, are evaluated at centers whereas the velocities, V_j, $j \in \{1, .., J+1\}$, are evaluated at edges. We set, for $j \in \{1, .., J\}$ and $i \in \{2, .., J\}$

$$\rho^0_{j+1/2} = \frac{1}{\delta x_{j+1/2}} \int_{x_j}^{x_{j+1}} \rho_0(x)\,\mathrm{d}x, \quad V^0_i = \frac{1}{\delta x_i} \int_{x_{i-1/2}}^{x_{i+1/2}} V_0(x)\,\mathrm{d}x. \tag{3}$$

The density is first updated with a FV approximation on the primal mesh

$$\delta x_{j+1/2}\big(\rho^{k+1}_{j+1/2} - \rho^k_{j+1/2}\big) + \delta t^k\big(\mathscr{F}^k_{j+1} - \mathscr{F}^k_j\big) = 0, \quad \forall j \in \{1, .., J\}. \tag{4}$$

Then, the velocity is updated with a FV approximation on the dual mesh:

$$\delta x_j\big(\rho^{k+1}_j V^{k+1}_j - \rho^k_j V^k_j\big) + \delta t^k\Big(\mathscr{G}^k_{j+1/2} - \mathscr{G}^k_{j-1/2} + \pi^{k+1/2}_{j+1/2} - \pi^{k+1/2}_{j-1/2}\Big) = 0, \tag{5}$$

for $j \in \{2, .., J\}$, while $V^{k+1}_1 = V^{k+1}_{J+1} = 0$. The density on the edges ρ^k_j is defined by

$$2\delta x_j \rho^k_j = \delta x_{j+1/2}\rho^k_{j+1/2} + \delta x_{j-1/2}\rho^k_{j-1/2}, \quad \forall j \in \{2, .., J\}.$$

The definition of the fluxes relies on the kinetic framework. We refer the reader to [1] for details. Let us introduce the two following functions $\mathscr{F}^+$ and $\mathscr{F}^-$

$$\mathscr{F}^\pm(\rho, V) = \frac{\rho}{2c(\rho)} \int_{\pm\xi>0} \xi\, \chi_{\rho,V}(\xi)\mathrm{d}\xi \text{ where } \chi_{\rho,V}(\xi) = \begin{cases} 1 \text{ if } |\xi - V| \leqslant c(\rho) \\ 0 \text{ otherwise} \end{cases}.$$

We adopt the following formulas for mass fluxes: $\mathscr{F}^k_1 = \mathscr{F}^k_{J+1} = 0$,

$$\mathscr{F}^k_j = \mathscr{F}^+(\rho^k_{j-1/2}, V^k_j) + \mathscr{F}^-(\rho^k_{j+1/2}, V^k_j), \quad \forall j \in \{2, .., J\}, \tag{6}$$

and, for momentum fluxes: $\mathscr{G}^k_{3/2} = \dfrac{V^k_2}{2}\mathscr{F}^-(\rho^k_{5/2}, V^k_2)$, $\mathscr{G}^k_{J+1/2} = \dfrac{V^k_J}{2}\mathscr{F}^+(\rho^k_{J-1/2}, V^k_J)$,

$$\mathscr{G}^k_{j+1/2} = \frac{V^k_j}{2}\big(\mathscr{F}^+(\rho^k_{j-1/2}, V^k_j) + \mathscr{F}^+(\rho^k_{j+1/2}, V^k_{j+1})\big) + \frac{V^k_{j+1}}{2}\big(\mathscr{F}^-(\rho^k_{j+1/2}, V^k_j) + \mathscr{F}^-(\rho^k_{j+3/2}, V^k_{j+1})\big),\ \forall j \in \{2, .., J-1\}. \tag{7}$$

The discrete pressure gradient combines space-centered-difference and time-semi-implicit discretization, namely it uses $\pi^{k+1/2}_{j+1/2} = \rho^k_{j+1/2}\Phi'(\rho^{k+1}_{j+1/2}) - \Phi(\rho^k_{j+1/2})$.

Properties of the scheme. The analysis is driven by the shapes of the functions $\mathscr{F}^\pm$, see [1, Lemma 3.2]. Here, we shall use the following properties

$$\begin{array}{l}\text{(i) Smoothness: } (\rho, V) \in (0,\infty)\times\mathbb{R} \mapsto \mathscr{F}^\pm(\rho, V) \text{ are of class } C^1,\\ \text{(ii) Consistency: } \mathscr{F}^+(\rho, V) + \mathscr{F}^-(\rho, V) = \rho V, \quad \forall V \in \mathbb{R},\ \forall \rho \geqslant 0.\end{array} \tag{8}$$

The following lemma, see [1], states the main properties of the scheme.

Lemma 1 *Let $N \in \mathbb{N}$. Assume $\min_i\left(\rho^0_{i+1/2}\right) > 0$. For all $k \in \{0, .., N-1\}$, there exists $\mathscr{V}^k > 0$, which depends only on the state (ρ^k, V^k), such that if*

$$\frac{\delta t^k}{\min_j\left(\delta x_{j+1/2}\right)}\mathscr{V}^k \leqslant 1, \tag{9}$$

then, $\min_i\left(\rho^k_{i+1/2}\right) > 0, \forall k \in \{0, .., N\}$ and

$$0 \leqslant \sum_{k=0}^{N-1}\sum_{j=2}^{J} D^k_j \leqslant C, \quad \text{with} \quad D^k_j = \frac{1}{4}\delta x_j \rho^{k+1}_j \big(V^{k+1}_j - V^k_j\big)^2, \tag{10}$$

$$\frac{\delta x_{j+1/2}}{\delta t^k}\Big[e^{k+1}_{j+1/2} - e^k_{j+1/2}\Big] + \overline{G}^k_{j+1} - \overline{G}^k_j + \pi^{k+1/2}_{j+1/2}\Big[V^{k+1}_{j+1} - V^{k+1}_j\Big] \leqslant \frac{D^k_j}{\delta t^k}, \tag{11}$$

$$\frac{\delta x_j}{\delta t^k}\Big[E^{k+1}_{K,j} - E^k_{K,j}\Big] + \Gamma^k_{j+1/2} - \Gamma^k_{j-1/2} + \Big[\pi^{k+1/2}_{j+1/2} - \pi^{k+1/2}_{j-1/2}\Big]V^{k+1}_j + \frac{D^k_j}{\delta t^k} \leqslant 0, \tag{12}$$

where $E^k_{K,j} = \frac{1}{2}\rho^k_j\big(V^k_j\big)^2$ and $e^k_{j+1/2} = \Phi(\rho^k_{j+1/2})$ are the kinetic and internal energies. The fluxes are defined by $\overline{G}^k_1 = \overline{G}^k_{J+1} = 0$ and

$$\overline{G}_j^k = \Phi(\rho_{j-1/2}^k)V_j^{k+1} - \frac{\delta x_{j-1/2}}{2\delta t^k}\Big[\bar{\Phi}\Big(\overline{\overline{\rho_{j-1/2}^{k+1}}}\Big) - \bar{\Phi}(\rho_{j-1/2}^k)\Big],\ \forall j \in \{2,..,J\},$$

$$\Gamma_{j+1/2}^k = \frac{1}{2}V_j^k V_{j+1}^k \frac{\mathscr{F}_j^k + \mathscr{F}_{j+1}^k}{2} + \frac{1}{2}(V_j^k - V_{j+1}^k)^2 \frac{\mathscr{F}_j^{k,|\cdot|} + \mathscr{F}_{j+1}^{k,|\cdot|}}{2},\ \forall j \in \{1,..,J\},$$

$$\overline{\overline{\rho_{j-1/2}^{k+1}}} = \rho_{j-1/2}^k - \frac{2\delta t^k}{\delta x_{j-1/2}}\Big(\mathscr{F}^-(\rho_{j+1/2}^k, V_j^k) - \mathscr{F}^-(\rho_{j-1/2}^k, V_j^k) - \rho_{j-1/2}^k\big(V_j^{k+1} - V_j^k\big)\Big),$$

and $\mathscr{F}_1^{k,|\cdot|} = \mathscr{F}_{J+1}^{k,|\cdot|} = 0$, $\mathscr{F}_j^{k,|\cdot|} = \mathscr{F}^+(\rho_{j-1/2}^k, V_j^k) - \mathscr{F}^-(\rho_{j+1/2}^k, V_j^k)$, $\forall j \in \{2,..,J\}$. *The function* $\bar{\Phi}$ *is a* $\mathscr{C}^2$ *extension of the function* Φ *(see [1, Section 4.3]).*

2 Consistency Analysis

Notation. Assuming that $\sum_{k=0}^{N-1} \delta t^k = T$, we define the reconstructions ($i = 0, 1$)

$$\rho_\delta^{(i)} = \sum_{k=0}^{N-1}\sum_{j=1}^{J} \rho_{j+1/2}^{k+i}\chi_{j+1/2}^{k+1/2},\ \pi_\delta = \sum_{k=0}^{N-1}\sum_{j=1}^{J} \pi_{j+1/2}^{k+1/2}\chi_{j+1/2}^{k+1/2},\ V_\delta = \sum_{k=0}^{N-1}\sum_{j=2}^{J} V_j^k \chi_j^{k+1/2},$$

where $\chi_j^{k+1/2} = \chi_{[x_{j-1/2},\, x_{j+1/2}[\times[t^k,\, t^{k+1}[}$, $\chi_{j+1/2}^{k+1/2} = \chi_{[x_j,\, x_{j+1}[\times[t^k,\, t^{k+1}[}$. We also set

$$||\rho_\delta||_{\infty,\mathscr{T}} = \max_{0\leqslant k\leqslant N}\ \max_{1\leqslant j\leqslant J} |\rho_{j+1/2}^k|, \qquad ||V_\delta||_{\infty,\mathscr{T}^\star} = \max_{0\leqslant k\leqslant N}\ \max_{2\leqslant j\leqslant J} |V_j^k|,$$

$$||\rho_\delta||_{1;\mathrm{BV},\mathscr{T}} = \sum_{k=0}^{N}\delta t^k \sum_{j=2}^{J} |\rho_{j+1/2}^k - \rho_{j-1/2}^k|, \quad ||V_\delta||_{1;\mathrm{BV},\mathscr{T}^\star} = \sum_{k=0}^{N}\delta t^k \sum_{j=1}^{J} |V_{j+1}^k - V_j^k|,$$

$$||\rho_\delta||_{\mathrm{BV};1,\mathscr{T}} = \sum_{j=1}^{J}\delta x_{j+1/2}\sum_{k=0}^{N-1} |\rho_{j+1/2}^{k+1} - \rho_{j+1/2}^k|.$$

For $\varphi \in \mathscr{C}_c^\infty\big((0,L)\times[0,T)\big)$, we set $\varphi_{j+1/2}^k = \varphi(x_{j+1/2}, t^k)$ and $\varphi_j^k = \varphi(x_j, t^k)$. The interpolate $\varphi_{\mathscr{T}}$ of φ on the primal mesh and its discrete derivatives are defined by

$$\varphi_{\mathscr{T}}(\cdot,0) = \sum_{j=1}^{J}\varphi_{j+1/2}^0\chi_{j+1/2}^{1/2}(\cdot,0), \qquad \varphi_{\mathscr{T}}(\cdot,t) = \sum_{k=0}^{N-1}\sum_{j=1}^{J}\varphi_{i+1/2}^{k+1}\chi_{j+1/2}^{k+1/2}(\cdot,t),\ \forall t > 0,$$

$$\eth_t\varphi_{\mathscr{T}} = \sum_{k=0}^{N-1}\sum_{j=1}^{J}\frac{\varphi_{j+1/2}^{k+1} - \varphi_{j+1/2}^k}{\delta t^k}\chi_{j+1/2}^{k+1/2}, \quad \eth_x\varphi_{\mathscr{T}} = \sum_{k=0}^{N-1}\sum_{j=2}^{J}\frac{\varphi_{j+1/2}^{k+1} - \varphi_{j-1/2}^{k+1}}{\delta x_j}\chi_j^{k+1/2}.$$

Similarly, the interpolate $\varphi_{\mathscr{T}^\star}$ of φ on $\mathscr{T}^\star$ and its discrete derivatives are given by

$$\varphi_{\mathscr{T}^\star}(\cdot,0)=\sum_{j=2}^{J}\varphi_j^0\chi_j^{1/2}(\cdot,0),\qquad \varphi_{\mathscr{T}^\star}(\cdot,t)=\sum_{k=0}^{N-1}\sum_{j=2}^{J}\varphi_j^{k+1}\chi_j^{k+1/2}(\cdot,t),\ \forall t>0,$$

$$\eth_t^\star\varphi_{\mathscr{T}^\star}=\sum_{k=0}^{N-1}\sum_{j=2}^{J}\frac{\varphi_j^{k+1}-\varphi_j^k}{\delta t^k}\chi_j^{k+1/2},\qquad \eth_x^\star\varphi_{\mathscr{T}^\star}=\sum_{k=0}^{N-1}\sum_{j=1}^{J}\frac{\varphi_{j+1}^{k+1}-\varphi_j^{k+1}}{\delta x_{j+1/2}}\chi_{j+1/2}^{k+1/2}.$$

Assumptions. Let $\left(\mathscr{T}_m\right)_{m\geqslant 1}$ be a sequence of meshes s. t. size$(\mathscr{T}_m)\to 0$ and a familly of time steps $\left(\delta t_m^k\right)_{k\geqslant 0,\ m\geqslant 1}$ verifying $\delta t_m\to 0$ and (9). Assume that there exists $N_m\in\mathbb{N}$ s. t. $\sum_{k=0}^{N_m-1}\delta t_m^k=T$. The scheme defines $(\rho_{\delta_m}^{(0)},V_{\delta_m})_{m\geqslant 1}$. Suppose that

$$||\rho_{\delta_m}^{(0)}||_{\infty,\mathscr{T}}+||V_{\delta_m}||_{\infty,\mathscr{T}^\star}\leqslant C_\infty,\qquad ||\rho_{\delta_m}^{(0)}||_{1;\mathrm{BV},\mathscr{T}}+||V_{\delta_m}||_{1;\mathrm{BV},\mathscr{T}^\star}\leqslant C_{\mathrm{BV}}\quad(13)$$

holds and, in the case $\left(\rho\mapsto\frac{p'(\rho)}{\rho}\right)\notin L^1_{\mathrm{loc}}(0,\infty)$, $||\left(\rho_{\delta_m}^{(0)}\right)^{-1}||_{\infty,\mathscr{T}}\leqslant C$. We assume that there exists $(\bar\rho,\bar V)\in L^\infty((0,T)\times(0,L))^2$ such that

$$\left(\rho_{\delta_m}^{(0)},V_{\delta_m}\right)\longrightarrow(\bar\rho,\bar V)\text{ in }L^r((0,T)\times(0,L))^2,\ 1\leqslant r<\infty.\quad(14)$$

Main results. The uniform bounds imply that there exists constants such that

$$\sup_{0\leqslant\rho,|V|\leqslant C_\infty}|\mathscr{A}(\rho,V)|\leqslant C_{\mathscr{A}},\quad\text{with }\mathscr{A}=\mathscr{F}^\pm,\ \partial_\rho\mathscr{F}^\pm\text{ and }\partial_V\mathscr{F}^\pm,$$

$$\sup_{0\leqslant\rho\leqslant C_\infty+4(C_\infty^2+C_{\mathscr{F}^\pm})}|\mathscr{B}(\rho)|\leqslant C_{\mathscr{B}},\quad\text{with }\mathscr{B}=\Phi,\ \Phi',\ \text{and }\bar\Phi'.$$

Note that $\left|\Phi\left(\rho_{j+1/2}^k\right)\right|\leqslant C_{\Phi,\rho}\rho_{j+1/2}^k,\ \forall j,k$. We also show that $||\rho_{\delta_m}^{(0)}||_{\mathrm{BV};1,\mathscr{T}}\leqslant C$ by using (4) which allows to dominate $\delta x_{j+1/2}|\rho_{j+1/2}^{k+1}-\rho_{j+1/2}^k|$ by

$$\delta t^k\Big[C_{\partial_\rho\mathscr{F}^\pm}\left(|\rho_{j+1/2}^k-\rho_{j-1/2}^k|+|\rho_{j+3/2}^k-\rho_{j+1/2}^k|\right)+2C_{\partial_V\mathscr{F}^\pm}|V_{j+1}^k-V_j^k|\Big].$$

Consequently, $\rho_{\delta_m}^{(1)}\to\bar\rho$ and $\pi_{\delta_m}\to p(\bar\rho)$ in $L^r((0,T)\times(0,L))$; with (3) and since $\rho_0,V_0\in L^\infty(0,L)$, we get $\rho_{\delta_m}^{(0)}(\cdot,0)\to\rho_0$ and $V_{\delta_m}(\cdot,0)\to V_0$ in $L^r((0,L))$, $1\leqslant r<\infty$.

In the sequel, when a function $\varphi\in\mathscr{C}_c^\infty\left((0,L)\times[0,T)\right)$ is given, we assume that δt_m and δx_m are sufficiently small so that $\varphi(x,\cdot)\equiv 0,\ \forall x\in[0,x_{3/2}]\cup[x_{J+1/2},L]$ and $\varphi(\cdot,t)\equiv 0,\ \forall t\in[t^{N-1},t^N]$. Moreover, since φ is smooth, $\varphi_{\mathscr{T}_m}$, $\varphi_{\mathscr{T}_m^\star}\to\varphi$, $\eth_t\varphi_{\mathscr{T}_m}$, $\eth_t^\star\varphi_{\mathscr{T}_m^\star}\to\partial_t\varphi$, and $\eth_x\varphi_{\mathscr{T}_m}$, $\eth_x^\star\varphi_{\mathscr{T}_m^\star}\to\partial_x\varphi$, in $L^r((0,T)\times(0,L))$, $1\leqslant r\leqslant\infty$.

Theorem 1 *Assume* (13) *and* (14). *Then,* $(\bar\rho,\bar V)$ *satisfies* (1) *in the distribution sense in* $\left(\mathscr{C}_c^\infty\left((0,L)\times[0,T)\right)\right)'$. *Moreover,* $(\bar\rho,\bar V)$ *satisfies the entropy inequality* (2).

Proof Let $\varphi \in \mathscr{C}_c^\infty\big((0,L)\times[0,T)\big)$. For the sake of simplicity, the index m is dropped.
Mass balance. We multiply (4) by $\varphi_{j+1/2}^{k+1}$ and sum the results to obtain

$$\underbrace{\sum_{k=0}^{N-1}\sum_{j=1}^{J}\delta x_{j+1/2}(\rho_{j+1/2}^{k+1}-\rho_{j+1/2}^{k})\varphi_{j+1/2}^{k+1}}_{:=T_1}+\underbrace{\sum_{k=0}^{N-1}\delta t^k\sum_{j=1}^{J}\big(\mathscr{F}_{j+1}^{k}-\mathscr{F}_{j}^{k}\big)\varphi_{j+1/2}^{k+1}}_{:=T_2}=0.$$

For the first term, since $\varphi_{j+1/2}^{N}=0$, a discrete integration by part w.r.t. time yields

$$T_1=-\int_0^T\int_0^L \rho_\delta^{(0)}\eth_t\varphi_{\mathscr{T}}\,\mathrm{d}x\,\mathrm{d}t-\int_0^L\rho_\delta^{(0)}(x,0)\varphi_{\mathscr{T}}(x,0)\,\mathrm{d}x.$$

For T_2, we combine the two following expressions of mass fluxes (see (8)-(ii))

$$\mathscr{F}_j^k=\rho_{j\pm1/2}^k V_j^k\mp R_j^{k,\pm}\ \text{with}\ R_j^{k,\pm}=\mathscr{F}^{\pm}(\rho_{j+1/2}^k,V_j^k)-\mathscr{F}^{\pm}(\rho_{j-1/2}^k,V_j^k).$$

Integrating by part w.r.t. space, we readily obtain $T_2=-T_{2,1}-T_{2,2}$ with

$$T_{2,1}=\int_0^T\int_{x_{3/2}}^{x_{J+1/2}}\rho_\delta^{(0)}V_\delta\eth_x\varphi_{\mathscr{T}}\,\mathrm{d}x\,\mathrm{d}t=\int_0^T\int_0^L\rho_\delta^{(0)}V_\delta\eth_x\varphi_{\mathscr{T}}\,\mathrm{d}x\,\mathrm{d}t,$$

$$T_{2,2}=\sum_{k=0}^{N-1}\delta t^k\sum_{j=2}^{J}\frac12\Big[\delta x_{j-1/2}R_j^{k,-}-\delta x_{j+1/2}R_j^{k,+}\Big]V_j^k\,\frac{\varphi_{j+1/2}^{k+1}-\varphi_{j-1/2}^{k+1}}{\delta x_j}.$$

With (14), we pass to the limit in T_1 and $T_{2,1}$. We prove that $(\bar\rho,\bar V)$ satisfies the mass conservation equation by showing that $T_{2,2}\to0$ since

$$|T_{2,2}|\leqslant C_{\partial_\rho\mathscr{F}^\pm}\,|\partial_x\varphi|_{L^\infty}\,||V_\delta||_{\infty,\mathscr{T}^\star}\,||\rho_\delta||_{1;\mathrm{BV},\mathscr{T}}\,\delta x\ \lesssim\ \delta x.$$

Momentum balance. We multiply (5) by φ_j^{k+1} and sum. We proceed as for the mass conservation, by using the following expression of the momentum flux

$$\begin{cases}\mathscr{G}_{j+1/2}^k &=\dfrac12\rho_{j+1/2}^k\Big[\big(V_j^k\big)^2+\big(V_{j+1}^k\big)^2\Big]+Q_{j+1/2}^k,\\ Q_{j+1/2}^k&=-\dfrac12V_j^kR_j^{k,+}+\dfrac12V_{j+1}^kR_{j+1}^{k,-}\\ &\quad-\dfrac12\big(V_{j+1}^k-V_j^k\big)\Big[\mathscr{F}^+(\rho_{j+1/2}^k,V_{j+1}^k)-\mathscr{F}^-(\rho_{j+1/2}^k,V_j^k)\Big].\end{cases}$$

Entropy inequality. We now assume that $\varphi\geqslant0$.
• Kinetic energy. We multiply (12) by $\delta t^k\varphi_j^{k+1}$ and sum to get $T_3+T_4+T_5\leqslant0$ with

$$T_3 = \sum_{k=0}^{N-1}\sum_{j=2}^{J}\delta x_j\Big[E_{K,j}^{k+1} - E_{K,j}^{k}\Big]\varphi_j^{k+1},\quad T_4 = \sum_{k=0}^{N-1}\delta t^k\sum_{j=2}^{J}\Big[\Gamma_{j+1/2}^{k} - \Gamma_{j-1/2}^{k}\Big]\varphi_j^{k+1},$$
$$T_5 = \sum_{k=0}^{N-1}\delta t^k\sum_{j=2}^{J}\Big[\pi_{j+1/2}^{k+1/2} - \pi_{j-1/2}^{k+1/2}\Big]V_j^{k+1}\varphi_j^{k+1} + \sum_{k=0}^{N-1}\sum_{j=2}^{J}D_j^k\varphi_j^{k+1}.$$

Integrating by part w.r.t. time yields

$$T_3 = -\int_0^T\int_0^L \frac{1}{2}\rho_\delta^{(0)}\big(V_\delta\big)^2\eth_t^\star\varphi_{\mathscr{T}^\star}\,\mathrm{d}x\,\mathrm{d}t - \int_0^L\frac{1}{2}\rho_\delta^{(0)}(x,0)\big(V_\delta(x,0)\big)^2\varphi_{\mathscr{T}^\star}(x,0)\,\mathrm{d}x.$$

For T_4, we write $\Gamma_{j+1/2}^k = \frac{1}{4}\rho_{j+1/2}^k\big[\big(V_j^k\big)^3 + \big(V_{j+1}^k\big)^3\big] + \frac{1}{4}S_{j+1/2}^k$ where

$$S_{j+1/2}^k = V_j^kV_{j+1}^k\Big[R_{j+1}^{k,-} - R_j^{k,+}\Big] + (V_{j+1}^k - V_j^k)^2\Big[\mathscr{F}_j^{k,|\cdot|} + \mathscr{F}_{j+1}^{k,|\cdot|} - \rho_{j+1/2}^k(V_j^k + V_{j+1}^k)\Big].$$

Integration by part w.r.t space leads to $T_4 = -T_{4,1} - T_{4,2}$ with

$$T_{4,1} = \int_0^T\int_0^L\frac{1}{2}\rho_\delta^{(0)}\big(V_\delta\big)^3\eth_x^\star\varphi_{\mathscr{T}^\star}\,\mathrm{d}x\,\mathrm{d}t,\quad T_{4,2} = \frac{1}{4}\sum_{k=0}^{N-1}\delta t^k\sum_{j=1}^{J}S_{j+1/2}^k\Big[\varphi_{j+1}^{k+1} - \varphi_j^{k+1}\Big].$$

Finally $|T_{4,2}| \lesssim \delta x$ since it is dominated by

$$\delta x|\partial_x\varphi|_{L^\infty}||V_\delta||_{\infty,\mathscr{T}^\star}\Big[\frac{C_{\partial_\rho\mathscr{F}^\pm}}{2}||V_\delta||_{\infty,\mathscr{T}^\star}||\rho_\delta||_{1;\mathrm{BV},\mathscr{T}} + \big(2C_{\mathscr{F}^\pm} + ||V_\delta||_{\infty,\mathscr{T}^\star}||\rho_\delta||_{\infty,\mathscr{T}}\big)||V_\delta||_{1;\mathrm{BV},\mathscr{T}^\star}\Big].$$

• Internal energy. Multiply (11) by $\delta t^k\varphi_{j+1/2}^{k+1}$ and sum to get $T_6+T_7+T_8\leqslant 0$ with

$$T_6 = \sum_{k=0}^{N-1}\sum_{j=1}^{J}\delta x_{j+1/2}\Big[e_{j+1/2}^{k+1} - e_{j+1/2}^k\Big]\varphi_{j+1/2}^{k+1},\ T_7 = \sum_{k=0}^{N-1}\delta t^k\sum_{j=1}^{J}\Big[\overline{G}_{j+1}^k - \overline{G}_j^k\Big]\varphi_{j+1/2}^{k+1},$$
$$T_8 = \sum_{k=0}^{N-1}\delta t^k\sum_{j=1}^{J}\pi_{j+1/2}^{k+1/2}(V_{j+1}^{k+1} - V_j^{k+1})\varphi_{j+1/2}^{k+1} - \sum_{k=0}^{N-1}\sum_{j=1}^{J}D_j^k\varphi_{j+1/2}^{k+1}.$$

Owing to integration by part w.r.t. time, we get

$$T_6 = -\int_0^T\int_0^L\Phi\big(\rho_\delta^{(0)}\big)\eth_t\varphi_{\mathscr{T}}\,\mathrm{d}x\,\mathrm{d}t - \int_0^L\Phi\big(\rho_\delta^{(0)}(x,0)\big)\varphi_{\mathscr{T}}(x,0)\,\mathrm{d}x.$$

We rewrite the flux as follows

$$\begin{cases} \overline{G}_j^k = \dfrac{1}{2\delta x_j}\Big[\delta x_{j-1/2}\Phi(\rho_{j-1/2}^k) + \delta x_{j+1/2}\Phi(\rho_{j+1/2}^k)\Big]V_j^k + U_{1,j}^k + U_{2,j}^k + U_{3,j}^k, \\ U_{1,j}^k = e_{j-1/2}^k\big(V_j^{k+1} - V_j^k\big), \qquad U_{2,j}^k = -\dfrac{\delta x_{j+1/2}}{2\delta x_j}\Big[e_{j+1/2}^k - e_{j-1/2}^k\Big]V_j^k, \\ U_{3,j}^k = -\dfrac{\delta x_{j-1/2}}{2\delta t^k}\Big[\bar{\Phi}\big(\overline{\overline{\rho_{j-1/2}^{k+1}}}\big) - \bar{\Phi}(\rho_{j-1/2}^k)\Big]. \end{cases}$$

It leads to $T_7 = -T_{7,0} - T_{7,1} - T_{7,2} - T_{7,3}$ with

$$\begin{cases} T_{7,0} = \displaystyle\int_0^T \int_0^L \Phi\big(\rho_\delta^{(0)}\big) V_\delta \eth_x \varphi_{\mathscr{T}} \,\mathrm{d}x\,\mathrm{d}t, \\ T_{7,i} = \displaystyle\sum_{k=0}^{N-1} \delta t^k \sum_{j=2}^{J} U_{i,j}^k\big(\varphi_{j+1/2}^{k+1} - \varphi_{j-1/2}^{k+1}\big), \quad i = 1,2,3. \end{cases}$$

The term $T_{7,1}$ can be bounded as follows

$$|T_{7,1}| \leqslant C_{\Phi,\rho}\big|\partial_x\varphi\big| \sum_{k=0}^{N-1} \delta t^k \sum_{j=2}^{J} \delta x_j \rho_{j-1/2}^k \big|V_j^{k+1} - V_j^k\big|. \tag{15}$$

Since $a \leqslant \min(a,b) + |b-a|$, we get $\rho_{j-1/2}^k \leqslant \rho_j^k + \big|\rho_{j+1/2}^k - \rho_{j-1/2}^k\big|$. This leads to

$$|T_{7,1}| \leqslant C_{\Phi,\rho}\big|\partial_x\varphi\big|_{L^\infty}\Bigg(\underbrace{\sum_{k=0}^{N-1} \delta t^k \sum_{j=2}^{J} \delta x_j \rho_j^k \big|V_j^{k+1} - V_j^k\big|}_{:=T^\star} + 2||V_\delta||_{\infty,\mathscr{T}^\star}||\rho_\delta||_{1;\mathrm{BV},\mathscr{T}}\delta x\Bigg).$$

Writing $\rho_j^k = \rho_j^{k+1} - (\rho_j^{k+1} - \rho_j^k)$ and using the Cauchy-Schwarz inequality yields

$$T^\star \leqslant 2\Big(TL||\rho_\delta||_{\infty,\mathscr{T}}\Big)^{1/2}\Big(\delta t \sum_{k=0}^{N-1}\sum_{j=2}^{J} D_j^k\Big)^{1/2} + 2||V_\delta||_{\infty,\mathscr{T}^\star}||\rho_\delta||_{\mathrm{BV};1,\mathscr{T}}\delta t \;\lesssim\; \delta t^{1/2}.$$

It finally leads to $|T_{7,1}| \lesssim \delta t^{\frac{1}{2}} + \delta x$. The term $T_{7,2}$ can be bounded as follows

$$|T_{7,2}| \leqslant C_{\Phi'}||V_\delta||_{\infty,\mathscr{T}^\star}|\partial_x\varphi|_{L^\infty}||\rho_\delta||_{1;\mathrm{BV},\mathscr{T}}\;\delta x \;\lesssim\; \delta x.$$

We now turn to $T_{7,3}$. We remark that

$$\Big|\overline{\overline{\rho_{j-1/2}^{k+1}}} - \rho_{j-1/2}^k\Big| \leqslant \frac{2\delta t^k}{\delta x_{j-1/2}}\Big(C_{\partial_\rho,\mathscr{F}^\pm}\big|\rho_{j+1/2}^k - \rho_{j-1/2}^k\big| + \rho_{j-1/2}^k\big|V_j^{k+1} - V_j^k\big|\Big).$$

Hence, using the same bound as for $T_{7,1}$ yields

$$\left|T_{7,3}\right| \leqslant C_{\tilde{\Phi}'}|\partial_x\varphi|_{L^\infty}\left(\left(C_{\partial_\rho\mathscr{F}^\pm}+2||V_\delta||_{\infty,\mathscr{T}^\star}\right)||\rho_\delta||_{1;\mathrm{BV},\mathscr{T}}\delta x+T^\star\right) \lesssim \delta t^{1/2}+\delta x.$$

• Pressure. It remains to get the limit of $T_5+T_8=-T_{9,0}-T_{9,1}-T_{9,2}-T_{9,3}$ with

$$T_{9,0}=\int_0^T\int_0^L \pi_\delta V_\delta \eth_x^\star\varphi_{\mathscr{T}^\star}\,\mathrm{d}x\,\mathrm{d}t, \qquad T_{9,1}=\sum_{k=0}^{N-1}\sum_{j=2}^{J} D_j^k\left(\varphi_{j+1/2}^{k+1}-\varphi_j^{k+1}\right),$$

$$T_{9,2}=\frac{1}{2}\sum_{k=0}^{N-1}\delta t^k\sum_{j=1}^{J}\pi_{j+1/2}^{k+1/2}\left(V_j^{k+1}-V_j^k+V_{j+1}^{k+1}-V_{j+1}^k\right)\left(\varphi_{j+1}^{k+1}-\varphi_j^{k+1}\right),$$

$$T_{9,3}=-\frac{1}{2}\sum_{k=0}^{N-1}\delta t^k\sum_{j=1}^{J}\pi_{j+1/2}^{k+1/2}\left(V_{j+1}^{k+1}-V_j^{k+1}\right)\left(2\varphi_{j+1/2}^{k+1}-\varphi_{j+1}^{k+1}-\varphi_j^{k+1}\right).$$

We bound $T_{9,1}$ and $T_{9,3}$ as follows

$$T_{9,1}\leqslant\frac{|\partial_x\varphi|_{L^\infty}}{2}\left(\sum_{k=0}^{N-1}\sum_{j=2}^{J}D_j^k\right)\delta x\lesssim\delta x, \quad T_{9,3}\leqslant\frac{C_\pi}{4}|\partial_{xx}\varphi|_{L^\infty}||V_\delta||_{1;\mathrm{BV},\mathscr{T}^\star}(\delta x)^2\lesssim(\delta x)^2.$$

Note that $\left|\pi_{j+1/2}^{k+1/2}\right|\leqslant\left(C_{\Phi'}+C_{\Phi,\rho}\right)\rho_{j+1/2}^k$. It readily leads to

$$\left|T_{9,2}\right|\leqslant\left(C_{\Phi'}+C_{\Phi,\rho}\right)|\partial_x\varphi|_{L^\infty}T^\star\lesssim\delta t^{\frac{1}{2}}.$$

With (14), we pass to the limit in T_3, $T_{4,1}$, T_6, $T_{7,0}$ and $T_{9,0}$. We arrive at (2) since the other terms tend to 0.

References

1. Berthelin, F., Goudon, T., Minjeaud, S.: Kinetic schemes on staggered grids for barotropic euler models: entropy-stability, analysis. Mathematics of Computation (2014)
2. Herbin, R., Latché, J.C., Nguyen, T.T.: Consistent explicit staggered schemes for compressible flows; part I: the barotropic Euler equations. Technical Report, LATP, University of Aix-Marseille & CNRS (2013)

An Asymptotic-Preserving Scheme for Systems of Conservation Laws with Source Terms on 2D Unstructured Meshes

C. Berthon, G. Moebs and R. Turpault

Abstract A finite volumes numerical scheme is here proposed for hyperbolic systems of conservation laws with source terms which degenerate into parabolic systems in large times when the source terms become stiff. In this framework, it is crucial that the numerical schemes are asymptotic-preserving i.e. that they degenerate accordingly. Here, an asymptotic-preserving numerical scheme is designed for any system within the aforementioned class on 2D unstructured meshes. This scheme is proved to be consistent and stable under a suitable CFL condition. Moreover, we show that it is also possible to prove that it preserves the set of (physically) admissible states under a geometrical property on the mesh. Finally, numerical examples are given to illustrate its behavior.

1 Introduction

The objective of this paper is to build a suitable numerical scheme for hyperbolic systems of conservation laws which can be written under the following form:

$$\partial_t \mathbf{U} + \operatorname{div}(\mathbf{F}(\mathbf{U})) = \gamma(\mathbf{U})(\mathbf{R}(\mathbf{U}) - \mathbf{U}),\ (t, x) \in \mathbb{R}_+ \times \mathbb{R}^2. \tag{1}$$

Here, the Jacobian of the flux $\mathbf{F}$ is assumed to be diagonalizable in $\mathbb{R}$. The set of admissible states is denoted $\mathscr{A}$. Moreover, $\mathbf{R}$ is a smooth function of $\mathbf{U}$ such that for all $\mathbf{U} \in \mathscr{A}$, $\mathbf{R}(\mathbf{U}) \in \mathscr{A}$. Finally, $\gamma(\mathbf{U})$ is a positive real function which represents the stiffness of the source term. The system (1) is assumed to fulfill the properties required in [3] so that it degenerates in long time and when the source term becomes stiff, more precisely when $\gamma(\mathbf{U})t \to \infty$, into a parabolic system. There are

C. Berthon · G. Moebs · R. Turpault (✉)
Laboratoire de Mathématiques Jean Leray, Université de Nantes,
2 rue de la Houssinière, 44322 Nantes Cedex 3, France
e-mail: rodolphe.turpault@univ-nantes.fr

J. Fuhrmann et al. (eds.), *Finite Volumes for Complex Applications VII - Methods and Theoretical Aspects*, Springer Proceedings in Mathematics & Statistics 77,
DOI: 10.1007/978-3-319-05684-5_9,

numerous examples of such systems, including the M_1 model for radiative transfer (see [17]) which is used here as when an illustration is required:

$$\mathbf{U} = \begin{pmatrix} E \\ Fx \\ Fy \\ T \end{pmatrix}, \quad \mathbf{F}(\mathbf{U}) = \begin{pmatrix} Fx & Fy \\ c^2 Pxx & c^2 Pxy \\ c^2 Pyx & c^2 Pyy \\ 0 & 0 \end{pmatrix}, \quad \mathbf{R}(\mathbf{U}) = \begin{pmatrix} \frac{\sigma(\mathbf{U}) a T^4 + \sigma_1(\mathbf{U})}{\sigma_m(\mathbf{U})} \\ \frac{\sigma_1(\mathbf{U}) Fx}{\sigma_m(\mathbf{U})} \\ \frac{\sigma_1(\mathbf{U}) Fy}{\sigma_m(\mathbf{U})} \\ \frac{\sigma(\mathbf{U}) E + \sigma_2(\mathbf{U}) \rho C_v T}{\rho C_v \sigma_m(\mathbf{U})} \end{pmatrix}, \tag{2}$$

where $\gamma(\mathbf{U}) = c\sigma_m(\mathbf{U})$ and

$$P = E\Big(\frac{1-\chi}{2} I_d + \frac{3\chi - 1}{2} \frac{F \otimes F}{\|F\|^2}\Big), \chi = \chi\Big(\xi = \frac{\|F\|}{cE}\Big) = \frac{3 + 4\xi^2}{5 + 2\sqrt{4 - 3\xi^2}}, \tag{3}$$

$$\sigma_m(\mathbf{U}) = \sigma(\mathbf{U}) \max\big(1, \frac{aT^3}{\rho C_v}\big), \sigma_1(\mathbf{U}) = \sigma_m(\mathbf{U}) - \sigma(\mathbf{U}), \sigma_2(\mathbf{U}) = \sigma_m(\mathbf{U}) - \sigma(\mathbf{U}) \frac{aT^3}{\rho C_v}. \tag{4}$$

The set of admissible states is:

$$\mathscr{A} = \big\{\mathbf{U} = (E, F_x, F_y, T)^\top \in \mathbb{R}^4 \,/\, E > 0,\ T > 0, \|F\| \leq cE\big\}. \tag{5}$$

When $\sigma_m(\mathbf{U})t \to \infty$, the M_1 model degenerates into the so-called *equilibrium diffusion equation*:

$$\partial_t\big(\rho C_v T + aT^4\big) - \operatorname{div}\Big(\frac{c}{3\sigma} \nabla aT^4\Big) = 0. \tag{6}$$

The main difficulty when designing a numerical scheme for such systems is to enforce the correct degeneracy in the diffusion limit. In other words, the limit of the scheme when $\gamma(\mathbf{U})t \to \infty$ shall be a consistent approximation of the limit diffusion equation. This property is generally not fulfilled by numerical schemes hence the design of *asymptotic-preserving* (AP) schemes has been an important issue during the last decade. For 1D applications, several asymptotic-preserving schemes were proposed in this context. The most explored way to do so is to use a modified HLL scheme and cleverly control the numerical diffusion in the spirit of the work of Gosse and Toscani for the telegraph equations [20]. This technique has been widely used for the M_1 model and Euler equations with friction (see for instance [8, 10]) and extended to general cases [5]. Other techniques have also been used, such as [1, 6, 11]. The situation is much more difficult for 2D applications however. While it is quite straightforward for Cartesian grids (see [2] for example), it is way more complex on unstructured grids. One of the reasons is that the classical two-point flux scheme (or FV4 [18]) which is the target of many AP schemes is not consistent anymore. The only exception is the MPFA-based AP scheme for Friedrich systems developed in [9]. Our

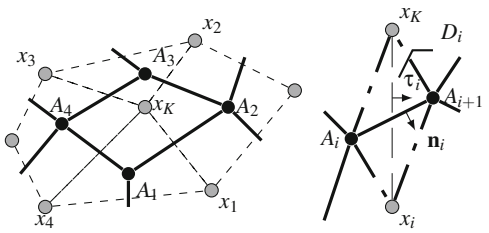


Fig. 1 Local mesh notations (l) one cell and the local notations (r) the diamond cell D_i associated with the ith interface of the cell $K \in \mathscr{M}$

goal is therefore to propose an AP finite volumes scheme for any system of the form (1). This scheme is a natural extension of the 1D scheme proposed in [5] based on the diamond scheme [12]. It is consistent and stable under a natural unrestrictive CFL condition. Moreover, it is also possible to enforce the preservation of the set of admissible states provided a geometrical property is satisfied by the mesh.

Notations. Since we intent to provide a finite volumes scheme which may be used in either *cell-centered* or *vertex-centered* contexts we call (*primary*) mesh $\mathscr{M}$ the set of all control volumes effectively used in the scheme. The *secondary* mesh is a set control volumes defined around the nodes of the *primary* mesh. Hence, the *primary* mesh is the primal mesh in the context of cell-centered schemes and the dual mesh in the context of vertex-centered schemes and the *secondary* mesh is the dual mesh in the context of cell-centered schemes and the primal mesh in the context of vertex-centered schemes. The notations used throughout this paper are summarized on Fig. 1:

- N_K is the number of nodes (and interfaces) of the cell $K \in \mathscr{M}$.
- x_K is the centroid of the cell K.
- The nodes of the cell K are locally denoted $\{A_i\}_{i=1\ldots N_K}$.
- The neighboring cells $(L_i)_i$ of the cell K are locally numbered from 1 to N_K such that $K \cap L_i = [A_i A_{i+1}]$. Their centroids are locally denoted $\{x_i\}_{i=1\ldots N_K}$.
- $d_i^K := \|x_K x_i\|$ and $e_i := \|A_i A_{i+1}\|$ is the length of the i^{th} interface of the cell K.
- $r^K := |K|/p_k$ where p_k is the perimeter of K.

2 Definition of the Scheme and Properties

The scheme proposed here is a direct generalization of the 1D scheme [5] where a Rusanov-type flux is selected for the hyperbolic part. As it was pointed in the introduction, the main difficulty is to select a scheme to degenerate into in the diffusive limit. The classical two-point finite volume scheme (a.k.a FV4 [18]) is not consistent with the diffusion equation on general meshes. The target scheme in the diffusive limit must therefore properly take into account the whole gradient. For the sake of consistency and simplicity, we choose to use the same gradient discretization in the hyperbolic part. Here, we adopt the *diamond scheme* strategy [12] but others could

be considered (see [7, 13–15, 19, 21] and references therein). With the diamond scheme to approximate the gradients, the resulting scheme is obtained:

$$\mathbf{U}_K^{n+1} = \mathbf{U}_i^n + \frac{\Delta t}{|K|}\sum_{i=1}^{N_K} e_i \alpha_{K,i}^n \mathscr{F}_{K,i}^n \cdot \mathbf{n}_i + \frac{\Delta t}{|K|}\sum_{i=1}^{N_K} e_i \alpha_{K,i}^n \mathbf{F}(\mathbf{U}_K^n)\cdot \mathbf{n}_i + \frac{\Delta t}{|K|}\sum_{i=1}^{N_K} e_i (1-\alpha_{K,i}^n) b_i^K (\mathbf{R}(\mathbf{U}_K^n) - \mathbf{U}_K^n), \tag{7}$$

$$\mathscr{F}_{K,i}^n = \frac{\mathbf{F}(\mathbf{U}_K^n) + \mathbf{F}(\mathbf{U}_i^n)}{2} - \frac{b_i^K \theta_i^K}{2}\nabla_i^K \mathbf{U}_K^n \cdot \mathbf{n}_i,\ \alpha_i^K = \frac{b_i^K}{b_i^K + \gamma_i^K r^K}, \tag{8}$$

$$\nabla_i^K \mathbf{U}_K^n \cdot \mathbf{n}_i = \frac{\mathbf{U}_i^n - \mathbf{U}_K^n}{2|D_i|} e_i + \frac{\mathbf{U}_{A_{i+1}}^n - \mathbf{U}_{A_i}^n}{2|D_i|} d_i^K \mathbf{n}_i \cdot \boldsymbol{\tau}_i, \tag{9}$$

where $\theta_i^K > 0$ is a parameter to be precised later and $\mathbf{U}_{A_i}^n$ is the value of the solution at the node A_i (see Fig. 1). This value is obtained as a mean value of the solution in the cells which share A_i as a node (see [12]).

Theorem 1 *Assume that $\theta_i^K \to 0$ when $r^K \to 0$, then the scheme (7)–(8) is consistent with (1).*

The proof of this theorem and the following can be found in [4]. In some applications, it is important to preserve the set of admissible states $\mathscr{A}$. It is all the more difficult since most finite volumes schemes for parabolic problems, including the diamond scheme, do not preserve the maximum principle. Only a few examples ensure this property (for example [16, 22]). Interestingly, it is sometimes possible to recover the maximum principle for our scheme under some geometric condition on the mesh.

Definition 1 The mesh is said to be δ-admissible if $\exists \delta > 0$ such that:

$$\forall K \in \mathscr{M},\ \forall i \in [1, N_K],\ 1 + \frac{e_{i-1}\overline{d^K}_{i-1}}{e_i^2}\frac{|D_i|}{3|D_{i-1}|} - \frac{e_{i+1}\overline{d^K}_{i+1}}{e_i^2}\frac{|D_i|}{3|D_{i+1}|} > \delta,$$
$$\overline{d^K}_i = d_i^K \mathbf{n}_i \cdot \boldsymbol{\tau}_i.$$

With this definition, an admissible mesh is δ-admissible for all $\delta \leq 1$ since all $\overline{d^K}_i$ are then equal to 0. This condition turned out to be satisfied by most of the meshes generated with reasonable constraints on the angles we tested. Equipped with this definition, we can obtain the following result.

Theorem 2 *Assume that the mesh is δ-admissible and that the secondary mesh is made of triangles. Let us also assume that α_i^K is constant inside each cell $K \in \mathscr{M}$ ($\alpha_i^K = \alpha^K$) and let us set $\theta_i^K = \dfrac{2|D_i|}{\delta e_i}$. Then, the scheme (7)–(8) preserves the set of admissible states $\mathscr{A}$ as soon as the following CFL condition holds:*

$$\max_{K\in\mathscr{A}, i\leq N_K} \{b_i^K \theta_i^K \delta_i^K\}\frac{\Delta t}{|K|} p_K \leq \frac{1}{2}. \tag{10}$$

Once again, a proof of this theorem is provided in [4]. There are two essential bricks here: the ability to write the 2D scheme as a convex combination of 1D schemes and the possibility to express the approximate gradient as $\sum_i \omega_{K,i}(U_K - U_i)$ with $\omega_{K,i} \geq 0$. Several comments have to be done concerning this theorem:

- The choice of θ_i^K tends to 0 when $r^K \to 0$ as it was requested for the sake of consistency.
- A similar theorem may be obtained on more general meshes. However, the geometrical condition quickly becomes cumbersome. On the other hand, other expressions of the discrete gradient such as [16] may also be used to ensure the property without any restriction on the mesh at the cost of a strongly nonlinear scheme.
- The main restriction is to consider α_i^K that are constant per cell. It is sometimes a severe limitation when the AP procedure defined in the following is applied.
- Other choices of θ_i^K allow to recover the same result e.g. $\theta_i^K = \max_{i \leq N_K} \dfrac{2|D_i|}{2\delta}$.

The scheme (7)–(8) is not AP in general but a simple procedure may be used to recover this property. It consists in appliying the scheme to the system:

$$\partial_t \mathbf{U} + \operatorname{div}(\mathbf{F}(\mathbf{U})) = (\gamma + \bar{\gamma})\big(\bar{\mathbf{R}}(\mathbf{U}) - \mathbf{U}\big), \ \bar{\mathbf{R}}(\mathbf{U}) = \frac{\gamma \mathbf{R}(\mathbf{U}) + \bar{\gamma}\mathbf{U}}{\gamma + \bar{\gamma}}, \tag{11}$$

which is obviously equivalent to (1). Then, a formal Chapmann-Enskog expansion leads to the following scheme in the diffusion limit:

$$\mathbf{U}_K^{n+1} = \mathbf{U}_K^n - \frac{\Delta t}{|K|} \sum_{i=1}^{N_K} e_i \frac{b_i^K}{(\gamma_i^K + \overline{\gamma^K}_i) r^K} \Big[\mathscr{F}(\mathbf{U}_K^n) \cdot \mathbf{n}_i - \mathbf{F}(\mathbf{U}_K^n) \cdot \mathbf{n}_i \Big]_{|\mathbf{R}(\mathbf{U}_K^n) = \mathbf{U}_K^n}. \tag{12}$$

Now, $\overline{\gamma^K}_i$ may be chosen so that the scheme is AP. This procedure is illustrated in the case of the M_1 model for radiative transfer.

AP correction for the M_1 model. For the M_1 model (2), $b_i^K = c$ and the equilibrium gives $F_x = F_y = 0$ and $E = aT^4$. The first and fourth equations of (12)–(8) hence become:

$$(\rho C_v + aT^4)_K^{n+1} = (\rho C_v T + aT^4)_K^n + \frac{\Delta t}{|K|} \sum_{i=1}^{N_K} \frac{c^2 e_i}{2(\sigma_{m,i}^K + \bar{\sigma}_i^K) r^K} \nabla_i^K (aT^4)^n \cdot \mathbf{n}_i.$$

In order to ensure that this scheme is consistent with the equilibrium diffusion equation (6), the terms $\bar{\sigma}_i^K$ have to be chosen accordingly. For example, if we take:

$$(\sigma_{m,i}^K + \bar{\sigma}_i^K) = \sigma_{m,i}^K \frac{3c\theta_i^K}{2r^K} > 0. \tag{13}$$

then the limit scheme in the diffusion regime is:

$$(\rho C_v T + aT^4)_K^{n+1} = (\rho C_v T + aT^4)_K^n + \frac{\Delta t}{|K|} \sum_{i=1}^{N_K} \frac{ce_i}{3\sigma_i^K} \nabla_i^K (aT^4)_i^n \cdot \mathbf{n}_i,$$

which is consistent with the diffusion equation (6). In fact, as expected, it is nothing but the diamond scheme applied to (6). Moreover, if σ is a constant and $\theta_i^K = \theta^K$ then $\bar{\sigma}_i^K = \bar{\sigma}^K$ and Theorem 2 can be applied. In order to meet such a requirement, one may choose: $\theta_i^K = \max_{i \le N_K} \frac{2|D_i|}{2\delta}$.

3 Numerical Results

Validation tests are performed in this paragraph in order to illustrate the behavior of the scheme. A Riemann problem for the M_1 model for radiative transfer is considered on $[0, 5] \times [0, 1]$ with:

$$(E, F_x, F_y, T)^\top (0, x) = \begin{cases} (aT_L^4, cf_{x,L} aT_L^4, 0, T_L)^\top, \text{ if } x < 1, \\ (aT_R^4, 0, 0, T_R)^\top, \text{ otherwise.} \end{cases}$$

In the following, $T_L = 10000$, $T_R = 300$ and $f_{x,L} = 0$, $\rho C_v = 10^{-2}$ and $c = 3.10^8$. First, σ is set to 0 since the preservation of admissible states is expected to be more difficult than in the presence of the (regularizing) source-term. Two different meshes are used: a "coarse" one (5152 triangles) and a "fine" one (132006 triangles). Both of these meshes are $\delta-$admissible with optimal $\delta = \delta_1 = 1.095$ for the coarse grid and $\delta_2 = 5.59910^{-2}$ for the fine one. The reference solution is the exact solution of the corresponding 1D Riemann problem. Figure 2 shows the solutions along $x = \frac{1}{2}$. Here, the conservation of admissible states is enforced by using $\theta_i^K = \max_{i \le N_K} \frac{2|D_i|}{2\delta}$ where $\delta = \delta_1$ on the coarse mesh and δ_2 on the fine one. The solution computed on the coarse grid is comparable to a 1D Rusanov scheme with a similar number of cells. On the other hand, since $\delta_2 \ll \delta_1$, the numerical diffusion of the scheme is way larger on the fine mesh than on the coarse one. The approximation is hence better on the coarse grid in this case. Now if $\delta = \delta_1$ on the fine mesh, as shown in the right of Fig. 2, the quality of the approximation behaves as expected, i.e. the approximation is better on the fine grid.

Next, we fix $\sigma = 1000$ to investigate the AP property. The results showed on Fig. 3 are compared with a grid-converged 1D approximation of the equilibrium diffusion equation. The tests are performed with and without the asymptotic-preserving correction on the fine grid. We immediately see that with the AP correction, the scheme provides an approximation which is nearly indistinguishable from the reference solution. On the other hand, as expected, if the AP correction is turned off i.e. $\overline{\gamma^K} = 0$),

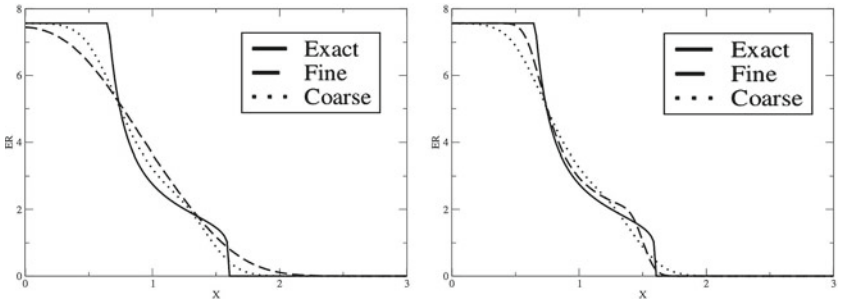


Fig. 2 Exact and computed E along $x = \frac{1}{2}$ with $\sigma = 0$ at $t = 2.10^{-9}$. (l) conservation of $\mathscr{A}$ enforced; (r) same value of δ for both meshes

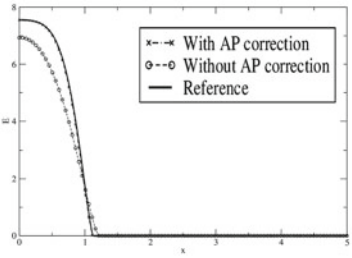


Fig. 3 Ref. and computed E with and w/o AP correction along $x = \frac{1}{2}$ with $\sigma = 1000$ at $t = 2.10^{-6}$

Fig. 4 Radiative flow in a channel (*top*) E (*bottom*) χ

there is a large discrepancy between the computed and the reference solution. Finally, a test-case involving the evolution of the radiation in a channel with multiple obstacles is performed. The entry condition on the left side of the channel models a beam of high energy ($F_L = cE_L = ca10000^4$) compared to the initial state of the domain ($F_0 = 0$, $E_0 = a10^4$), $\sigma = 1$ and 11 obstacles (with wall boundary conditions) are scattered in the channel. A vertex-centered approached is used on a mesh consisting of 15348 cells refined near the obstacles (see Fig. 4). Let us emphasize that this case is numerically very challenging and that it is all the more critical to preserve the set of admissible states here since very small numerical errors may yield

unadmissible values, which immediately cause the code to crash. Several values of θ_i^K were tested and even a value 5 % larger than the choice stated in the theorem produces unadmissible results. In this sense, it seems that the condition of Theorem 2 is optimal.

Acknowledgments This work was supported by ANR-12-IS01-0004-01 GEONUM.

References

1. Aregba-Driollet, D., Briani, M., Natalini, R.: Time asymptotic high order schemes for dissipative bgk hyperbolic systems. arXiv:1207.6279v1 (2012)
2. Berthon, C., Dubois, J., Dubroca, B., Nguyen-Bui, T.H., Turpault, R.: A free streaming contact preserving scheme for the M_1 model. Adv. Appl. Math. Mech. **2**(3), 259–285 (2010)
3. Berthon, C., LeFloch, P.G., Turpault, R.: Late time/stiff relaxation asymptotic preserving approximations of hyperbolic equations. Math. Comp. **82**, 831–860 (2013)
4. Berthon, C., Moebs, G., Sarazin-Desbois, C., Turpault, R.: An asymptotic-preserving scheme for systems of conservation laws with source terms on 2D unstructured meshes. submitted
5. Berthon, C., Turpault, R.: AP HLL schemes. NMPDE **27**, 1396–1422 (2011)
6. Bouchut, F., Ounaissa, H., Perthame, B.: Upwinding of the source term at interfaces for Euler equations with high friction. Comput. Math. Appl. **53**, 361–375 (2007)
7. Boyer, F., Hubert, F.: Finite volume method for 2D linear and nonlinear elliptic problems with discontinuities. SIAM J. Numer. Anal. **46**(6), 3032–3070 (2008)
8. Buet, C., Després, B.: AP and positive schemes for radiation hydrodynamics. J. Comp. Phys. **215**, 717–740 (2006)
9. Buet, C., Després, B., Frank, E.: Design of asymptotic preserving finite volume schemes for the hyperbolic heat equation on unstructured meshes. Numer. Math **122**(2), 227–278 (2012)
10. Chalons, C., Coquel, F., Godlewski, E., Raviart, P., Seguin, N.: Godunov-type schemes for hyperbolic systems with parameter dependent source. The case of Euler system with friction. Math. Models Meth. Appl. Sci. **20**(11), 2109–2166 (2010)
11. Chalons, C., Girardin, M., Kokh, S.: Large time-step and asymptotic-preserving numerical schemes for the gas dynamics equations with source terms. SIAM J. Sci Comput. (2014). http://hal.archives-ouvertes.fr/hal-00718022
12. Coudière, Y., Vila, J., Villedieu, P.: Convergence rate of a finite volume scheme for a two dimensional convection-diffusion problem. M2AN **33**(3), 493–516 (1999)
13. Domelevo, K., Omnès, P.: A finite volume method for the Laplace equation on almost arbitrary two-dimensional grids. Math. Model. Numer. Anal. **39**(6), 1203–1249 (2005)
14. Droniou, J., Eymard, R., Gallouët, T., Herbin, R.: A unified approach to mimetic finite difference, hybrid finite volume and mixed finite volume methods. M3AS **20**(2), 265–295 (2010)
15. Droniou, J., Eymard, R., Gallouët, T., Herbin, R.: Gradient schemes: a generic framework for the discretisation of linear, nonlinear & nonlocal elliptic & parabolic equations. M3AS **23**(13), 2395–2432 (2013)
16. Droniou, J., Le Potier, C.: Construction and convergence study of schemes preserving the elliptic local maximum principle. SIAM J. Numer. Anal. **49**, 459–490 (2011)
17. Dubroca, B., Feugeas, J.: Entropic moment closure hierarchy for the radiative transfer equation. C. R. Acad. Sci. Paris, Ser. I **329**, 915–920 (1999)
18. Eymard, R., Gallouët, T., Herbin, R.: Finite Volume Methods: Handbook of Numerical Analysis, vol. 7. North-Holland, Amsterdam (2000)
19. Eymard, R., Gallouët, T., Herbin, R.: Discretisation of heterogeneous and anisotropic diffusion problems on general non-conforming meshes. SUSHI: a scheme using stabilisation and hybrid interfaces. IMA J. Numer. Anal. **30**(1), 1009–1043 (2009)

20. Gosse, L., Toscani, G.: AP well-balanced scheme for the hyperbolic heat equations. C. R. Math. Acad. Sci. Paris **334**, 337–342 (2002)
21. Hermeline, F.: Approximation of diffusion operators with discontinuous tensor coefficients on distorted meshes. Comput. Meth. Appl. Mech. Eng. **192**, 1939–1959 (2003)
22. Le Potier, C.: A nonlinear correction and maximum principle for diffusion operators discretized using cell-centered finite volume schemes. C.R. Math. **348**(11–12), 691–695 (2010)

Numerical Dissipation and Dispersion of the Homogeneous and Complete Flux Schemes

J. H. M. ten Thije Boonkkamp and M. J. H. Anthonissen

Abstract We analyse numerical dissipation and dispersion of the homogeneous flux (HF) and complete flux (CF) schemes, finite volume methods introduced in [4]. To that purpose we derive the modified equation of both schemes. We show that the HF scheme suffers from numerical diffusion for dominant advection, which is effectively removed in the CF scheme. The latter scheme, however, is prone to numerical dispersion. We validate both schemes for a model problem.

1 Introduction

Conservation laws are often of advection-diffusion-reaction type, describing the interplay between different processes such as, e.g., drift, diffusion and generation/recombination. They occur in disciplines like combustion theory, plasma physics, transport in porous media etc. A model equation for these conservation laws is the advection-diffusion-reaction equation. In [4] we introduced the complete flux scheme for this equation. For steady problems, the complete flux approximation is based on the solution of a local BVP for the *entire* equation, including the source term. Consequently, the numerical flux consists of a homogeneous component, depending on the advection-diffusion operator, and an inhomogeneous component, depending on the source term. In many applications, the homogeneous component is known as the exponential scheme, however, we refer to it as the homogeneous flux scheme. The inclusion of the inhomogeneous flux is especially of importance for dominant

J. H. M. ten Thije Boonkkamp (✉) · M. J. H. Anthonissen (✉)
Department of Mathematics and Computer Science, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands
e-mail: j.h.m.tenthijeboonkkamp@tue.nl

M. J. H. Anthonissen
e-mail: m.j.h.anthonissen@tue.nl

J. Fuhrmann et al. (eds.), *Finite Volumes for Complex Applications VII - Methods and Theoretical Aspects*, Springer Proceedings in Mathematics & Statistics 77,
DOI: 10.1007/978-3-319-05684-5_10,

advection, since it it guarantees that the flux approximation remains second order accurate, even for infinite Péclet number [1].

For time-dependent problems we include the time-derivative in the inhomogeneous flux. We will demonstrate that also this term is very important for advection-dominant problems, since it effectively removes the artificial diffusion introduced by the homogeneous flux scheme, at least up to second order, albeit at the cost of a small dispersion error. In order to analyse this, we consider the model equation

$$\frac{\partial \varphi}{\partial t} + \frac{\partial}{\partial x}\left(u\varphi - \varepsilon \frac{\partial \varphi}{\partial x}\right) = s, \tag{1}$$

where u is the advection velocity, $\varepsilon \geq \varepsilon_{\min} > 0$ the diffusion coefficient and s the source term. Associated with (1) we introduce the the flux f, defined by

$$f = u\varphi - \varepsilon \frac{\partial \varphi}{\partial x}. \tag{2}$$

To analyse the dissipation and dispersion errors of the HF and CF scheme, we derive the modified equation for both schemes, which is roughly speaking the partial differential equation that is *exactly* solved by the numerical solution. The modified equation of the HF scheme contains an artificial diffusion term, which suppresses spurious oscillations, however, it makes the scheme very dissipative for dominant advection. This artificial diffusion term is eliminated by the time derivative term in the inhomogeneous flux, making the scheme nondissipative, which is a very benificial property especially for long time integration. Instead, the modified equation contains a leading order dispersion term, responsible for possible spurious oscillations.

We have organised our paper as follows. The derivation of the HF and CF schemes is briefly outlined in Sect. 2. In Sect. 3 we derive the modified equation for both schemes and interpret these in terms of dissipation and dispersion. The performance of both schemes is demonstrated in Sect. 4, and finally in Sect. 5, we present a discussion and conclusions.

2 Numerical Approximation of the Flux

In this section we outline the complete flux scheme for Eq. (1), which is a special case of the scheme introduced in [4].

Equation (1) can be written as $\partial\varphi/\partial t + \partial f/\partial x = s$ with the flux f defined in (2). Assume this equation is defined on the domain $\Omega = [0, 1]$. In the finite volume method we cover Ω with a finite number of control volumes (cells) Ω_j of size h. We adopt the vertex-centred approach, i.e., we introduce the grid points $x_j = jh$ $(j = 0, 1, 2, \ldots, N)$ with $Nh = 1$, where the variable φ has to be approximated, and choose $\Omega_j = [x_{j-1/2}, x_{j+1/2}]$ with $x_{j\pm1/2} = \frac{1}{2}\left(x_j + x_{j\pm1}\right)$. Integrating the equation over Ω_j and applying the midpoint rule for the integrals of s and $\partial\varphi/\partial t$, we obtain the semi-discrete equation

$$h\,\dot{\varphi}_j(t) + F_{j+1/2}(t) - F_{j-1/2}(t) = h\,s_j(t), \tag{3}$$

where $\dot{\varphi}_j(t) \approx \partial\varphi/\partial t(x_j, t)$ and $F_{j+1/2}(t)$ is the numerical approximation of the flux at the interface $x = x_{j+1/2}$. In the following we will suppress the dependence of all variables on t.

We determine the numerical flux $F_{j+1/2}$ from the following quasi-stationary boundary value problem

$$\frac{\partial f}{\partial x} = \frac{\partial}{\partial x}\Big(u\varphi - \varepsilon\frac{\partial\varphi}{\partial x}\Big) = s - \frac{\partial\varphi}{\partial t}, \quad x_j < x < x_{j+1}, \tag{4a}$$

$$\varphi(x_j) = \varphi_j, \quad \varphi(x_{j+1}) = \varphi_{j+1}, \tag{4b}$$

where we have put the time derivative in the right hand side of Eq. (4a). As a consequence, the numerical flux will depend on the time derivative, and this will turn out to be of importance for dominant advection. In the derivation that follows we assume u and ε constant. Moreover, we introduce the following variables

$$a = \frac{u}{\varepsilon}, \quad P = ah, \quad \sigma(x) = \frac{x - x_j}{h}. \tag{5}$$

P is the well-known (grid) Péclet number and $\sigma(x)$ is the normalized coordinate on $[x_j, x_{j+1}]$ $(0 \le \sigma(x) \le 1)$. Integrating equation (4a) from $x_{j+1/2}$ to $x \in [x_j, x_{j+1}]$, we obtain the integral balance

$$f(x) - f(x_{j+1/2}) = \int_{x_{j+1/2}}^{x} \hat{s}(\xi)\,\mathrm{d}\xi, \quad \hat{s} = s - \frac{\partial\varphi}{\partial t}. \tag{6}$$

Next, substituting the integrating factor formulation

$$f(x) = -\varepsilon\,\mathrm{e}^{ax}\frac{\partial}{\partial x}\big(\mathrm{e}^{-ax}\varphi\big) \tag{7}$$

in (6) and integrating the resulting equation form x_j to x_{j+1}, we get

$$\begin{aligned}&\varepsilon\big(\mathrm{e}^{-ax_{j+1}}\varphi_{j+1} - \mathrm{e}^{-ax_j}\varphi_j\big) + \tfrac{1}{a}\big(\mathrm{e}^{-ax_j} - \mathrm{e}^{-ax_{j+1}}\big)f(x_{j+1/2})\\ &\quad = \int_{x_j}^{x_{j+1}}\int_{x_{j+1/2}}^{x}\mathrm{e}^{-ax}\hat{s}(\xi)\,\mathrm{d}\xi\,\mathrm{d}x.\end{aligned} \tag{8}$$

Finally, changing the order of integration in the double integral in the right hand side of (8), we find the following expressions for the flux

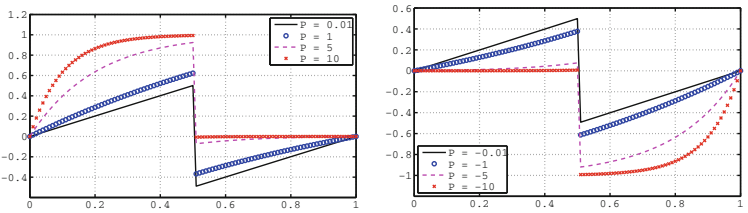


Fig. 1 Green's function $G(\sigma;\, P)$ for the flux for $P > 0$ (*left*) and $P < 0$ (*right*).

$$f(x_{j+1/2}) = f^{\mathrm{h}}(x_{j+1/2}) + f^{\mathrm{i}}(x_{j+1/2}), \tag{9a}$$

$$f^{\mathrm{h}}(x_{j+1/2}) = \frac{\varepsilon}{h}\big(B(-P)\varphi_j - B(P)\varphi_{j+1}\big), \tag{9b}$$

$$f^{\mathrm{i}}(x_{j+1/2}) = h\int_0^1 G(\sigma;\, P)\hat{s}(x(\sigma))\,\mathrm{d}\sigma, \tag{9c}$$

where $B(z) = z/\big(\mathrm{e}^z - 1\big)$ and where $G(\sigma;\, P)$ is the so-called *Green's function for the flux*, given by

$$G(\sigma;\, P) = \begin{cases} \dfrac{1-\mathrm{e}^{-P\sigma}}{1-\mathrm{e}^{-P}} & \text{for} \quad 0 \le \sigma \le \frac{1}{2}, \\[2ex] -\dfrac{1-\mathrm{e}^{P(1-\sigma)}}{1-\mathrm{e}^{P}} & \text{for} \quad \frac{1}{2} < \sigma \le 1; \end{cases} \tag{10}$$

see Fig. 1. From (9) it is evident that the flux is the sum of the homogeneous component f^{h}, depending on the advection-diffusion operator, and the inhomogeneous component f^{i}, depending the modified source $\hat{s}$.

Obviously, the numerical flux is the sum of a homogeneous component, $F^{\mathrm{h}}_{j+1/2}$, and an inhomogeneous component, $F^{\mathrm{i}}_{j+1/2}$. For the homogeneous component we simply take (9b), i.e., $F^{\mathrm{h}}_{j+1/2} = f^{\mathrm{h}}(x_{j+1/2})$. For the inhomogeneous component we need to evaluate the integral in (9c). Note that for dominant diffusion ($|P| \ll 1$) the integral (average) of $G(\sigma;\, P)$ is small, whereas for dominant advection ($|P| \gg 1$) $G(\sigma;\, P)$ has a clear bias towards the upwind side of the interval. For this reason we replace $\hat{s}(x(\sigma))$ in (9c) by its upwind value and evalute the resulting integral. This way we obtain

$$F_{j+1/2} = F^{\mathrm{h}}_{j+1/2} + h\left(\tfrac{1}{2} - W(P)\right)\big(s_{\mathrm{u},\,j+1/2} - \dot{\varphi}_{\mathrm{u},\,j+1/2}\big), \tag{11}$$

where $W(z) = \big(\mathrm{e}^z - 1 - z\big)/\big(z\big(\mathrm{e}^z - 1\big)\big)$ and where $s_{\mathrm{u},\,j+1/2}$ denotes the upwind value of s relative to the interface $x_{j+1/2}$, i.e., $s_{\mathrm{u},\,j+1/2} = s_j$ if $u \ge 0$ and $s_{\mathrm{u},\,j+1/2} = s_{j+1}$ if $u < 0$, and likewise for $\dot{\varphi}$.

We refer to the flux approximation in (11) as the complete flux (CF) scheme, as opposed to the homogeneous flux (HF) scheme, where we omit the source term

and time derivative. Finally, substituting this expression in (3) we obtain the semi-discretisation

$$\begin{aligned}\left(\tfrac{1}{2}+W(|P|)\right)\dot{\varphi}_j+\left(\tfrac{1}{2}-W(|P|)\right)\dot{\varphi}_{j(\mathrm{u})}+\tfrac{1}{h}\left(F^{\mathrm{h}}_{j+1/2}-F^{\mathrm{h}}_{j-1/2}\right)=\\ \left(\tfrac{1}{2}+W(|P|)\right)s_j+\left(\tfrac{1}{2}-W(|P|)\right)s_{j(\mathrm{u})},\end{aligned} \tag{12}$$

where $j(\mathrm{u})$ is the index of the grid point upwind of j, i.e., $j(\mathrm{u})=j-1$ if $u\geq 0$ and $j(\mathrm{u})=j+1$ is $u<0$. If we set $W(|P|)=\frac{1}{2}$ in (12), we get the HF semidisretisation.

3 Numerical Dissipation and Dispersion

In this section we investigate the semi-discrete system (12) in terms of dissipation and dispersion. We consider both the HF and CF scheme. In [3] we presented a detailed analysis based on the evolution of a planar wave solution, here we adopt a different approach, viz., we derive the modified equation for both schemes.

Roughly speaking, the modified equation of a finite difference scheme is defined as the partial differential equation that is *actually* solved by the numerical solution, apart from rounding errors [5]. Assume ψ to be a sufficiently smooth function coinciding with the numerical solution on the space-time grid. Expanding all differences in Taylor series, we obtain the original differential equation with an extra local discretisation error in the right hand side, which is an infinite sequence of derivatives. If we subsequently eliminate all time derivatives except the first order, we obtain the so-called modified equation. We will slightly adapt this procedure to our purpose.

We introduce the following difference operators for a generic grid function v_j

$$\begin{aligned}\delta_x^- v_j&=\tfrac{1}{h}\left(v_j-v_{j-1}\right),\quad \delta_x v_j=\tfrac{1}{2h}\left(v_{j+1}-v_{j-1}\right),\\ \delta_{xx}v_j&=\tfrac{1}{h^2}\left(v_{j+1}-2v_j+v_{j-1}\right).\end{aligned} \tag{13}$$

In the following we assume that $s=0$ and $u>0$. We first consider the HF scheme, which can be written as

$$\dot{\varphi}_j+\frac{\varepsilon}{h^2}\left(B^-\left(\varphi_j-\varphi_{j-1}\right)-B^+\left(\varphi_{j+1}-\varphi_j\right)\right)=0, \tag{14}$$

where $B^{\pm}=B(\pm P)$. Rearranging terms, we see that the scheme is equivalent to

$$\dot{\varphi}_j+u\delta_x\varphi_j-\varepsilon_{\mathrm{art}}\delta_{xx}\varphi_j=0,\quad \varepsilon_{\mathrm{art}}=D\left(\tfrac{1}{2}P\right)\varepsilon, \tag{15}$$

with $D(z)=z\coth(z)$, see Fig. 2. The artificial diffusion coefficient $\varepsilon_{\mathrm{art}}$ is the sum of ε and the numerical diffusion coefficient $\varepsilon_{\mathrm{num}}=P\left(\frac{1}{2}-W(P)\right)\varepsilon$. To derive the modified equation of the *semi-discretisation* (15), we substitute a (sufficiently smooth) function ψ, satisfying $\psi(x_j,t)=\varphi_j(t)$ for all grid points x_j, and expand all *spatial* differences in Taylor series. Moreover, we discard all $\mathcal{O}\left(h^2\right)$-terms. This

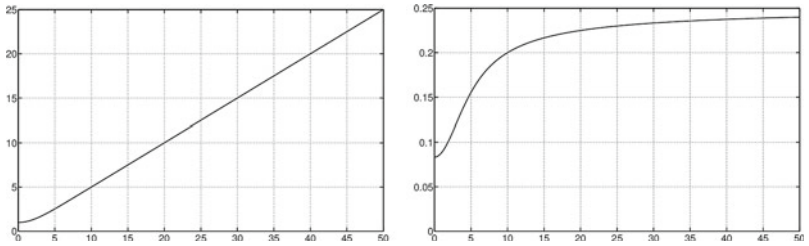


Fig. 2 The functions $D\left(\frac{1}{2}z\right)$ *(left)* and $E(z)$ *(right)*

way we obtain the modified equation

$$\frac{\partial \psi}{\partial t} + u\frac{\partial \psi}{\partial x} = \varepsilon_{\text{art}}\frac{\partial^2 \psi}{\partial x^2}. \tag{16}$$

So the HF scheme suffers from numerical diffusion, where the artificial diffusion coefficient ε_{art} increases from ε for $P = 0$ to $\frac{1}{2}uh$ for $P \to \infty$ (upwind limit). Unlike standard central differences, the scheme does not display spurious oscillations since $0 < uh/\varepsilon_{\text{art}} < 2$.

Next, we consider the CF scheme, which reads

$$\left(\tfrac{1}{2} - W^+\right)\dot{\varphi}_{j-1} + \left(\tfrac{1}{2} + W^+\right)\dot{\varphi}_j + \frac{\varepsilon}{h^2}\left(B^-\left(\varphi_j - \varphi_{j-1}\right) - B^+\left(\varphi_{j+1} - \varphi_j\right)\right) = 0, \tag{17}$$

where $W^+ = W(P)$. Again, rearranging terms we obtain

$$\left(\text{id} - \alpha\delta_x^-\right)\dot{\varphi}_j + u\delta_x\varphi_j - \varepsilon_{\text{art}}\delta_{xx}\varphi_j = 0, \quad \alpha = h\left(\tfrac{1}{2} - W^+\right), \tag{18}$$

where id is the identity operator. Note that for $\alpha \neq 0$ this scheme defines an implicit ODE-system, which is a result of the quasi-stationary assumption made in (4). To derive the modified equation we first make the ODE-system explicit by applying the inverse operator $\left(\text{id} - \alpha\delta_x^-\right)^{-1}$ to (18), for which the following relation holds

$$\left(\text{id} - \alpha\delta_x^-\right)^{-1} = \text{id} + \alpha\delta_x^- + \mathcal{O}\left(h^2\right). \tag{19}$$

Applying the inverse to (18), the advection term gives rise to the antidiffusion term

$$-\alpha u\delta_x^-\delta_x\varphi_j = P\left(W(P) - \tfrac{1}{2}\right)\varepsilon\delta_{xx}\varphi_j + \mathcal{O}\left(h^2\right),$$

which exactly cancels the numerical diffusion introduced by the HF scheme up to $\mathcal{O}\left(h^2\right)$, and the diffusion term gives rise to the dispersive term $\alpha\varepsilon_{\text{art}}\delta_x^-\delta_{xx}\varphi_j$. Thus the CF scheme is equivalent to

$$\dot{\varphi}_j + u\delta_x\varphi_j = \varepsilon\delta_{xx}\varphi_j + \nu\delta_x^-\delta_{xx}\varphi_j + \mathcal{O}\left(h^2\right), \quad \nu = E(P)uh^2, \tag{20}$$

where $E(z) = D\left(\frac{1}{2}z\right)\left(\frac{1}{2} - W(z)\right)/z$; see Fig. 2. ν is the numerical dispersion coefficient. Since $\delta_x^- \delta_{xx}\varphi_j$ is a first order approximation of $\partial^3\varphi/\partial x^3(x_j)$ we conclude that the CF scheme is a second order accurate approximation of the modified equation

$$\frac{\partial \psi}{\partial t} + u\frac{\partial \psi}{\partial x} = \varepsilon\frac{\partial^2 \psi}{\partial x^2} + \nu\frac{\partial^3 \psi}{\partial x^3}. \tag{21}$$

The third order term $\nu\partial^3\varphi/\partial x^3$ is reponsible for dispersion, i.e., waves of different frequencies propagate at different speed.

It is instructive to consider the advection-reaction equation as special case, i.e., $\varepsilon = 0$. In this case $P \to \infty$ and $\nu = \frac{1}{4}uh^2$. Substituting the planar wave $\psi(x,t) = \mathrm{e}^{\mathrm{i}(\kappa x - \omega t)}$ in (21), with κ the wave number and ω the frequency, we obtain the dispersion relation $\omega(\kappa) = u\kappa + \nu\kappa^3$ from which we infer that the numerical phase velocity $c_\mathrm{p}(\kappa) = \omega(\kappa)/\kappa = u\left(1 + \frac{1}{4}(\kappa h)^2\right)$. Since $\nu > 0$, the numerical solution propagates (a little) too fast.

To compute the full numerical solution, we have to apply a time integrator to (12). Since the CF scheme is nondissipative, the trapezoidal rule is an obvious choice. Combined with the CF scheme for the advection equation, the trapezoidal rule reduces to the box scheme, which is second order and nondissipative indeed. However, the dispersion error of the combined scheme differs from the dispersion error of the CF scheme alone, this is due to the trapzoidal rule [2, pp. 379–381].

4 Numerical Example

In this section we apply the HF and CF scheme to a model problem. In [3] we investigated the order of convergence of the schemes, and we will not repeat this here. The main conclusions, however, are the following. For dominant diffusion both the HF and CF scheme exhibit second order convergence. On the other hand, for dominant advection, the HF scheme reduces to the first order upwind scheme, whereas the CF scheme remains second order convergent. Thus, inclusion of the source term and time derivative is important for dominant advection!

Numerical dissipation and dispersion can be conveniently demonstrated for the advection equation, therefore we choose $\varepsilon = 0$ and $s = 0$. Furthermore, we choose the initial and boundary conditions such that the exact solution is given by the wave packet

$$\varphi(x,t) = \mathrm{e}^{-\lambda\xi^2}\sin(\kappa\xi), \quad \xi = x - x_0 - ut. \tag{22}$$

Numerical approximations of (22) together with its (Gaussian) envelope are shown in Fig. 3. Clearly, the HF scheme is very dissipative. At $t = 5\times10^{-2}$, the amplitude has decreased to approximately 0.2 times its initial value while at $t = 0.5$ the solution has completely vanished. On the other hand, the CF numerical solution has not dissipated and is clearly much better than the HF solution. In fact for this choice of h and time

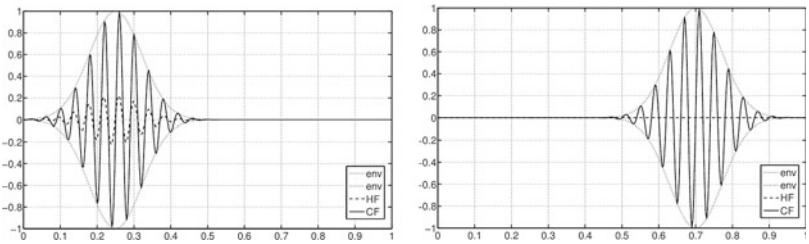


Fig. 3 HF and CF numerical solutions for the advection equation at $t = 0.05$ (*left*) and at $t = 0.5$ (*right*). Parameter values are $u = 1$, $\lambda = 10^2$, $\kappa = 50\pi$, $x_0 = 0.2$ and $h = \Delta t = 2.5 \times 10^{-3}$

step Δt, the Courant number $c = u\Delta t/h = 1$, and the exact solution is recovered. Other choices of Δt give rise to numerical dispersion, which should be attributed to the time integrator, however.

5 Concluding Remarks and Discussion

We have presented the HF and CF scheme for an advection-diffusion-reaction model problem. The CF scheme is based on a quasi-stationary boundary value problem, including the time derivative in the source term. Consequently, the numerical flux consists of a homogeneous component, depending on the advection-diffusion operator, and an inhomogeneous component, containing the source and time derivative. We have derived the modified equation for both schemes and demonstrated that the HF scheme is prone to numerical diffusion, which is completely removed in the CF scheme. Instead, the CF scheme suffers from a dispersion error. We have shown that the CF scheme is much more accurate than the HF scheme when applied to the advection equation.

Both schemes are applicable to complex applications like plasma systems. In [1] we have shown that the CF scheme is much more accurate than the HF scheme for the numerical simulation of a glow discharge. The conservation laws are in this case nonlinear and coupled, which is no objection since the original HF and CF schemes developed in [4] only assume coefficients that are functions of space and time.

References

1. Liu, L., van Dijk, J., ten Thije Boonkkamp, J., Mihailova, D., van de Mullen, J.: The complete flux scheme—error analysis and application to plasma simulation. J. Comput. Appl. Math. **250**, 229–243 (2013)
2. Mattheij, R., Rienstra, S., ten Thije Boonkkamp, J.: Partial Differential Equations, Modeling, Analysis, Computation. SIAM, Philadelphia (2005)

3. ten Thije Boonkkamp, J., Anthonissen, M.: Extension of the complete flux scheme to time-dependent conservation laws. In: Kreiss, G et al. (ed.) Numerical mathematics and advanced applications 2009, pp. 865–873. (2010)
4. ten Thije Boonkkamp, J., Anthonissen, M.: The finite volume-complete flux scheme for advection-diffusion-reaction equations. J. Sci. Comput. **46**(1), 47–70 (2011)
5. Warming, R., Hyett, B.: The modified equation approach to the stability and accuracy analysis of finite-difference methods. J. Comput. Phys. **14**, 159–179 (1974)

A New Finite Volume Scheme for a Linear Schrödinger Evolution Equation

Abdallah Bradji

Abstract We consider the linear Schrödinger evolution equation with a time dependent potential in several space dimension. We provide a new implicittime finite volume scheme, using the general nonconforming meshes of [2] as discretization in space. We prove that the convergence order is $h_{\mathscr{D}} + k$, where $h_{\mathscr{D}}$ (resp. k) is the mesh size of the spatial (resp. time) discretization, in discrete norms $\mathbb{L}^{\infty}(0, T; H_0^1(\Omega))$ and $\mathscr{W}^{1,\infty}(0, T; L^2(\Omega))$. These error estimates are useful because they allow to obtain approximations to the exact solution and its first derivatives of order $h_{\mathscr{D}} + k$.

1 Motivation and Aim of This Paper

Let us consider the following linear time dependent Schrödinger problem. We seek a complex valued function u defined on $\Omega \times [0, T]$ satisfying

$$i\, u_t(x,t) + \Delta u(x,t) - V(x,t)u(x,t) = f(x,t), \ \ (x,t) \in \Omega \times (0,T), \quad (1)$$

where Ω is an open bounded polyhedral subset in $\mathbb{R}^d$, with $d \in \mathbb{N} \setminus \{0\}$, $T > 0$, $i \in \mathbb{C}$ (the set of complex numbers) is the imaginary unit, V is a time dependent potential and f is a given function.

An initial condition is given by:

$$u(x,0) = u^0(x), \ \ x \in \Omega, \quad (2)$$

A. Bradji(✉)
Department of Mathematics, Faculty of Sciences, University of Badji Mokhtar-Annaba, Annaba, Algeria
e-mail: abdallah-bradji@univ-annaba.org, bradji@cmi.univ-mrs.fr

J. Fuhrmann et al. (eds.), *Finite Volumes for Complex Applications VII - Methods and Theoretical Aspects*, Springer Proceedings in Mathematics & Statistics 77,
DOI: 10.1007/978-3-319-05684-5_11,

with homogeneous Dirichlet boundary conditions, that is

$$u(x,t) = 0, \ \ (x,t) \in \partial\Omega \times (0,T), \tag{3}$$

The form (1)–(3) of Schrödinger equation occurs, for example, when $d = 1$ in underwater acoustics, cf. [1]. The model (1)–(3) is studied for instance in [1] when a Galerkin finite element method is used as discretization in space. The stationary case of Schrödinger equation is also considered using finite volume methods in [3] where there are some interesting numerical tests. In this work we analyze a new finite volume scheme for the Schrödinger evolution problem (1)–(3).

2 Definition of the Scheme and Statement of the Main Result

The discretization of Ω is performed using the mesh $\mathscr{D} = (\mathscr{M}, \mathscr{E}, \mathscr{P})$ described in [2, Definition 2.1] which we recall here for the sake of completeness.

Definition 1 (Definition of the spatial mesh) Let Ω be a polyhedral open bounded subset of $\mathbb{R}^d$, where $d \in \mathbb{N} \setminus \{0\}$, and $\partial\Omega = \overline{\Omega} \setminus \Omega$ its boundary. A discretisation of Ω, denoted by $\mathscr{D}$, is defined as the triplet $\mathscr{D} = (\mathscr{M}, \mathscr{E}, \mathscr{P})$, where:

1. $\mathscr{M}$ is a finite family of non empty connected open disjoint subsets of Ω (the "control volumes") such that $\overline{\Omega} = \cup_{K\in\mathscr{M}} \overline{K}$. For any $K \in \mathscr{M}$, let $\partial K = \overline{K} \setminus K$ be the boundary of K; let $\mathrm{m}(K) > 0$ denote the measure of K and h_K denote the diameter of K.
2. $\mathscr{E}$ is a finite family of disjoint subsets of $\overline{\Omega}$ (the "edges" of the mesh), such that, for all $\sigma \in \mathscr{E}$, σ is a non empty open subset of a hyperplane of $\mathbb{R}^d$, whose $(d-1)$–dimensional measure is strictly positive. We also assume that, for all $K \in \mathscr{M}$, there exists a subset $\mathscr{E}_K$ of $\mathscr{E}$ such that $\partial K = \cup_{\sigma\in\mathscr{E}_K} \overline{\sigma}$. For any $\sigma \in \mathscr{E}$, we denote by $\mathscr{M}_\sigma = \{K;\ \sigma \in \mathscr{E}_K\}$. We then assume that, for any $\sigma \in \mathscr{E}$, either $\mathscr{M}_\sigma$ has exactly one element and then $\sigma \subset \partial\Omega$ (the set of these interfaces, called boundary interfaces, denoted by $\mathscr{E}_{\mathrm{ext}}$) or $\mathscr{M}_\sigma$ has exactly two elements (the set of these interfaces, called interior interfaces, denoted by $\mathscr{E}_{\mathrm{int}}$). For all $\sigma \in \mathscr{E}$, we denote by x_σ the barycentre of σ. For all $K \in \mathscr{M}$ and $\sigma \in \mathscr{E}$, we denote by $\mathbf{n}_{K,\sigma}$ the unit vector normal to σ outward to K.
3. $\mathscr{P}$ is a family of points of Ω indexed by $\mathscr{M}$, denoted by $\mathscr{P} = (x_K)_{K\in\mathscr{M}}$, such that for all $K \in \mathscr{M}$, $x_K \in K$ and K is assumed to be x_K–star-shaped, which means that for all $x \in K$, the property $[x_K, x] \subset K$ holds. Denoting by $d_{K,\sigma}$ the Euclidean distance between x_K and the hyperplane including σ, one assumes that $d_{K,\sigma} > 0$. We then denote by $\mathscr{D}_{K,\sigma}$ the cone with vertex x_K and basis σ.

The time discretization is performed with a constant time step $k = \frac{T}{N+1}$, where $N \in \mathbb{N}^\star$, and we shall denote $t_n = nk$, for $n \in [\![0, N+1]\!]$. Throughout this paper, the letter C stands for a positive constant independent of the parameters of the space and time discretizations and its values may be different in different appearance.

Since we deal with a complex valued solution, one has to seek for an approximation in discrete spaces over the field of complex numbers $\mathbb{C}$. Some slight modifications should be made on the discrete spaces used in [2]. In particular, we define the space $X_{\mathcal{D}}$ as the set of all $\left((v_K)_{K\in\mathcal{M}}\,,\ (v_\sigma)_{\sigma\in\mathcal{E}}\right)$, where $v_K, v_\sigma \in \mathbb{C}$ for all $K \in \mathcal{M}$ and for all $\sigma \in \mathcal{E}$, and $X_{\mathcal{D},0} \subset X_{\mathcal{D}}$ is the set of all $v \in X_{\mathcal{D}}$ such that $v_\sigma = 0$ for all $\sigma \in \mathcal{E}_{\rm ext}$. Let $H_{\mathcal{M}}(\Omega, \mathbb{C})$ be the space of functions from Ω to $\mathbb{C}$ which are constant over each control volume of the mesh. For all $v \in X_{\mathcal{D}}$, we denote by $\Pi_{\mathcal{M}}\, v \in H_{\mathcal{M}}(\Omega, \mathbb{C})$ the function defined by $\Pi_{\mathcal{M}}\, v(x) = v_K$, for a.e. $x \in K$, for all $K \in \mathcal{M}$.

For all $\varphi \in \mathcal{C}(\Omega, \mathbb{C})$, we define $\mathcal{P}_{\mathcal{D}}\varphi = \left((\varphi(x_K))_{K\in\mathcal{M}}\,,\ (\varphi(x_\sigma))_{\sigma\in\mathcal{E}}\right) \in X_{\mathcal{D}}$. We denote by $\mathcal{P}_{\mathcal{M}}\varphi \in H_{\mathcal{M}}(\Omega, \mathbb{C})$ the function defined by $\mathcal{P}_{\mathcal{M}}\,\varphi(x) = \varphi(x_K)$, for a.e. $x \in K$, for all $K \in \mathcal{M}$. We will use the norm $\|\cdot\|_{1,2,\mathcal{M}}$ given by [2, (4.5), p. 1026].

In order to analyze the convergence, we need to consider the size of the discretization $\mathcal{D}$ defined by $h_{\mathcal{D}} = \sup\{\mathrm{diam(K)},\ K \in \mathcal{M}\}$ and the regularity of the mesh given by $\theta_{\mathcal{D}} = \max\left(\max_{\sigma\in\mathcal{E}_{\rm int},K,L\in\mathcal{M}} \dfrac{d_{K,\sigma}}{d_{L,\sigma}}, \max_{K\in\mathcal{M},\sigma\in\mathcal{E}_K} \dfrac{h_K}{d_{K,\sigma}}\right)$.

The scheme we want to consider in this note is based on the use of the discrete gradient given in [2]. For $u \in X_{\mathcal{D}}$, we define, for all $K \in \mathcal{M}$

$$\nabla_{\mathcal{D}}\, u(x) = \nabla_{K,\sigma}\, u, \quad \text{a. e. } x \in \mathcal{D}_{K,\sigma}, \tag{4}$$

where $\mathcal{D}_{K,\sigma}$ is the cone with vertex x_K and basis σ and

$$\nabla_{K,\sigma} u = \nabla_K\, u + \left(\frac{\sqrt{d}}{d_{K,\sigma}}\,(u_\sigma - u_K - \nabla_K\, u \cdot (x_\sigma - x_K))\right)\mathbf{n}_{K,\sigma}, \tag{5}$$

where $\nabla_K\, u = \dfrac{1}{\mathrm{m}(K)} \displaystyle\sum_{\sigma\in\mathcal{E}_K} \mathrm{m}(\sigma)\,(u_\sigma - u_K)\,\mathbf{n}_{K,\sigma}$ and d is the space dimension.

We define the finite volume approximation for (1)–(3) as $\left(u^n_{\mathcal{D}}\right)_{n=0}^{N+1} \in X^{N+2}_{\mathcal{D},0}$ with $u^n_{\mathcal{D}} = \left(\left(u^n_K\right)_{K\in\mathcal{M}}, \left(u^n_\sigma\right)_{\sigma\in\mathcal{E}}\right)$, for all $n \in \{0, \ldots, N+1\}$ and

1. discretization of the initial conditions (2): for all $v \in X_{\mathcal{D},0}$

$$\langle u^0_{\mathcal{D}}, v\rangle_F + \left(V(0)\Pi_{\mathcal{M}}\, u^0_{\mathcal{D}}, \Pi_{\mathcal{M}}\, v\right)_{\mathbb{L}^2(\Omega)} = \left(-\Delta\, u^0 + V(0)u^0, \Pi_{\mathcal{M}}\, v\right)_{\mathbb{L}^2(\Omega)}, \tag{6}$$

2. discretization of Eq. (1): for any $n \in \{1, \ldots, N\}$, for all $v \in X_{\mathcal{D},0}$

$$\begin{aligned} &i\left(\partial^1\,\Pi_{\mathcal{M}}\, u^{n+1}_{\mathcal{D}}, \Pi_{\mathcal{M}}\, v\right)_{\mathbb{L}^2(\Omega)} - \langle u^{n+1}_{\mathcal{D}}, v\rangle_F - \left(V(t_{n+1})\Pi_{\mathcal{M}}\, u^{n+1}_{\mathcal{D}}, \Pi_{\mathcal{M}}\, v\right)_{\mathbb{L}^2(\Omega)} \\ &\quad = \left(\frac{1}{k}\int_{t_n}^{t_{n+1}} f(t)dt, \Pi_{\mathcal{M}}\, v\right)_{\mathbb{L}^2(\Omega)}, \end{aligned} \tag{7}$$

where $\langle u, v\rangle_F = \displaystyle\int_\Omega \nabla_{\mathscr{D}} u(x) \cdot \nabla_{\mathscr{D}} \bar{v}(x)dx$, $\partial^1 v^n = \dfrac{v^n - v^{n-1}}{k}$, and $(\cdot,\cdot)_{\mathbb{L}^2(\Omega)}$ denotes the $\mathbb{L}^2$-inner product of the space $\mathbb{L}^2(\Omega, \mathbb{C})$.

The main result of the present contribution is the following theorem.

Theorem 1 (Error estimates for the finite volume scheme (6)– (7)) *Let Ω be a polyhedral open bounded subset of $\mathbb{R}^d$, where $d \in \mathbb{N} \setminus \{0\}$, and $\partial\Omega = \overline{\Omega} \setminus \Omega$ its boundary. Assume that the solution of the Schrödinger evolution problem of (1)–(3) satisfies $u \in \mathscr{C}^2([0,T];\ \mathscr{C}^2(\overline{\Omega}, \mathbb{C}))$ and the time dependent potential V is satisfying $V \in \mathscr{C}([0,T];\ \mathbb{L}^\infty(\Omega, \mathbb{R}))$ and $V(t)(x) \geq 0$ for all $t \in [0,T]$ and for a.e. $x \in \Omega$. Let $k = \frac{T}{N+1}$, with $N \in \mathbb{N}^\star$, and denote by $t_n = nk$, for $n \in \{0, \ldots, N+1\}$. Let $\mathscr{D} = (\mathscr{M}, \mathscr{E}, \mathscr{P})$ be a discretization in the sense of [2, Definition 2.1]. Assume that $\theta_{\mathscr{D}}$ satisfies $\theta \geq \theta_{\mathscr{D}}$.*

Then there exists a unique solution $\left(u^n_{\mathscr{D}}\right)_{n=0}^{N+1} \in \mathscr{X}^{N+2}_{\mathscr{D},0}$ for problem (6)–(7). Assume in addition that $V \in \mathscr{C}^j([0,T];\ \mathbb{L}^\infty(\Omega, \mathbb{R}))$ for all $j \in \{1,2\}$. Then, the following error estimates hold:

- *Discrete $\mathbb{L}^\infty(0,T; H^1_0(\Omega))$–estimate: for all $n \in \{0, \ldots, N+1\}$*

$$\| \mathscr{P}_{\mathscr{M}} u(t_n) - \Pi_{\mathscr{M}} u^n_{\mathscr{D}} \|_{1,2,\mathscr{M}} \leq C(k + h_{\mathscr{D}}) \| u \|_{\mathscr{C}^2([0,T];\ \mathscr{C}^2(\overline{\Omega}))}. \tag{8}$$

- *Discrete $\mathscr{W}^{1,\infty}(0,T; \mathbb{L}^2(\Omega))$–estimate: for all $n \in \{1, \ldots, N+1\}$*

$$\| \partial^1 (\mathscr{P}_{\mathscr{M}} u(t_n) - \Pi_{\mathscr{M}} u^n_{\mathscr{D}}) \|_{\mathbb{L}^2(\Omega)} \leq C(k + h_{\mathscr{D}}) \| u \|_{\mathscr{C}^2([0,T];\ \mathscr{C}^2(\overline{\Omega}))}. \tag{9}$$

- *Error estimate in the gradient approximation: for all $n \in \{0, \ldots, N+1\}$*

$$\| \nabla_{\mathscr{D}} u^n_{\mathscr{D}} - \nabla u(t_n) \|_{\mathbb{L}^2(\Omega)} \leq C(k + h_{\mathscr{D}}) \| u \|_{\mathscr{C}^2([0,T];\ \mathscr{C}^2(\overline{\Omega}))}. \tag{10}$$

The following lemma will help us to prove Theorem 1:

Lemma 1 (A new a priori estimate) *We consider the same discretizations as in Theorem 1. Assume that $\theta_{\mathscr{D}}$ satisfies $\theta \geq \theta_{\mathscr{D}}$ and that there exists $\left(\eta^n_{\mathscr{D}}\right)_{n=0}^{N+1} \in \mathscr{X}^{N+2}_{\mathscr{D},0}$ such that $\eta^0_{\mathscr{D}} = 0$ and for any $n \in \{0, \ldots, N\}$, for all $v \in \mathscr{X}_{\mathscr{D},0}$*

$$\begin{aligned} & i\left(\Pi_{\mathscr{M}} \partial^1 \eta^{n+1}_{\mathscr{D}}, \Pi_{\mathscr{M}} v\right)_{\mathbb{L}^2(\Omega)} - \langle \eta^{n+1}_{\mathscr{D}}, v\rangle_F - \left(V(t_{n+1}) \Pi_{\mathscr{M}} \eta^{n+1}_{\mathscr{D}}, \Pi_{\mathscr{M}} v\right)_{\mathbb{L}^2(\Omega)} \\ & = \left(\mathscr{S}^n, \Pi_{\mathscr{M}} v\right)_{\mathbb{L}^2(\Omega)}, \end{aligned} \tag{11}$$

where $\mathscr{S}^n \in \mathbb{L}^2(\Omega, \mathbb{C})$, for all $n \in \{0, \ldots, N\}$ and $V \in \mathscr{C}([0,T];\ \mathbb{L}^\infty(\Omega, \mathbb{R}))$ satisfying $V(t)(x) \geq 0$ for all $t \in [0,T]$ and for a.e. $x \in \Omega$. Assume in addition that $V \in \mathscr{C}^j([0,T];\ \mathbb{L}^\infty(\Omega, \mathbb{R}))$ for all $j \in \{1,2\}$. Then the following estimate holds, for all $j \in \{0, \ldots, N\}$

$$\| \Pi_{\mathscr{M}} \partial^1 \eta_{\mathscr{D}}^{j+1} \|_{\mathbb{L}^2(\Omega)} + \| \Pi_{\mathscr{M}} \eta_{\mathscr{D}}^{j+1} \|_{1,2,\mathscr{M}} + \| \nabla_{\mathscr{D}} \eta_{\mathscr{D}}^{j+1} \|_{(\mathbb{L}^2(\Omega))^d} \leq C(\mathscr{S} + \mathscr{S}_1), \tag{12}$$

where $\mathscr{S} = \max_{n=0}^{N} \| \mathscr{S}^n \|_{\mathbb{L}^2(\Omega)}$ *and* $\mathscr{S}_1 = \max_{n=1}^{N} \| \partial^1 \mathscr{S}^n \|_{\mathbb{L}^2(\Omega)}$.

Proof **1. Estimate on** $\| \Pi_{\mathscr{M}} \partial^1 \eta_{\mathscr{D}}^{j+1} \|_{\mathbb{L}^2(\Omega)}$. Acting the discrete operator ∂^1 on both sides of (11) and using the formula $\partial^1(r^n s^n) = s^n \partial^1 r^n + r^{n-1} \partial^1 s^n$ yields

$$\begin{aligned} & i \left(\Pi_{\mathscr{M}} \partial^2 \eta_{\mathscr{D}}^{n+1}, \Pi_{\mathscr{M}} v \right)_{\mathbb{L}^2(\Omega)} - \langle \partial^1 \eta_{\mathscr{D}}^{n+1}, v \rangle_F - \left(V(t_{n+1}) \partial^1 \Pi_{\mathscr{M}} \eta_{\mathscr{D}}^{n+1}, \Pi_{\mathscr{M}} v \right)_{\mathbb{L}^2(\Omega)} \\ & = \left(\partial^1 \mathscr{S}^n, \Pi_{\mathscr{M}} v \right)_{\mathbb{L}^2(\Omega)} + \left(\partial^1 V(t_{n+1}) \Pi_{\mathscr{M}} \eta_{\mathscr{D}}^{n}, \Pi_{\mathscr{M}} v \right)_{\mathbb{L}^2(\Omega)}. \end{aligned} \tag{13}$$

Choosing $v = \partial^1 \eta_{\mathscr{D}}^{n+1}$ in (13) and taking the imaginary part of the result, we get

$$\begin{aligned} & \mathrm{Re} \left(\Pi_{\mathscr{M}} \partial^2 \eta_{\mathscr{D}}^{n+1}, \Pi_{\mathscr{M}} v \right)_{\mathbb{L}^2(\Omega)} \\ & = \mathrm{Im} \left(\left(\partial^1 \mathscr{S}^n, \Pi_{\mathscr{M}} v \right)_{\mathbb{L}^2(\Omega)} + \left(\partial^1 V(t_{n+1}) \Pi_{\mathscr{M}} \eta_{\mathscr{D}}^{n}, \Pi_{\mathscr{M}} v \right)_{\mathbb{L}^2(\Omega)} \right). \end{aligned} \tag{14}$$

Some calculations lead to the expression, for all function $(\omega^n)_{n=0}^{N+1} \in (\mathbb{L}^2(\Omega, \mathbb{C}))^{N+2}$:

$$\begin{aligned} 2k \left(\partial^1 \omega^{n+1}, \omega^{n+1} \right)_{\mathbb{L}^2(\Omega)} = & \| \omega^{n+1} - \omega^n \|^2_{\mathbb{L}^2(\Omega)} \\ & + \| \omega^{n+1} \|^2_{\mathbb{L}^2(\Omega)} - \| \omega^n \|^2_{\mathbb{L}^2(\Omega)} + 2i \, \mathrm{Im} \left(\omega^{n+1}, \omega^n \right)_{\mathbb{L}^2(\Omega)} \end{aligned} \tag{15}$$

Gathering (14) with (15) when $\omega^{n+1} = \Pi_{\mathscr{M}} \partial^1 \eta_{\mathscr{D}}^{n+1}$, using the fact that $V \in \mathscr{C}^1([0, T]; \mathbb{L}^\infty(\Omega, \mathbb{R}))$, and the Cauchy Schwarz inequality, we get

$$\begin{aligned} & \| \Pi_{\mathscr{M}} \partial^1 \eta_{\mathscr{D}}^{n+1} \|^2_{\mathbb{L}^2(\Omega)} - \| \Pi_{\mathscr{M}} \partial^1 \eta_{\mathscr{D}}^{n} \|^2_{\mathbb{L}^2(\Omega)} \\ & \leq 2kC(\mathscr{S}_1 + \| \Pi_{\mathscr{M}} \eta_{\mathscr{D}}^{n} \|_{\mathbb{L}^2(\Omega)}) \| \Pi_{\mathscr{M}} \partial^1 \eta_{\mathscr{D}}^{n+1} \|_{\mathbb{L}^2(\Omega)}. \end{aligned} \tag{16}$$

Let us prove an $\mathbb{L}^\infty(0, T; \mathbb{L}^2(\Omega, \mathbb{C}))$-estimate. Taking $v = \eta_{\mathscr{D}}^{n+1}$ in (11) and using the fact that V is a real valued function, and taking the imaginary part to get

$$\mathrm{Re} \left(\Pi_{\mathscr{M}} \partial^1 \eta_{\mathscr{D}}^{n+1}, \Pi_{\mathscr{M}} \eta_{\mathscr{D}}^{n+1} \right)_{\mathbb{L}^2(\Omega)} = \mathrm{Im} \left(\mathscr{S}^n, \Pi_{\mathscr{M}} v \right)_{\mathbb{L}^2(\Omega)}. \tag{17}$$

This with (15) when $\omega^{n+1} = \Pi_{\mathscr{M}} \eta_{\mathscr{D}}^{n+1}$, and the Cauchy Schwarz inequality yields

$$\| \Pi_{\mathscr{M}} \eta_{\mathscr{D}}^{n+1} \|^2_{\mathbb{L}^2(\Omega)} - \| \Pi_{\mathscr{M}} \eta_{\mathscr{D}}^{n} \|^2_{\mathbb{L}^2(\Omega)} \leq 2k\mathscr{S} \| \Pi_{\mathscr{M}} \eta_{\mathscr{D}}^{n+1} \|_{\mathbb{L}^2(\Omega)}. \tag{18}$$

Summing (18) over $n \in \{0, \ldots, j\}$, where $j \in \{0, \ldots, N\}$, and using the fact $\eta^0_{\mathscr{D}} = 0$ yields $\| \Pi_{\mathscr{M}} \eta^{j+1}_{\mathscr{D}} \|^2_{\mathbb{L}^2(\Omega)} \leq 2k\mathscr{S} \sum_{n=0}^{j} \| \Pi_{\mathscr{M}} \eta^{n+1}_{\mathscr{D}} \|_{\mathbb{L}^2(\Omega)}$. Applying a Young's inequality (as applied in (20) below) and using the discrete Gronwall's lemma yields

$$\| \Pi_{\mathscr{M}} \eta^{j+1}_{\mathscr{D}} \|_{\mathbb{L}^2(\Omega)} \leq C\mathscr{S}. \tag{19}$$

Inserting this estimate in (16) and summing the result over $n \in \{1, \ldots, j\}$, where $j \in \{1, \ldots, N\}$ yields

$$\begin{aligned} &\| \Pi_{\mathscr{M}} \partial^1 \eta^{j+1}_{\mathscr{D}} \|^2_{\mathbb{L}^2(\Omega)} - \| \Pi_{\mathscr{M}} \partial^1 \eta^1_{\mathscr{D}} \|^2_{\mathbb{L}^2(\Omega)} \\ &\qquad \leq 2kC(\mathscr{S} + \mathscr{S}_1) \sum_{n=1}^{j} \| \Pi_{\mathscr{M}} \partial^1 \eta^{n+1}_{\mathscr{D}} \|_{\mathbb{L}^2(\Omega)}. \end{aligned}$$

This with a Young's inequality leads to

$$\begin{aligned} \| \Pi_{\mathscr{M}} \partial^1 \eta^{j+1}_{\mathscr{D}} \|^2_{\mathbb{L}^2(\Omega)} \leq \frac{2k}{T} \sum_{n=2}^{j} \| \Pi_{\mathscr{M}} \partial^1 \eta^{n}_{\mathscr{D}} \|^2_{\mathbb{L}^2(\Omega)} + 2\| \Pi_{\mathscr{M}} \partial^1 \eta^1_{\mathscr{D}} \|^2_{\mathbb{L}^2(\Omega)} \\ + 8T^2 \, (C)^2 \, (\mathscr{S} + \mathscr{S}_1)^2 \, . \end{aligned} \tag{20}$$

We now estimate $\| \Pi_{\mathscr{M}} \partial^1 \eta^1_{\mathscr{D}} \|^2_{\mathbb{L}^2(\Omega,\mathbb{C})}$. To this end, we set $n = 0$ and $v = \partial^1 \eta^1_{\mathscr{D}}$ in (11) and we use the fact that $\partial^1 \eta^1_{\mathscr{D}} = \dfrac{\eta^1_{\mathscr{D}}}{k}$ (this stems from $\eta^0_{\mathscr{D}} = 0$)

$$\begin{aligned} &i \| \Pi_{\mathscr{M}} \partial^1 \eta^1_{\mathscr{D}} \|^2_{\mathbb{L}^2(\Omega)} - \frac{1}{k} \langle \eta^1_{\mathscr{D}}, \eta^1_{\mathscr{D}} \rangle_F - \frac{1}{k} \left(V(t_1) \Pi_{\mathscr{M}} \eta^1_{\mathscr{D}}, \Pi_{\mathscr{M}} \eta^1_{\mathscr{D}} \right)_{\mathbb{L}^2(\Omega)} \\ &\quad = \left(\mathscr{S}^0, \Pi_{\mathscr{M}} \partial^1 \eta^1_{\mathscr{D}} \right)_{\mathbb{L}^2(\Omega)}. \end{aligned} \tag{21}$$

Taking the imaginary part in (21) and using the Cauchy Schwarz inequality implies $\| \Pi_{\mathscr{M}} \partial^1 \eta^1_{\mathscr{D}} \|_{\mathbb{L}^2(\Omega)} \leq \mathscr{S}$. This with inequality (20) and the discrete version of the Gronwall's lemma yields the desired estimate $\mathscr{W}^{1,\infty}(0, T; \mathbb{L}^2)$-estimate in (12).

2. Estimate on $\| \Pi_{\mathscr{M}} \eta^{j+1}_{\mathscr{D}} \|_{1,2,\mathscr{M}}$. Choosing $v = \partial^1 \eta^{n+1}_{\mathscr{D}}$ in (11) and taking the real part yields

$$\begin{aligned} &\mathrm{Re} \left(\langle \eta^{n+1}_{\mathscr{D}}, \partial^1 \eta^{n+1}_{\mathscr{D}} \rangle_F + \left(V(t_{n+1}) \Pi_{\mathscr{M}} \eta^{n+1}_{\mathscr{D}}, \partial^1 \Pi_{\mathscr{M}} \eta^{n+1}_{\mathscr{D}} \right)_{\mathbb{L}^2(\Omega)} \right) \\ &\quad = \mathrm{Re} \left(-\mathscr{S}^n, \Pi_{\mathscr{M}} v \right)_{\mathbb{L}^2(\Omega)}. \end{aligned} \tag{22}$$

Writing $\langle \eta_{\mathscr{D}}^{n+1}, \partial^1 \eta_{\mathscr{D}}^{n+1} \rangle_F$ and $\left(V(t_{n+1}) \Pi_{\mathscr{M}} \eta_{\mathscr{D}}^{n+1}, \partial^1 \Pi_{\mathscr{M}} \eta_{\mathscr{D}}^{n+1} \right)_{\mathbb{L}^2(\Omega)}$ in a similar manner to that of (15) and gathering this with (22) leads to

$$\begin{aligned} \langle \eta_{\mathscr{D}}^{n+1}, \eta_{\mathscr{D}}^{n+1} \rangle_F - \langle \eta_{\mathscr{D}}^{n}, \eta_{\mathscr{D}}^{n} \rangle_F + \left(V(t_{n+1}) \Pi_{\mathscr{M}} \eta_{\mathscr{D}}^{n+1}, \Pi_{\mathscr{M}} \eta_{\mathscr{D}}^{n+1} \right)_{\mathbb{L}^2(\Omega)} \\ - \left(V(t_{n+1}) \Pi_{\mathscr{M}} \eta_{\mathscr{D}}^{n}, \Pi_{\mathscr{M}} \eta_{\mathscr{D}}^{n} \right)_{\mathbb{L}^2(\Omega)} \\ \leq 2k \, \mathrm{Re} \left(-\mathscr{S}^n, \Pi_{\mathscr{M}} v \right)_{\mathbb{L}^2(\Omega)}. \end{aligned} \tag{23}$$

Summing (23) over $n \in \{0, \ldots, j\}$, where $j \in \{0, \ldots, N\}$, using the Cauchy Schwarz inequality and [2, Lemma 4.2] yields $|\eta_{\mathscr{D}}^{j+1}|_{\mathscr{X}}^2 \leq Ck\mathscr{S} \sum_{n=0}^{j} \| \partial^1 \Pi_{\mathscr{M}} \eta_{\mathscr{D}}^{n+1} \|_{\mathbb{L}^2(\Omega)}$. This with the estimate on $\| \Pi_{\mathscr{M}} \partial^1 \eta_{\mathscr{D}}^{j+1} \|_{\mathbb{L}^2(\Omega)}$ (it is proved in in the previous item) yields

$$|\eta_{\mathscr{D}}^{j+1}|_{\mathscr{X}} \leq C(\mathscr{S} + \mathscr{S}_1). \tag{24}$$

This with the inequality norm [2, (4.6), p. 1026] implies the desired estimate $\mathbb{L}^\infty(0, T; \ H^1(\Omega))$-estimate in (12).

3. Estimate $\| \nabla_{\mathscr{D}} \eta_{\mathscr{D}}^{j+1} \|_{(\mathbb{L}^2(\Omega))^d}$. Estimate (24) with [2, Lemma 4.2] implies the estimate concerning $\| \nabla_{\mathscr{D}} \eta_{\mathscr{D}}^{j+1} \|_{(\mathbb{L}^2(\Omega))^d}$ in (12). ■

Sketch of the proof of Theorem 1: The uniqueness of $\left(u_{\mathscr{D}}^n \right)_{n \in \{0, \ldots, N+1\}}$ satisfying (6)–(7) can be deduced using the [2, Lemma 4.2]. As usual, we use this uniqueness to prove the existence. To prove the error estimates (8)–(10), we compare the solution $u_{\mathscr{D}}^n$ with the solution: for any $n \in \{0, \ldots, N+1\}$, find $\hat{u}_{\mathscr{D}}^n \in \mathscr{X}_{\mathscr{D},0}$ such that

$$\begin{aligned} \langle \hat{u}_{\mathscr{D}}^n, v \rangle_F + \left(V(t_n) \Pi_{\mathscr{M}} \hat{u}_{\mathscr{D}}^n, \Pi_{\mathscr{M}} v \right)_{\mathbb{L}^2(\Omega)} \\ = \left(-\Delta u(t_n) + V(t_n) u(t_n), \Pi_{\mathscr{M}} v \right)_{\mathbb{L}^2(\Omega)}, \quad \forall v \in \mathscr{X}_{\mathscr{D},0}. \end{aligned} \tag{25}$$

Step 1: Comparison between u and $\hat{u}_{\mathscr{D}}^n$. Using techniques of the proof of [2, Theorem 4.8] yields, for all $v \in \mathscr{X}_{\mathscr{D},0}$

$$\begin{aligned} \left\langle \mathscr{P}_{\mathscr{D}} u(t_n) - \hat{u}_{\mathscr{D}}^n, v \right\rangle_F + \left(V(t_n) (\mathscr{P}_{\mathscr{M}} u(t_n) - \Pi_{\mathscr{M}} \hat{u}_{\mathscr{D}}^n), \Pi_{\mathscr{M}} v \right)_{\mathbb{L}^2(\Omega)} \\ = \sum_{K \in \mathscr{M}} \sum_{\sigma \in \mathscr{E}_K} \mathscr{R}_{K,\sigma} (u(t_n)) (\bar{v}_K - \bar{v}_\sigma) + \left(V(t_n) r(u(t_n)), \Pi_{\mathscr{M}} v \right)_{\mathbb{L}^2(\Omega)}, \end{aligned} \tag{26}$$

where the expression $\mathbb{E}_{\mathscr{D}}(u(t_n)) = \left(\sum_{K \in \mathscr{M}} \sum_{\sigma \in \mathscr{E}_K} \frac{d_{K,\sigma}}{\mathrm{m}(\sigma)} \left| \mathscr{R}_{K,\sigma} (u(t_n)) \right|^2 \right)^{\frac{1}{2}}$ is satisfying the estimate $\mathbb{E}_{\mathscr{D}}(u(t_n)) \leq Ch_{\mathscr{D}} \| u \|_{\mathscr{C}([0,T]; \ \mathscr{C}^2(\overline{\Omega}))}$ and $r(u) = \mathscr{P}_{\mathscr{M}} u - u$. Taking $v = \mathscr{P}_{\mathscr{D}} u(t_n) - \hat{u}_{\mathscr{D}}^n$ in (26) yields

$$\langle v, v\rangle_F + \left(V(t_n)\Pi_{\mathscr{M}} v, \Pi_{\mathscr{M}} v\right)_{\mathbb{L}^2(\Omega)} = \sum_{K\in\mathscr{M}}\sum_{\sigma\in\mathscr{E}_K} \mathscr{R}_{K,\sigma}\left(u(t_n)\right)\left(\bar{v}_K - \bar{v}_\sigma\right) + \left(V(t_n) r\left(u(t_n)\right), \Pi_{\mathscr{M}} v\right)_{\mathbb{L}^2(\Omega)}. \quad (27)$$

This with [2, Lemma 4.2], the Cauchy Schwarz inequality, the Sobolev inequality of [2, Lemma 5.4], and the inequality norm [2, (4.6), p. 1026] yields that

$$|\mathscr{P}_{\mathscr{D}} u(t_n) - \hat{u}^n_{\mathscr{D}}|_{\mathscr{X}} \leq Ch_{\mathscr{D}} \|u\|_{\mathscr{C}([0,T];\, \mathscr{C}^2(\overline{\Omega}))}. \quad (28)$$

This with [2, (4.6), p. 1026], [2, Lemma 4.2], and [2, Lemma 4.4] implies the error estimate:

$$\begin{aligned} &\|\mathscr{P}_{\mathscr{M}} u(t_n) - \Pi_{\mathscr{M}}\hat{u}^n_{\mathscr{D}}\|_{1,2,\mathscr{M}} + \|\nabla u(t_n) - \nabla_{\mathscr{D}}\hat{u}^n\|_{\mathbb{L}^2(\Omega)} \\ &\quad \leq Ch_{\mathscr{D}} \|u\|_{\mathscr{C}([0,T];\, \mathscr{C}^2(\overline{\Omega}))}. \end{aligned} \quad (29)$$

We will now derive an $\mathscr{W}^{1,\infty}(0, T; \mathbb{L}^2)$–estimate. Acting the discrete operator ∂^1 on Eq. (26) to get, for any $n \in \{1, \ldots, N+1\}$

$$\begin{aligned} &\left\langle \partial^1\left(\mathscr{P}_{\mathscr{D}} u(t_n) - \hat{u}^n_{\mathscr{D}}\right), v\right\rangle_F + \left(V(t_n)\partial^1\left(\left(\mathscr{P}_{\mathscr{M}} u(t_n) - \Pi_{\mathscr{M}}\hat{u}^n_{\mathscr{D}}\right)\right), \Pi_{\mathscr{M}} v\right)_{\mathbb{L}^2(\Omega)} \\ &= \sum_{K\in\mathscr{M}}\sum_{\sigma\in\mathscr{E}_K} \mathscr{R}_{K,\sigma}\left(\partial^1 u(t_n)\right)\left(\bar{v}_K - \bar{v}_\sigma\right) + \left(\mathbb{T}_1 + \mathbb{T}_2 - \mathbb{T}_3, \Pi_{\mathscr{M}} v\right)_{\mathbb{L}^2(\Omega)}, \end{aligned} \quad (30)$$

where $\mathbb{T}_1 = \partial^1\left(V(t_n)\right)\left(\mathscr{P}_{\mathscr{M}} u(t_n) - u(t_n)\right)$, $\mathbb{T}_2 = V(t_{n-1})\partial^1\left(\left(\mathscr{P}_{\mathscr{M}} u(t_n) - u(t_n)\right)\right)$, and $\mathbb{T}_3 = \partial^1\left(V(t_n)\right)\left(\mathscr{P}_{\mathscr{M}} u(t_{n-1}) - \Pi_{\mathscr{M}}\hat{u}^{n-1}_{\mathscr{D}}\right)$. Thanks to Taylor expansions and $\mathbb{L}^\infty(0, T; H^1_0(\Omega))$-estimate in (29) with [2, Lemma 5.4], we have

$$\|\mathbb{T}_i\|_{\mathbb{L}^2(\Omega)} \leq Ch_{\mathscr{D}} \|u\|_{\mathscr{C}^1([0,T];\, \mathscr{C}^2(\overline{\Omega}))}, \quad \forall i \in \{1, 2, 3\}. \quad (31)$$

Taking $v = \partial^1\left(\mathscr{P}_{\mathscr{D}} u(t_n) - \hat{u}^n_{\mathscr{D}}\right)$ in (30), using [2, Lemma 4.2], and gathering this with the Cauchy Schwarz inequality, [2, Lemma 5.4], [2, (4.6), p. 1026], and (31) to get

$$\|\partial^1\left(\mathscr{P}_{\mathscr{M}} u(t_n) - \Pi_{\mathscr{M}}\hat{u}^n_{\mathscr{D}}\right)\|_{\mathbb{L}^2(\Omega)} \leq Ch_{\mathscr{D}} \|u\|_{\mathscr{C}^1([0,T];\, \mathscr{C}^2(\overline{\Omega}))}. \quad (32)$$

Using the same techniques followed in (30)–(32), we are able to prove

$$\|\partial^2\left(\mathscr{P}_{\mathscr{M}} u(t_n) - \Pi_{\mathscr{M}}\hat{u}^n_{\mathscr{D}}\right)\|_{\mathbb{L}^2(\Omega)} \leq Ch_{\mathscr{D}} \|u\|_{\mathscr{C}^2([0,T];\, \mathscr{C}^2(\overline{\Omega}))}. \quad (33)$$

Step 2: Comparison between $\hat{u}^n_{\mathscr{D}}$ and $u^n_{\mathscr{D}}$. Writing (25) in the step $n+1$, summing the result with (7) and using (1) yields, for all $n \in \{0, \ldots, N\}$

$$i\left(\partial^1 \Pi_{\mathscr{M}}\, \eta_{\mathscr{D}}^{n+1}, \Pi_{\mathscr{M}}\, v\right)_{\mathbb{L}^2(\Omega)} - \langle\, \eta_{\mathscr{D}}^{n+1}, v\rangle_F - \left(V(t_{n+1})\Pi_{\mathscr{M}}\eta_{\mathscr{D}}^{n+1}, \Pi_{\mathscr{M}}\, v\right)_{\mathbb{L}^2(\Omega)}$$
$$= \left(\mathscr{S}^n, \Pi_{\mathscr{M}}\, v\right)_{\mathbb{L}^2(\Omega)}, \tag{34}$$

where $\eta_{\mathscr{D}}^n = u_{\mathscr{D}}^n - \hat{u}_{\mathscr{D}}^n$ and $\mathscr{S}^n$ is given by

$$\mathscr{S}^n = i\,\partial^1(u(t_{n+1}) - \Pi_{\mathscr{M}}\,\hat{u}_{\mathscr{D}}^{n+1}) + \frac{1}{k}\int_{t_n}^{t_{n+1}} \Delta\, u(t)dt - \Delta\, u(t_{n+1})$$
$$- \frac{1}{k}\int_{t_n}^{t_{n+1}} V(t)u(t)dt + V(t_{n+1})u(t_{n+1}). \tag{35}$$

Thanks to suitable Taylor expansions and (32)–(33), we are able to justify that $\mathscr{S} + \mathscr{S}_1 \leq C(k + h_{\mathscr{D}})\| u\|_{\mathscr{C}^2([0,T];\, \mathscr{C}^2(\overline{\Omega}))}$, where $\mathscr{S}$ and $\mathscr{S}_1$ are defined in Lemma 1. In addition to this, $\eta_{\mathscr{D}}^0 = 0$ (it stems from (2)). One remarks that $\left(\eta_{\mathscr{D}}^n\right)_{n=0}^{N+1}$ is satisfying hypothesis of Lemma 1, one can apply estimate (12) of Lemma 1 to obtain

$$\|\Pi_{\mathscr{M}}\,\partial^1\,\eta_{\mathscr{D}}^{j+1}\|_{\mathbb{L}^2(\Omega)} + \|\,\Pi_{\mathscr{M}}\eta_{\mathscr{D}}^{j+1}\|_{1,2,\mathscr{M}} + \|\,\nabla_{\mathscr{D}}\,\eta_{\mathscr{D}}^{j+1}\|_{(\mathbb{L}^2(\Omega))^d}$$
$$\leq C(k + h_{\mathscr{D}})\| u\|_{\mathscr{C}^2([0,T];\, \mathscr{C}^2(\overline{\Omega}))}. \tag{36}$$

This with estimates (29) and (32) implies estimates of Theorem 1. ■

3 Conclusion and a Perspective

We considered the linear Schrödinger evolution equation. A convergence analysis of a new finite volume scheme is provided. We plan to consider the case when the spatial spatial domain is not bounded and to use the absorbing boundary conditions.

References

1. Akrivis, G. D., Dougalis, V. A.: On a class of conservative, highly accurate Galerkin methods for the Schrödinger equation. RAIRO Modél. Math. Anal. Numér. **25** /6, 643–670 (1991)
2. Eymard, R., Gallouët, T., Herbin, R.: Discretization of heterogeneous and anisotropic diffusion problems on general nonconforming meshes SUSHI: a scheme using stabilization and hybrid interfaces. IMA J. Numer. Anal. **30**(4), 1009–1043 (2010)
3. Koprucki, T., Eymard, R., Fuhrmann, J.: Convergence of a finite volume scheme to the eigenvalues of a Schrödinger operator. WIAS Preprint NO. 1260 (2007)

A Note on a New Second Order Approximation Based on a Low–Order Finite Volume Scheme for the Wave Equation in One Space Dimension

Abdallah Bradji

Abstract This note is an extension of our previous work [1] which dealt with a first order finite volume scheme for the wave equation. We construct a new second order approximation for the solution of the wave equation in one space dimension. This new high-order approximation can be computed using the same simple scheme used in [1] and its formulation includes an approximation which together with their first and second time derivatives converge towards their corresponding time derivatives of the second spatial derivative of the exact solution. The analysis provided in this note is based on the use of a new a priori estimate.

MSC2010: 65M08, 65M12, 65M15, 35L10

1 Preliminaries and Aim of This Paper

We consider the following one dimensional wave problem:

$$u_{tt}(x,t) - u_{xx}(x,t) = f(x,t), \ \ (x,t) \in \mathbf{I} \times (0,T), \tag{1}$$

where $\mathbf{I} = (0,1)$, $T > 0$, and f is a given function.
Initial conditions are defined by, for given functions u^0 and u^1:

$$u(x,0) = u^0(x) \ \text{ and } \ u_t(x,0) = u^1(x), \ \ x \in \mathbf{I}, \tag{2}$$

Homogeneous Dirichlet boundary conditions are given by

$$u(0,t) = u(1,t) = 0, \ \ t \in (0,T). \tag{3}$$

A. Bradji (✉)
Department of Mathematics, Faculty of Sciences,
University of Badji Mokhtar-Annaba, Annaba, Algeria
e-mail: bradji@cmi.univ-mrs.fr; abdallah-bradji@univ-annaba.org

J. Fuhrmann et al. (eds.), *Finite Volumes for Complex Applications VII - Methods and Theoretical Aspects*, Springer Proceedings in Mathematics & Statistics 77,
DOI: 10.1007/978-3-319-05684-5_12,

The time discretization is performed with a constant time step $k = \frac{T}{M+1}$, where $M \in \mathbb{N} \setminus \{0\}$, and we shall denote $t_n = nk$, for $n \in [\![0, M+1]\!]$. The spatial domain $\mathbf{I}$ is discretized using the admissible one–dimensional mesh of [2] which we recall here for the sake of completeness.

Definition 1 (*Admissible mesh*) An admissible mesh $\mathscr{T}$ of $\mathbf{I} = (0, 1)$ is given by a family $\{K_i;\ i \in [\![1, N]\!]\}$, $N \in \mathbb{N}^\star$ of control volumes, such that $K_i = (x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}})$ and a family $\{x_i;\ i \in [\![0, N+1]\!]\}$ such that $x_0 = x_{\frac{1}{2}} = 0 < x_1 < x_{\frac{3}{2}} < \cdots < x_{i-\frac{1}{2}} < x_i < x_{i+\frac{1}{2}} < \cdots < x_N < x_{N+\frac{1}{2}} = x_{N+1} = 1$ and, for $i \in [\![1, N]\!]$, $h_i = \mathrm{m}(K_i) = x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}}$, $h_i^- = x_i - x_{i-\frac{1}{2}}$ and $h_i^+ = x_{i+\frac{1}{2}} - x_i$.
We set $h_{i+\frac{1}{2}} = x_{i+1} - x_i$, for all $i \in [\![0, N]\!]$, and $h = \max_{i \in [\![1,N]\!]} h_i$.

Define $\mathscr{X}(\mathscr{T})$ as the set of functions from $\mathbf{I}$ to $\mathbb{R}$ which are constant on each control volume K_i, $i \in [\![1, N]\!]$, of the mesh. We shall sometime identify $\mathscr{X}(\mathscr{T})$ with $\mathbb{R}^N$. For each $u \in \mathscr{X}(\mathscr{T})$, we define the discrete H_0^1–norm by $\|u\|_{1,\mathscr{T}} = \left(\sum_{i=1}^{N-1} \frac{(u_{i+1}-u_i)^2}{h_{i+\frac{1}{2}}} + \frac{(u_1)^2}{h_{\frac{1}{2}}} + \frac{(u_N)^2}{h_{N+\frac{1}{2}}}\right)^{\frac{1}{2}}$, where u_i denotes the value taken by $u \in \mathscr{X}(\mathscr{T})$ on the control volume $K_i = (x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}})$. For all function $\varphi \in \mathscr{C}(\bar{\mathbf{I}})$, we denote by $\Pi_{\mathscr{T}}\varphi \in \mathscr{X}(\mathscr{T})$ the function defined by $\Pi_{\mathscr{T}}\varphi(x) = \varphi(x_i)$, for a.e. $x \in K_i$, for all $i \in [\![1, N]\!]$. Throughout this paper, the letter C stands for a positive constant independent of the parameters of the space and time discretizations. For $u = (u_i)_{i=1}^N \in \mathscr{X}(\mathscr{T})$, we define:

$$\mathbb{F}_{i+\frac{1}{2}}(u) = \frac{u_{i+1} - u_i}{h_{i+\frac{1}{2}}} \quad \text{and} \quad \mathbb{D}_{i+\frac{1}{2}}(u) = \mathbb{F}_{i+\frac{1}{2}}(u) - \mathbb{F}_{i-\frac{1}{2}}(u). \tag{4}$$

The following scheme is given in [1]: for all $n \in [\![1, M]\!]$, find $u^{n+1} = \left(u_i^n\right)_{i=1}^N \in \mathscr{X}(\mathscr{T})$ such that

$$h_i \partial^2 u_i^{n+1} - \mathbb{D}_{i+\frac{1}{2}}(u^n) = \frac{1}{k} \int_{nk}^{(n+1)k} \int_{K_i} f(x,t)\, dx dt, \quad \forall i \in [\![1, N]\!], \tag{5}$$

with $u_0^n = u_{N+1}^n = 0$, for all $n \in [\![0, M+1]\!]$ and for all $i \in [\![1, N]\!]$

$$-\mathbb{D}_{i+\frac{1}{2}}(u^0) = -\int_{K_i} (u^0)_{xx}(x) dx \quad \text{and} \quad -\mathbb{D}_{i+\frac{1}{2}}(\partial^1 u^1) = -\int_{K_i} (u^1)_{xx}(x) dx. \tag{6}$$

and the operator ∂^1 denotes the discrete temporal derivative $\partial^1 v^n = \frac{v^n - v^{n-1}}{k}$, and ∂^2 denotes the discrete second temporal derivative $\partial^2 v^n = \partial^1(\partial^1 v^n)$. It is proved in [1] that scheme (5)–(6) is of first order. This contribution deals with a new second order finite volume approximation for problem (1)–(3). This second order approximation can be computed using the same scheme (5)–(6).

2 Formulation of a New Second Order Approximation

We first provide approximations for the following derivatives of u:

1. *A convenient approximation* $(\alpha^n)_n$ for u_{xx}, we follow the following steps:
   1.1. First step: a convenient approximation for $\psi = u_{xxtt}$. By differentiating (1) twice with respect to x, we get $\psi = f_{xx} + u_{xxxx}$. An approximation for u_{xxxx} yields then an approximation for $\psi = u_{xxtt}$. The function $\varphi = u_{xxxx}$ is satisfying $\varphi_{tt} - \varphi_{xx} = f_{xxxx}$. In addition, using (2) yields

$$\varphi(x,0) = (u^0)_{xxxx}(x) \text{ and } \varphi_t(x,0) = (u^1)_{xxxx}(x), \ x \in \mathbf{I}. \tag{7}$$

   To get the boundary conditions of φ, we have (thanks to some suitable differentiation for (1)) $0 = f_{xx}(x,t) + f_{tt}(x,t) + \varphi(x,t) - u_{tttt}(x,t)$ which implies that, thanks to (3) to get $\varphi(0,t) = F_1(t)$ and $\varphi(1,t) = F_2(1,t)$, where $F_1(t) = -f_{xx}(0,t) - f_{tt}(0,t)$ and $F_2(t) = -f_{xx}(1,t) - f_{tt}(1,t)$.
   Since $\varphi = u_{xxxx}$ satisfies equation $\varphi_{tt} - \varphi_{xx} = f_{xxxx}$ which is similar to the Eq. (1) satisfied by u with different boundary and initial conditions, hence $\varphi = u_{xxxx}$ can be approximated using the same scheme (5)–(6): for all $n \in [\![1, M]\!]$, find $\varphi^{n+1} = \left(\varphi_i^{n+1}\right)_{i=1}^N \in \mathscr{X}(\mathscr{T})$ such that

$$h_i \partial^2 \varphi_i^{n+1} - \mathbb{D}_{i+\frac{1}{2}}\left(\varphi^{n+1}\right) = \frac{1}{k}\int_{nk}^{(n+1)k}\int_{K_i} f_{xxxx}(x,t)\,dxdt, \tag{8}$$

   with $\varphi_0^n = F_1(t_n)$ and $\varphi_{N+1}^n = F_2(t_n)$, for all $n \in [\![0, M+1]\!]$, and for all $i \in [\![1, N]\!]$

$$-\mathbb{D}_{i+\frac{1}{2}}\left(\varphi^0\right) = -\int_{K_i}\frac{d^6u^0}{dx^6}(x)dx \quad \text{and} \quad -\mathbb{D}_{i+\frac{1}{2}}\left(\partial^1\varphi^1\right) = -\int_{K_i}\frac{d^6u^1}{dx^6}(x)dx. \tag{9}$$

   We now define an approximation for $\psi = u_{xxtt}$ on the mesh points by $\psi_i^n = f_{xx}(x_i, t_n) + \varphi_i^n$.
   1.2. Second step: a convenient approximation for $v = u_{xx}$. One remarks that the unknown function $v = u_{xx}$ is a second integration in time of $\psi = u_{xxtt}$, one can attempt to look for an approximation for $v = u_{xx}$ using a *second numerical integration in time* for $\psi^n = (\psi_i^n)_{i\in[\![1,N]\!]}$ which is an approximation for $\psi = u_{xxtt}$. Let $(\alpha^n)_{n\in[\![0,M+1]\!]} \in (\mathscr{X}(\mathscr{T}))^{M+2}$ be defined as $\partial^2\alpha^n = \psi^n$, for all $n \in [\![2, M+1]\!]$. Some computations lead to, for $l \in [\![2, M+1]\!]$

$$\alpha^l = k^2\sum_{j=2}^{l}\sum_{n=2}^{j}\psi^n + t_l\partial^1\alpha^1 + \alpha^0, \tag{10}$$

where we choose $\alpha_i^0 = (u^0)_{xx}(x_i)$ and $\alpha_i^1 = k(u^1)_{xx}(x_i) + (u^0)_{xx}(x_i)$, for all $i \in [\![1, N]\!]$.

2. *Approximation of* u_{ttx}. One remarks that $u_{ttx} = f_x + u_{xxx}$, one can deduce an approximation for u_{ttx} from an approximation of u_{xxx}. To derive an approximation for u_{xxx}, we remark that the function $w = u_{xxx}$ is satisfying $w_{tt} - w_{xx} = f_{xxx}$. Using boundary conditions of φ yields the Neumann boundary conditions $w_x(0, t) = F_1(t)$ and $w_x(1, t) = F_2(1, t)$, $t \in (0, T)$. In this way, $w_{tt} - w_{xx} = f_{xxx}$ is discretized as: for all $n \in [\![1, M]\!]$, find $w^{n+1} = \left(w_i^n\right)_{i=1}^N \in \mathscr{X}(\mathscr{T})$ such that

$$h_i \partial^2 w_i^{n+1} - \mathbb{D}_{i+\frac{1}{2}}\left(w^{n+1}\right) = \frac{1}{k}\int_{nk}^{(n+1)k}\int_{K_i} f_{xxx}(x,t)dx\,dt, \quad \forall\, i \in [\![2, N-1]\!] \tag{11}$$

$$h_1 \partial^2 w_1^{n+1} - \mathbb{F}_{\frac{3}{2}}\left(w^{n+1}\right) = \frac{1}{k}\int_{nk}^{(n+1)k}\int_{K_1} f_{xxx}(x,t)dx\,dt - F_1(t_{n+1}), \tag{12}$$

$$h_N \partial^2 w_N^{n+1} + \mathbb{F}_{N-\frac{1}{2}}\left(w^{n+1}\right) = \frac{1}{k}\int_{nk}^{(n+1)k}\int_{K_N} f_{xxx}(x,t)dx\,dt + F_2(t_{n+1}). \tag{13}$$

Using (2) yields $w(x, 0) = (u^0)_{xxx}(x)$ for all $x \in \mathbf{I}$. This initial condition can be discretized by

$$-\mathbb{D}_{i+\frac{1}{2}}\left(w^0\right) = -\int_{K_i}(u^0)_{xxxxx}(x)dx, \ \ \forall i \in [\![2, N-1]\!], \tag{14}$$

$$-\mathbb{F}_{\frac{3}{2}}\left(w^0\right) = -\int_{K_1}(u^0)_{xxxxx}(x)dx - (u^0)_{xxxx}(0), \tag{15}$$

$$\mathbb{F}_{N-\frac{1}{2}}\left(w^0\right) = -\int_{K_N}(u^0)_{xxxxx}(x)dx + (u^0)_{xxxx}(1). \tag{16}$$

We add the following condition in order to get the well-posedness of (14)–(16):

$$\sum_{i=1}^N h_i w_i^0 = \int_0^1 (u^0)_{xxx}(x)dx. \tag{17}$$

We also have $w_t(x, 0) = (u^1)_{xxx}(x)$ for all $x \in \mathbf{I}$. This second initial condition is discretized by

$$-\mathbb{D}_{i+\frac{1}{2}}\left(\partial^1 w^1\right) = -\int_{K_i} (u^1)_{xxxxx}(x)dx, \quad \forall i \in [\![2, N-1]\!], \tag{18}$$

$$-\mathbb{F}_{\frac{3}{2}}\left(\partial^1 w^1\right) = -\int_{K_1} (u^1)_{xxxxx}(x)dx - \partial^1 F_1(t_1), \tag{19}$$

$$\mathbb{F}_{N-\frac{1}{2}}\left(\partial^1 w^1\right) = -\int_{K_N} (u^1)_{xxxxx}(x)dx + \partial^1 F_2(t_1). \tag{20}$$

An additional condition should be added on the mean value:

$$\sum_{i=1}^{N} h_i \partial^1 w_i^1 = \int_0^1 (u^1)_{xxx}(x)dx. \tag{21}$$

As approximation for u_{ttx} on the mesh points, we suggest $l_i^n = f_x(x_i, t_n) + w_i^n$.

3. *Approximation of* u_{xxt}. As an approximation for $z(x_i, t_n) = u_{xxt}(x_i, t_n)$, we suggest $z_i^n = \partial^1 \alpha_i^{n+1}$.
4. *Approximation of* u_{ttt}. As an approximation for $s(x_i, t_n) = u_{ttt}(x_i, t_n)$, we suggest, since $u_{ttt} = f_t + u_{xxt}$, $s_i^n = f_t(x_i, t_n) + z_i^n$.

Let us consider the expression

$$d_i^n = -h_i \frac{h_i^+ - h_i^-}{2} l_i^n - k\frac{h_i}{2} z_i^n - k\frac{h_i}{2} s_i^n - \frac{h_{i+1}^- - h_i^+}{2}\alpha_{i+1}^{n+1} + \frac{h_i^- - h_{i-1}^+}{2}\alpha_i^{n+1}. \tag{22}$$

Definition 2 (*Definition of a new second order approximation*) We define the new approximation $u^{n,1} = (u_i^{n,1})_{i\in[\![1,N]\!]}$, for any $n \in [\![0, M+1]\!]$ as

$$h_i\, \partial^2 u_i^{n+1,1} - \mathbb{D}_{i+\frac{1}{2}}\left(u^{n+1,1}\right) = \frac{1}{k}\int_{nk}^{(n+1)k}\int_{K_i} f(x,t)dx\, dt + d_i^n, \quad \forall\, i \in [\![1, N]\!] \tag{23}$$

with $u_0^{n,1} = u_{N+1}^{n,1} = 0$ for all $n \in [\![0, M+1]\!]$ and, for any $i \in [\![1, N]\!]$

$$-\mathbb{D}_{i+\frac{1}{2}}\left(u^{0,1}\right) = -\int_{K_i} (u^0)_{xx}(x)dx - \frac{1}{2}\delta_1\left((h_{i+1}^- - h_i^+)\alpha_{i+1}^0\right), \tag{24}$$

$$-\mathbb{D}_{i+\frac{1}{2}}\left(u^{1,1}\right) = -\int_{K_i} (u^0 + k\bar{u}^1)_{xx}(x)dx - \frac{1}{2}\delta_1\left((h_{i+1}^- - h_i^+)\alpha_{i+1}^1\right), \tag{25}$$

where $\bar{u}^1 = u^1 + \frac{k}{2}(f(0) + (u^0)_{xx})$ and $\delta_1 v_i = v_i - v_{i-1}$.

One of the main results of this contribution is the following theorem:

Theorem 1 (Error estimate of the approximation) *(23)–(25) Let $\boldsymbol{I} = (0, 1)$ and $T > 0$ be given. Assume that the solution of (1)–(3) satisfies $u \in \mathscr{C}^4([0, T];\ \mathscr{C}^6(\bar{\boldsymbol{I}}))$. Let $\mathscr{T}$ be an admissible mesh in the sense of Definition 1. Let $k = \frac{T}{M+1}$, with $M \in \mathbb{N}^\star$, and denote by $t_n = nk$, for $n \in [\![0, M+1]\!]$. There exists unique solutions $(\varphi^n)_{n\in[\![0,M+1]\!]}$ and $(w^n)_{n\in[\![0,M+1]\!]}$ for respectively (8)–(9), and (11)–(21). We define α^l by (10). There exists a unique solution $\left(u_i^{n,1}\right)_{i\in[\![1,N]\!],n\in[\![0,M+1]\!]}$ for (23)–(25). For each $n \in [\![0, M+1]\!]$, let $e_{\mathscr{T}}^{n,1} \in \mathscr{X}(\mathscr{T})$ be defined by $e_{\mathscr{T}}^{n,1}(x) = u(x_i, t_n) - u_i^{n,1}$, for a.e. $x \in K_i$, for all $K_i \in \mathscr{T}$. Then, the following error estimates hold:*

- *Discrete $\mathbb{L}^\infty(0, T; H_0^1(\boldsymbol{I}))$–estimate: for all $n \in [\![0, M+1]\!]$*

$$\| e_{\mathscr{T}}^{n,1} \|_{1,\mathscr{T}} \leq C(h+k)^2 \| u \|_{\mathscr{C}^4([0,T];\ \mathscr{C}^6(\bar{\boldsymbol{I}}))}. \tag{26}$$

- *$\mathscr{W}^{1,\infty}(0, T; \mathbb{L}^2(\boldsymbol{I}))$–estimate: for all $n \in [\![1, M+1]\!]$*

$$\| \partial^1 e_{\mathscr{T}}^{n,1} \|_{\mathbb{L}^2(\boldsymbol{I})} \leq C(h+k)^2 \| u \|_{\mathscr{C}^4([0,T];\ \mathscr{C}^6(\bar{\boldsymbol{I}}))}. \tag{27}$$

The following new *a priori estimate* will help us to prove Theorem 1:

Lemma 1 (A new a priori estimate result) *We assume the same discretizations as in Theorem 1. Let $\left(\mathscr{S}_i^n\right)_{n\in[\![0,M+1]\!];\ i\in[\![1,N]\!]}$ and $\left(\mathscr{F}_{i+\frac{1}{2}}^n\right)_{n\in[\![0,M+1]\!];\ i\in[\![0,N]\!]}$ be given. There exists a unique solution $(\eta^n)_{n=0}^{M+1} \in (\mathscr{X}(\mathscr{T}))^{M+2}$ such that for any $n \in [\![1, M]\!]$, for all $i \in [\![1, N]\!]$*

$$h_i \partial^2 \eta_i^{n+1} - \mathbb{D}_{i+\frac{1}{2}}\left(\eta^{n+1}\right) = h_i \mathscr{S}_i^{n+1} + \mathscr{F}_{i+\frac{1}{2}}^{n+1} - \mathscr{F}_{i-\frac{1}{2}}^{n+1}, \tag{28}$$

with, for all $j \in \{0, 1\}$

$$-\mathbb{D}_{i+\frac{1}{2}}\left(\eta^j\right) = h_i \mathscr{S}_i^j + \mathscr{F}_{i+\frac{1}{2}}^j - \mathscr{F}_{i-\frac{1}{2}}^j, \ \forall i \in [\![1, N]\!] \tag{29}$$

and $\eta_0^n = \eta_{N+1}^n = 0$, for all $n \in [\![0, M+1]\!]$, and η_i^n denotes the value taken by η^n on the control volume K_i. Then, the following estimate holds: for all $n \in [\![1, M+1]\!]$

$$\| \eta^n \|_{1,\mathscr{T}} + \| \eta^0 \|_{1,\mathscr{T}} + \| \partial^1 \eta^n \|_{\mathbb{L}^2(\boldsymbol{I})} \leq C\left(\mathscr{S} + \mathscr{S}_1 + \mathscr{F}_0 + \mathscr{F}_2\right), \tag{30}$$

where $\mathscr{S} = \max_{n=0}^{M+1}(\sum_{i=1}^N h_i(\mathscr{S}_i^n)^2)^{\frac{1}{2}}$, $\mathscr{S}_1 = (\sum_{i=1}^N h_i(\partial^1 \mathscr{S}_i^1)^2)^{\frac{1}{2}}$, and $\mathscr{F}_j = \max_{n=j}^{M+1}(\sum_{i=0}^N h_{i+\frac{1}{2}}(\partial^j \mathscr{F}_{i+\frac{1}{2}}^n)^2)^{\frac{1}{2}}$.

Sketch of the proof of Lemma 1: The well-posedness of (28)–(29) is given in [1]. To prove the desired estimate (30), we follow the following steps:
Decomposition for η^n. For all $n \in [\![0, M+1]\!]$, let $\theta^n = \left(\theta_i^n\right)_{i=1}^N \in \mathscr{X}(\mathscr{T})$ be the solution of

$$-\mathbb{D}_{i+\frac{1}{2}}\left(\theta^n\right) = \mathscr{F}^n_{i+\frac{1}{2}} - \mathscr{F}^n_{i-\frac{1}{2}}, \tag{31}$$

with $\theta_0^n = \theta_{N+1}^n = 0$, and we consider the solution of $(\psi^n)_{n=0}^{M+1} \in (\mathscr{X}(\mathscr{T}))^{M+2}$, with $\psi^n = \left(\psi_i^n\right)_{i=1}^N$, such that $\psi_0^n = \psi_{N+1}^n = 0$, for all $n \in [\![0, M+1]\!]$, and for all $j \in \{0, 1\}$

$$-\mathbb{D}_{i+\frac{1}{2}}\left(\psi^j\right) = h_i \mathscr{S}_i^j, \tag{32}$$

and for any $n \in [\![1, M]\!]$, for all $i \in [\![1, N]\!]$

$$h_i \partial^2 \psi_i^{n+1} - \mathbb{D}_{i+\frac{1}{2}}\left(\psi^{n+1}\right) = h_i(\mathscr{S}_i^{n+1} - \partial^2 \theta_i^{n+1}). \tag{33}$$

Thanks to the existence and uniqueness arguments, we can justify that, for all $n \in [\![0, M+1]\!]$

$$\eta^n = \theta^n + \psi^n. \tag{34}$$

Estimate on θ^n. Acting the operator ∂^j, where $j \in \{0, 1, 2\}$, on the both sides of (31), multiplying both sides of the result by $\partial^j \theta_i^{n+1}$, summing over $i \in [\![1, N]\!]$, re-ordering the sums, and using the Cauchy Schwarz inequality to get, for $n \in [\![j, M+1]\!]$

$$\| \partial^j \theta^n \|_{1,\mathscr{T}} \leq C \mathscr{F}_j. \tag{35}$$

Estimate on ψ^n. Acting the operator ∂^l, where $l \in \{0, 1\}$, on the both sides of (32), multiplying both sides of the result by $\partial^l \psi_i^j$, summing over $i \in [\![1, N]\!]$, re-ordering the sums, using the Cauchy Schwarz inequality, and using the discrete Poincaré inequality to get

$$\| \psi^0 \|_{1,\mathscr{T}} + \| \psi^1 \|_{1,\mathscr{T}} + \| \partial^1 \psi^1 \|_{1,\mathscr{T}} \leq C(\mathscr{S} + \mathscr{S}_1). \tag{36}$$

Using the a priori estimate result of [1, Lemma 4.7] on (33) yields that, for all $n \in [\![2, M+1]\!]$

$$\| \psi^n \|_{1,\mathscr{T}} + \| \partial^1 \psi^n \|_{\mathbb{L}^2(\mathbf{I})} \leq C(\mathscr{S} + \max_{n=1}^{M} \| \partial^2 \theta^{n+1} \|_{\mathbb{L}^2(\mathbf{I})} + \| \psi^1 \|_{1,\mathscr{T}} + \| \partial^1 \psi^1 \|_{\mathbb{L}^2(\mathbf{I})}). \tag{37}$$

Using the discrete Poincaré inequality and (35) (when $j = 2$) yields $\max_{n=1}^{M} \|\partial^2\theta^{n+1}\|_{\mathbb{L}^2(\mathbf{I})} \leq C\mathscr{F}_2$. Consequently, estimates (36)–(37) become as, for all $n \in [\![1, M+1]\!]$

$$\| \psi^n \|_{1,\mathscr{T}} + \| \psi^0 \|_{1,\mathscr{T}} + \| \partial^1 \psi^n \|_{\mathbb{L}^2(\mathbf{I})} \leq C(\mathscr{S} + \mathscr{S}_1 + \mathscr{F}_2). \tag{38}$$

Proof of estimate (30). Gathering (34), (35), (38) yields (30). ■

The following lemma provides a convergence analysis for α^l given by (10).

Lemma 2 (convergence of (10)) *We assume the same discretizations as in Theorem 1 and that the solution of (1)–(3) satisfies* $u \in \mathscr{C}^3([0,T];\ \mathscr{C}^6(\bar{\mathbf{I}}))$. *Assume that there exists a unique solution* $(\varphi^n)_{n\in[\![0,M+1]\!]} \in (\mathscr{X}(\mathscr{T}))^{M+2}$ *for (8)–(9). Let* $(\psi^n)_{n\in[\![0,M+1]\!]} \in (\mathscr{X}(\mathscr{T}))^{M+2}$ *be defined by* $\psi^n = (f_{xx}(x_i,t_n))_{i=1}^N + \varphi^n$. *We define* α^l *by (10) and* $\alpha_i^0 = (u^0)_{xx}(x_i)$ *and* $\alpha_i^1 = k(u^1)_{xx}(x_i) + (u^0)_{xx}(x_i)$, *for all* $i \in [\![1, N]\!]$. *Then, the following error estimates hold:*

- $\mathscr{W}^{2,\infty}(0,T;H_0^1(\mathbf{I}))$*–estimate: for all* $n \in [\![2, M+1]\!]$

$$\| \partial^2 \alpha^n - \Pi_{\mathscr{T}} u_{xxtt}(t_n) \|_{1,\mathscr{T}} \leq C(h+k)\| u \|_{\mathscr{C}^3([0,T];\ \mathscr{C}^6(\bar{\mathbf{I}}))}, \tag{39}$$

- $\mathscr{W}^{j,\infty}(0,T;\mathbb{L}^\infty(\mathbf{I}))$*–estimate, with* $j \in [\![0, 2]\!]$*: for all* $n \in [\![j, M+1]\!]$

$$\| \partial^j \left(\alpha^n - \Pi_{\mathscr{T}} u_{xx}(t_n) \right) \|_{\mathbb{L}^\infty(\mathbf{I})} \leq C(h+k)\| u \|_{\mathscr{C}^3([0,T];\ \mathscr{C}^6(\bar{\mathbf{I}}))}. \tag{40}$$

An overview on the proof of Lemma 2: The proof is based on the fact that α^l is a *second numerical integration in time* for $\psi^n = (\psi_i^n)_{i\in[\![1,N]\!]}$ which is an approximation for u_{xxtt}. However, estimates (39)–(40) demand a rather longer proof. We will detail this in future papers. ■

Sketch of the proof of Theorem 1: The well-posedness of the schemes involved in the scheme (23)–(25) can be justified using techniques of [1] and [2, pp. 794–797]. Integrating equation (1) over $K_i \times (t_n, t_{n+1})$ and using some convenient Taylor expansions to get

$$h_i\, \partial^2 u(x_i,t_{n+1}) - \mathbb{D}_{i+\frac{1}{2}}(u(t_{n+1})) = \frac{1}{k}\int_{nk}^{(n+1)k}\int_{K_i} f(x,t)dx\,dt + \overline{d}_i^n,\ \forall\, n \in [\![1, M]\!] \tag{41}$$

where, we denoted $u(t_{n+1}) = (u(x_i,t_{n+1}))_{i=1}^N$, and

$$\begin{aligned}\overline{d}_i^n = &-h_i\frac{h_i^+ - h_i^-}{2}u_{ttx}(x_i,t_n) - k\frac{h_i}{2}u_{xxt}(x_i,t_{n+1})\\ &- k\frac{h_i}{2}u_{ttt}(x_i,t_n) - \frac{h_{i+1}^- - h_i^+}{2}u_{xx}(x_{i+\frac{1}{2}},t_{n+1})\\ &+ \frac{h_i^- - h_{i-1}^+}{2}u_{xx}(x_{i-\frac{1}{2}},t_{n+1}) + h_i r_i^{n,1} + \mathscr{R}^1_{i+\frac{1}{2}}(u(t_{n+1})) - \mathscr{R}^1_{i+\frac{1}{2}}(u(t_{n+1})),\end{aligned} \tag{42}$$

with $|\, r_i^{n,1}|$ and $|\, \partial^j \mathscr{R}^1_{i+\frac{1}{2}}(u(t_{n+1}))|$, for $j \in [\![0,2]\!]$, are bounded above by $C(k+h)^2 \| u\|_{\mathscr{C}^4([0,T];\, \mathscr{C}^3(\bar{\mathbf{I}}))}$. Thanks to the initial condition (2) of the exact solution with a convenient Taylor expansion, we get

$$\begin{aligned}-\mathbb{D}_{i+\frac{1}{2}}(u(0)) = &-\int_{K_i}(u^0)_{xx}(x)dx - \frac{h_{i+1}^- - h_i^+}{2}(u^0)_{xx}(x_{i+\frac{1}{2}})\\ &+ \frac{h_i^- - h_{i-1}^+}{2}(u^0)_{xx}(x_{i-\frac{1}{2}})\\ &+ \mathscr{R}^1_{i+\frac{1}{2}}(u(0)) - \mathscr{R}^1_{i-\frac{1}{2}}(u(0)),\end{aligned} \tag{43}$$

$$\begin{aligned}-\mathbb{D}_{i+\frac{1}{2}}(u(t_1)) = &-\int_{K_i}\left(u^0 + k\bar{u}^1\right)_{xx}(x)dx - \frac{h_{i+1}^- - h_i^+}{2}u_{xx}(x_{i+\frac{1}{2}},t_1)\\ &+ \frac{h_i^- - h_{i-1}^+}{2}u_{xx}(x_{i-\frac{1}{2}},t_1)\\ &+ \mathscr{R}^1_{i+\frac{1}{2}}(u(t_1)) - \mathscr{R}^1_{i-\frac{1}{2}}(u(t_1)) - k\int_{K_i}\gamma_{xx}(x)dx,\end{aligned} \tag{44}$$

where $\gamma = \dfrac{1}{2k}\displaystyle\int_0^{t_1}(t_1-t)^2 u_{ttt}(t)dt$. We will use estimate $\| \gamma_{xx}\|_{\mathscr{C}(\bar{\mathbf{I}})} \le \dfrac{k^2}{2}$ $\| u\|_{\mathscr{C}^3([0,T];\, \mathscr{C}^2(\bar{\mathbf{I}}))}$.

Subtracting (23), (24), and (25) (multiplied by k) from (41), (43), and (44), respectively, leads to

$$h_i\, \partial^2 e_i^{n+1,1} - \mathbb{D}_{i+\frac{1}{2}}\left(e^{n+1,1}\right) = h_i \mathscr{S}_i^{n+1} + \mathscr{F}_{i+\frac{1}{2}}^{n+1} - \mathscr{F}_{i-\frac{1}{2}}^{n+1},\ \forall\, i \in [\![1,N]\!],\ \forall\, n \in [\![1,M]\!] \tag{45}$$

with, for all $j \in \{0,1\}$

$$-\mathbb{D}_{i+\frac{1}{2}}\left(e^{j,1}\right) = h_i \mathscr{S}_i^j + \mathscr{F}_{i+\frac{1}{2}}^j - \mathscr{F}_{i-\frac{1}{2}}^j,\ \forall i \in [\![1,N]\!] \tag{46}$$

and $e_0^{n,1} = e_{N+1}^{n,1} = 0$, for all $n \in [\![0, M+1]\!]$ and $e_i^{n,1} = u(x_i, t_n) - u_i^{n,1}$, whereas the terms $\mathscr{S}_i^n$ and $\mathscr{F}_{i+\frac{1}{2}}^n$ are given by $\mathscr{S}_i^1 = -\dfrac{k}{h_i} \displaystyle\int_{K_i} \gamma_{xx}(x)dx$, $\mathscr{S}_i^0 = 0$, and

$$
\begin{aligned}
Ss_i^{n+1} = & -\frac{h_i^+ - h_i^-}{2} \left(u_{ttx}(x_i, t_n) - l_i^n \right) - \frac{k}{2} \left(u_{xxt}(x_i, t_{n+1}) - z_i^{n+1} \right) \\
& - \frac{k}{2} \left(u_{ttt}(x_i, t_n) - s_i^n \right) + r_i^{n,1}, \\
\mathscr{F}_{i+\frac{1}{2}}^n = & -\frac{h_{i+1}^- - h_i^+}{2} \left(u_{xx}(x_{i+\frac{1}{2}}, t_n) - \alpha_{i+1}^n \right) + \mathscr{R}_{i+\frac{1}{2}}^1 (u(t_n)).
\end{aligned}
$$

One remarks that $\left(e_{\mathscr{T}}^{n,1} \right)_{n=0}^{M+1}$ satisfies hypothesis of Lemma 1, one can apply estimate (30) to get

$$
\| e_{\mathscr{T}}^{n,1} \|_{1,\mathscr{T}} + \| e_{\mathscr{T}}^{0,1} \|_{1,\mathscr{T}} + \| \partial^1 e_{\mathscr{T}}^{n,1} \|_{\mathbb{L}^2(\mathbf{I})} \leq C \left(\mathscr{S} + \mathscr{S}_1 + \mathscr{F}_0 + \mathscr{F}_2 \right), \quad \forall n \in [\![1, M+1]\!], \tag{47}
$$

where $\mathscr{S}$, $\mathscr{S}_1$, $\mathscr{F}_0$, $\mathscr{F}_2$ are defined in Lemma 1. Using Lemma 2 implies that $\mathscr{F}_0 + \mathscr{F}_2 \leq C(h+k)^2 \| u \|_{\mathscr{C}^4([0,T];\, \mathscr{C}^6(\bar{\mathbf{I}}))}$. One remarks that (8)–(9) (resp. (11)–(21)) is an approximation for the wave equation $\varphi_{tt} - \varphi_{xx} = f_{xxxx}$ (resp. $w_{tt} - w_{xx} = f_{xxx}$) with Dirichlet (resp. Neumann) boundary conditions, one can apply [1, Theorem 4.1] to obtain error estimates for schemes (8)–(9) and (11)–(21) which yields that $\mathscr{S} + \mathscr{S}_1 \leq C(h+k)^2 \| u \|_{\mathscr{C}^4([0,T];\, \mathscr{C}^6(\bar{\mathbf{I}}))}$. ■

3 A Numerical Simulation on a Non-Uniform Spatial Mesh

Consider the example of problem (1)–(3) when $u(x,t) = \sin(\pi x)\cos(\pi t)$, $(x,t) \in (0,1) \times (0,1)$. We set $h_i = h$ for even i, $h_i = h/2$ for odd i, and $x_i = (x_{i-\frac{1}{2}} + x_{i+\frac{1}{2}})/2$, for $i \in [\![1, N]\!]$. The step is taken as $k = h/2$. The following table shows that the order of scheme (23)–(25) is two:

h	Error in $W^{1,\infty}(L^2)$		Error in $L^\infty(H_0^1)$	
	Error	Order	Error	Order
1/225	0.0000946	–	0.0000770	–
1/300	0.0000533	1.9942888	0.0000434	1.9929847
1/375	0.0000341	2.0015768	0.0000278	1.9961295
1/450	0.0000237	1.9954982	0.0000193	2.0015786

References

1. Bradji, A.: A theoretical analysis of a new finite volume scheme for second order hyperbolic equations on general nonconforming multidimensional spatial meshes. Numer. Meth. Partial Differ. Equ. **29**(1), 1–39 (2013)
2. Eymard, R., Gallouët, T., Herbin, R.: Finite volume methods. In: Ciarlet, P.G., Lions, J.L. (eds.) Handbook of Numerical Analysis, vol. VII, pp. 723–1020. Elsevier, NorthHolland, Amsterdam (2000)

Note on the Convergence of a Finite Volume Scheme Using a General Nonconforming Mesh for an Oblique Derivative Boundary Value Problem

Abdallah Bradji

Abstract This note is an extension of our work [1], which dealt with the convergence of a finite volume approximation using the admissible mesh of [3], for oblique derivative boundary value problems. In this note, we provide a finite volume scheme for the Laplace equation with an oblique boundary condition, using the general nonconforming meshes and the discrete gradient introduced recently in [4]. A convergence order for an approximation for the gradient of the exact solution is proved.

MSC2010: 65M08, 65M12, 65M15

1 Motivation and Aim of This Paper

In this note, we are interested with the finite volume approximation of the following Laplace problem, on an open bounded polygonal connected subset Ω of $\mathbb{R}^2$, with oblique boundary condition:

$$-\Delta u(\mathbf{x}) = f(\mathbf{x}), \mathbf{x} \in \Omega, \tag{1}$$

and, for the sake of simplicity, we consider homogeneous boundary conditions

$$u_n(\mathbf{x}) + \alpha u_t(\mathbf{x}) = 0, \ \ \mathbf{x} \in \partial\Omega. \tag{2}$$

where $\mathbf{x} = (x, y)$ is the current point of $\mathbb{R}^2$, $u_n = \nabla u \cdot \mathbf{n}$ and $u_t = \nabla u \cdot \mathbf{t}$, with $\mathbf{n} = (\mathbf{n}_x, \mathbf{n}_y)^t$ (resp. $\mathbf{t} = (-\mathbf{n}_y, \mathbf{n}_x)^t$) is the normal vector to the boundary $\partial\Omega$ and outward to Ω (resp. is a tangential derivative), and α is a given constant. We will assume that $\alpha > 0$. The case when $\alpha < 0$ can be handled in a similar way whereas

A. Bradji (✉)
Department of Mathematics, Faculty of Sciences, University of Badji Mokhtar-Annaba, Annaba, Algeria
e-mail: bradji@cmi.univ-mrs.fr; abdallah-bradji@univ-annaba.org

J. Fuhrmann et al. (eds.), *Finite Volumes for Complex Applications VII - Methods and Theoretical Aspects,* Springer Proceedings in Mathematics & Statistics 77,
DOI: 10.1007/978-3-319-05684-5_13,

$\alpha = 0$ (which is the case of Laplace equation with Neumann boundary conditions) can be treated as in [4].

Problem (1)–(2) appears, for instance, in a method developed in [2] for improving the convergence order of numerical schemes for the classical Dirichlet problem (it can probably also appear in the modeling of some mechanical problems, but perhaps not directly under the form (1)–(2), see [1, 5] and references therein).

We give now a sens for the operators of normal and tangential derivatives which will allow us to define the weak formulation (6)–(7) of Theorem 1 for problem (1)–(2). The definition (3) (resp. (4)) of normal derivative operator (resp. tangential derivative operator) as well as Theorem 1 are already provided in [1].

We denote by $H^1(\Omega)$ the Sobolev space of functions which together with their first generalized derivatives are in $L^2(\Omega)$. The norm of $H^1(\Omega)$ is defined by

$$\|w\|_{1,\Omega}^2 = \|w\|_{L^2(\Omega)}^2 + \|\nabla w\|_{L^2(\Omega)}^2.$$

We denote by $H_0^1(\Omega)$ the space $\{v \in H^1(\Omega) : \tilde{\gamma}(v) = 0\}$, where $\tilde{\gamma}$ is the linear trace operator from $H^1(\Omega)$ to $L^2(\partial\Omega)$.

We define $\mathscr{D}(\Omega)$ as the linear space of infinitely differentiable functions, with compact support on Ω, and we set $\mathscr{D}(\overline{\Omega}) = \{\varphi|_\Omega, \varphi \in \mathscr{D}(\mathbb{R}^2)\}$.

We recall the density results $\overline{\mathscr{D}(\Omega)} = H_0^1(\Omega)$and $\overline{\mathscr{D}(\overline{\Omega})} = H^1(\Omega)$. We denote by $H^{\frac{1}{2}}(\partial\Omega)$ the space of the traces of the elements of $H^1(\Omega)$ equipped with the norm

$$\|w\|_{\frac{1}{2},\partial\Omega} = \inf_{\tilde{\gamma}(v)=w} \|v\|_{1,\Omega},$$

We define the operator of normal derivative acting on u as an element u_n of $H^{-\frac{1}{2}}(\partial\Omega)$ defined as: if $u \in H^1(\Omega)$ and $-\Delta u = f \in L^2(\Omega)$,

$$\langle u_n, v\rangle_{H^{-\frac{1}{2}}(\partial\Omega), H^{\frac{1}{2}}(\partial\Omega)} = \int_\Omega \nabla u \cdot \nabla \tilde{v} d\mathbf{x} - \int_\Omega f\tilde{v} d\mathbf{x}, \forall v \in H^{\frac{1}{2}}(\partial\Omega). \quad (3)$$

We define the operator of tangential derivative acting on u as an element u_t of $H^{-\frac{1}{2}}(\partial\Omega)$ defined as: if $u \in H^1(\Omega)$,

$$\langle u_t, v\rangle_{H^{-\frac{1}{2}}(\partial\Omega), H^{\frac{1}{2}}(\partial\Omega)} = \int_\Omega \tilde{v}_x u_y d\mathbf{x} - \int_\Omega u_x \tilde{v}_y d\mathbf{x}, \forall v \in H^{\frac{1}{2}}(\partial\Omega), \quad (4)$$

where $\tilde{v}$, in (3) and (4), is an element of $H^1(\Omega)$ such that $\tilde{\gamma}(\tilde{v}) = v$.

It is easy to justify that the operators of normal and tangential derivatives are well defined.

We are able now to define the variational formulation to (1)–(2). To get the well-posedness to (1)–(2), we assume that $f \in \mathbb{L}^2(\Omega)$ and the following compatibility condition is satisfied:

$$\int_\Omega f(\mathbf{x})\, d\mathbf{x} = 0. \quad (5)$$

The following theorem gives the existence of a weak solution for problem (1)–(2)

Theorem 1 (Existence for problem (1)–(2), cf. [1]) *Under the assumption that the function f is satisfying $f \in \mathbb{L}^2(\Omega)$ and the compatibility condition (5), the problem (1)–(2) has a unique solution in the following sense:*

$$u \in H^1(\Omega), \int_\Omega u(\boldsymbol{x})d\boldsymbol{x} = 0, \tag{6}$$

and,

$$\int_\Omega \nabla u.\nabla v \, d\boldsymbol{x} + \alpha \int_\Omega (v_x u_y - u_x v_y)\, d\boldsymbol{x} = \int_\Omega f v \, d\boldsymbol{x}, \forall v \in H^1(\Omega). \tag{7}$$

The problem (1)–(2) is approximated in [1] using the admissible mesh of [3]. In this work, we provide a finite volume scheme for (1)–(2) using the general mesh introduced recently in [4] along with a convergence order for the scheme. To derive a finite volume scheme for (1)–(2) and to prove its convergence, we use the following lemma (see [1])

Lemma 1 (Useful lemma) *Let a and b be two points in $\mathbb{R}^2$ and $(a, b) = \{sa + (1 - s)b, s \in (0, 1)\}$. Let $f \in \mathscr{C}^1(\mathbb{R}^2)$ and $t = \frac{b-a}{|b-a|}$. Let $f_t = \nabla f \cdot \boldsymbol{t}$ (it is the tangential derivative of f). Then*

$$\int_{(a,b)} f_t(\boldsymbol{x})d\gamma(\boldsymbol{x}) = f(b) - f(a). \tag{8}$$

Remark 1 (Another way to approximate (1)–(2)) The problem (1)–(2) can be written as a Neumann problem (with, if $\alpha \neq 0$, a non symmetric operator):

$$\begin{cases} -\mathrm{div}(\mathscr{A} \nabla u(\mathbf{x})) = f(\mathbf{x}), \mathbf{x} \in \Omega, \\ (\mathscr{A} \nabla u(\mathbf{x})) \cdot \mathbf{n}(\mathbf{x}) = 0, \ \mathbf{x} \in \partial\Omega, \end{cases} \tag{9}$$

where the positive definite matrix $\mathscr{A}$ is given by $\mathscr{A} = \begin{pmatrix} 1 & \alpha \\ -\alpha & 1 \end{pmatrix}$. But the finite volume scheme we shall present will be derived directly from the problem (1)–(2) (as followed in [1]) and not from (9).

2 Definition of the Scheme and Statement of the Main Result

The discretization of Ω is performed using the mesh $\mathscr{D} = (\mathscr{M}, \mathscr{E}, \mathscr{P})$ described in [4, Definition 2.1] which we recall here for the sake of completeness.

Definition 1 (*Definition of a general mesh for a general domain*) Let Ω be a "general" polyhedral open bounded subset of $\mathbb{R}^d$, where $d \in \mathbb{N} \setminus \{0\}$, and

$\partial\Omega = \overline{\Omega} \setminus \Omega$ its boundary. A discretization of Ω, denoted by $\mathcal{D}$, is defined as the triplet $\mathcal{D} = (\mathcal{M}, \mathcal{E}, \mathcal{P})$, where:

1. $\mathcal{M}$ is a finite family of non empty connected open disjoint subsets of Ω (the "control volumes") such that $\overline{\Omega} = \cup_{K\in\mathcal{M}} \overline{K}$. For any $K \in \mathcal{M}$, let $\partial K = \overline{K} \setminus K$ be the boundary of K; let $\mathrm{m}(K) > 0$ denote the measure of K and h_K denote the diameter of K.
2. $\mathcal{E}$ is a finite family of disjoint subsets of $\overline{\Omega}$ (the "edges" of the mesh), such that, for all $\sigma \in \mathcal{E}$, σ is a non empty open subset of a hyperplane of $\mathbb{R}^d$, whose $(d-1)$–dimensional measure is strictly positive. We also assume that, for all $K \in \mathcal{M}$, there exists a subset $\mathcal{E}_K$ of $\mathcal{E}$ such that $\partial K = \cup_{\sigma\in\mathcal{E}_K} \overline{\sigma}$. For any $\sigma \in \mathcal{E}$, we denote by $\mathcal{M}_\sigma = \{K;\ \sigma \in \mathcal{E}_K\}$. We then assume that, for any $\sigma \in \mathcal{E}$, either $\mathcal{M}_\sigma$ has exactly one element and then $\sigma \subset \partial\Omega$ (the set of these interfaces, called boundary interfaces, denoted by $\mathcal{E}_{\mathrm{ext}}$) or $\mathcal{M}_\sigma$ has exactly two elements (the set of these interfaces, called interior interfaces, denoted by $\mathcal{E}_{\mathrm{int}}$). For all $\sigma \in \mathcal{E}$, we denote by $\mathbf{x}_\sigma$ the barycentre of σ. For all $K \in \mathcal{M}$ and $\sigma \in \mathcal{E}$, we denote by $\mathbf{n}_{K,\sigma}$ the unit vector normal to σ outward to K.
3. $\mathcal{P}$ is a family of points of Ω indexed by $\mathcal{M}$, denoted by $\mathcal{P} = (\mathbf{x}_K)_{K\in\mathcal{M}}$, such that for all $K \in \mathcal{M}$, $\mathbf{x}_K \in K$ and K is assumed to be $\mathbf{x}_K$–star-shaped, which means that for all $\mathbf{x} \in K$, the property $[\mathbf{x}_K, \mathbf{x}] \subset K$ holds. Denoting by $d_{K,\sigma}$ the Euclidean distance between $\mathbf{x}_K$ and the hyperplane including σ, one assumes that $d_{K,\sigma} > 0$. We then denote by $\mathcal{D}_{K,\sigma}$ the cone with vertex $\mathbf{x}_K$ and basis σ.

The following definition will help us to define the finite volume scheme we shall present and to prove its convergence

Definition 2 Let $\sigma \in \mathcal{E}_{\mathrm{ext}}$ and $\mathbf{n}$ be the normal vector to σ, outward to Ω. Recall that $\mathbf{t} = (-\mathbf{n}_y, \mathbf{n}_x)^t$ where $\mathbf{n} = (\mathbf{n}_x, \mathbf{n}_y)^t$, then $\sigma = (a, b) = \{sa + (1-s)b, s \in [0,1]\}$ where a, b are chosen such that $|b-a|\mathbf{t} = b - a$. We denote by σ^- (resp. σ^+) the element of $\mathcal{E}_{\mathrm{ext}}$ such that a is in the closure of σ^- (resp. b is in the closure of σ^+) and $\sigma^- \neq \sigma$ (resp. $\sigma^+ \neq \sigma$). We also set $\sigma_e = b$ and $\sigma_b = a$ (so that $|\sigma_e - \sigma_b|\mathbf{t} = \sigma_e - \sigma_b$).

We define the space $\mathcal{X}_\mathcal{D}$ as the set of all $\left((\nu_K)_{K\in\mathcal{M}}, (\nu_\sigma)_{\sigma\in\mathcal{E}}\right)$, where $\nu_K, \nu_\sigma \in \mathbb{R}$ for all $K \in \mathcal{M}$ and for all $\sigma \in \mathcal{E}$. The space $\mathcal{X}_\mathcal{D}$ is equipped with the semi–norm $|\nu|^2_{\mathcal{X}} = \sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{E}_K} \frac{\mathrm{m}(\sigma)}{d_{K,\sigma}} (\nu_\sigma - \nu_K)^2$. Let $H_\mathcal{M}(\Omega)$ be the space of functions from Ω to $\mathbb{R}$ which are constant over each control volume of the mesh. For all $\nu \in \mathcal{X}_\mathcal{D}$, we denote by $\Pi_\mathcal{M} \nu \in H_\mathcal{M}(\Omega)$ the function defined by $\Pi_\mathcal{M} \nu(x) = \nu_K$, for a.e. $x \in K$, for all $K \in \mathcal{M}$. For all $\varphi \in \mathcal{C}(\Omega)$, we define $\mathcal{P}_\mathcal{D}\varphi = \left((\varphi(x_K))_{K\in\mathcal{M}}, (\varphi(x_\sigma))_{\sigma\in\mathcal{E}}\right) \in \mathcal{X}_\mathcal{D}$. We denote by $\mathcal{P}_\mathcal{M}\varphi \in H_\mathcal{M}(\Omega)$ the function defined by $\mathcal{P}_\mathcal{M}\varphi(x) = \varphi(x_K)$, for a.e. $x \in K$, for all $K \in \mathcal{M}$. In order to analyze the convergence, we need to consider the size of the discretization $\mathcal{D}$ defined by $h_\mathcal{D} = \sup\{\mathrm{diam}(K), K \in \mathcal{M}\}$ and the regularity of the mesh given by $\theta_\mathcal{D} = \max\left(\max_{\sigma\in\mathcal{E}_{\mathrm{int}}, K, L\in\mathcal{M}} \frac{d_{K,\sigma}}{d_{L,\sigma}}, \max_{K\in\mathcal{M}, \sigma\in\mathcal{E}_K} \frac{h_K}{d_{K,\sigma}}\right)$.

Throughout this paper, the letters C_i stand for positive constants which are independent of the parameters of the discretization.

The scheme we want to consider in this note is based on the use of the discrete gradient given in [4]. For $u = \left((u_K)_{K\in\mathcal{M}}, (u_\sigma)_{\sigma\in\mathcal{E}}\right) \in \mathcal{X}_\mathcal{D}$, we define, for all $K \in \mathcal{M}$

$$\nabla_\mathcal{D} u(x) = \nabla_{K,\sigma} u, \quad \text{a. e. } x \in \mathcal{D}_{K,\sigma}, \tag{10}$$

where $\mathcal{D}_{K,\sigma}$ is the cone with vertex x_K and basis σ and

$$\nabla_{K,\sigma} u = \nabla_K u + \left(\frac{\sqrt{d}}{d_{K,\sigma}} (u_\sigma - u_K - \nabla_K u \cdot (x_\sigma - x_K)) \right) \mathbf{n}_{K,\sigma}, \tag{11}$$

where $\nabla_K u = \dfrac{1}{\mathrm{m}(K)} \sum_{\sigma\in\mathcal{E}_K} \mathrm{m}(\sigma)\,(u_\sigma - u_K)\,\mathbf{n}_{K,\sigma}$ and d is the space dimension. We define the finite volume approximation for (1)–(2) as $u_\mathcal{D} \in \mathcal{X}_\mathcal{D}$ such that

$$\langle u_\mathcal{D}, v\rangle_F + \alpha \sum_{\sigma\in\mathcal{E}_{\mathrm{ext}}} (u_\sigma - u_{\sigma^-}) v_\sigma = (f, \Pi_\mathcal{M} v)_{\mathbb{L}^2(\Omega)}, \quad \forall v \in \mathcal{X}_\mathcal{D} \tag{12}$$

and

$$\sum_{K\in\mathcal{M}} \mathrm{m}(K) u_K = 0, \tag{13}$$

where $\langle u, v\rangle_F = \displaystyle\int_\Omega \nabla_\mathcal{D} u(x) \cdot \nabla_\mathcal{D} v(x) dx$ and $(\cdot,\cdot)_{\mathbb{L}^2(\Omega)}$ denotes the $\mathbb{L}^2$-inner product. The main result of the present contribution is the following theorem:

Theorem 2 (Error estimate for scheme (12)–(13)) *Let Ω be an open bounded polygonal connected subset Ω of $\mathbb{R}^2$. Let $\theta > 0$ and let $\mathcal{D} = (\mathcal{M}, \mathcal{E}, \mathcal{P})$ be a discretization in the sense of Definition 1, such that $\theta_\mathcal{D}$ satisfies $\theta \geq \theta_\mathcal{D}$. Under the assumption that the function f is satisfying $f \in \mathbb{L}^2(\Omega)$ and the compatibility condition (5), the finite volume scheme (12)–(13) has a unique solution $u = \left((u_K)_{K\in\mathcal{M}}, (u_\sigma)_{\sigma\in\mathcal{E}}\right) \in \mathcal{X}_\mathcal{D}$.*
Assume in addition that the weak solution u of (6)–(7) is satisfying $u \in \mathcal{C}^2(\overline{\Omega})$. Then, the following error estimate holds:

$$\| \nabla_\mathcal{D} u_\mathcal{D} - \nabla u\|_{\mathbb{L}^2(\Omega)} \leq C_1 \sqrt{h_\mathcal{D}} \| u\|_{\mathcal{C}^2(\overline{\Omega})}. \tag{14}$$

Proof The well-posedness of scheme (12)–(13) can be justified using a reasoning similar to that of [1, p. 8] combined with [4, Lemma 4.2, p. 1026] and the following useful rule (see [1]):

$$\sum_{\sigma\in\mathcal{E}_{\mathrm{ext}}} (u_\sigma - u_{\sigma^-}) u_\sigma = \frac{1}{2} \sum_{\sigma\in\mathcal{E}_{\mathrm{ext}}} (u_\sigma - u_{\sigma^-})^2. \tag{15}$$

Integrating both sides of Eq. (1) over each control volume K, using an integration by parts, multiplying the result by v_K, and summing over the control volumes $K \in \mathcal{M}$ yields that

$$-\sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{E}_K} v_K \int_\sigma \nabla u(x)\cdot\mathbf{n}_{K,\sigma}(x)d\gamma(x) = (f, \Pi_{\mathcal{M}} v)_{\mathbb{L}^2(\Omega)}, \quad \forall v \in \mathcal{X}_{\mathcal{D}}. \tag{16}$$

Since $\int_\sigma \nabla u(\mathbf{x})\cdot\mathbf{n}_{K,\sigma}(\mathbf{x})d\gamma(\mathbf{x}) + \int_\sigma \nabla u(\mathbf{x})\cdot\mathbf{n}_{L,\sigma}(\mathbf{x})d\gamma(\mathbf{x})$, for all $\sigma \in \mathcal{E}$ such that $\mathcal{M}_\sigma = \{K, L\}$ (it stems from the fact that $\mathbf{n}_{K,\sigma} = -\mathbf{n}_{L,\sigma}$), the equality (16) implies that

$$\begin{aligned}&-\sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{E}_K}(v_K - v_\sigma)\int_\sigma \nabla u(\mathbf{x})\cdot\mathbf{n}_{K,\sigma}(\mathbf{x})d\gamma(\mathbf{x})\\ &-\sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{E}_K\cap\mathcal{E}_{\rm ext}} v_\sigma\int_\sigma \nabla u(\mathbf{x})\cdot\mathbf{n}_{K,\sigma}(\mathbf{x})d\gamma(\mathbf{x}) = (f, \Pi_{\mathcal{M}} v)_{\mathbb{L}^2(\Omega)}.\end{aligned} \tag{17}$$

This with the oblique boundary condition (2) and Lemma 1 leads to

$$\begin{aligned}&-\sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{E}_K}(v_K - v_\sigma)\int_\sigma \nabla u(\mathbf{x})\cdot\mathbf{n}_{K,\sigma}(\mathbf{x})d\gamma(\mathbf{x}) + \alpha\sum_{\sigma\in\mathcal{E}_{\rm ext}} v_\sigma(u(\sigma_e) - u(\sigma_b))\\ &\qquad = (f, \Pi_{\mathcal{M}} v)_{\mathbb{L}^2(\Omega)}.\end{aligned} \tag{18}$$

Inserting $(f, \Pi_{\mathcal{M}} v)_{\mathbb{L}^2(\Omega)}$ by its value of (18) in (12) gives

$$\begin{aligned}&\langle \mathcal{P}_{\mathcal{D}} u - u_{\mathcal{D}}, v\rangle_F + \alpha\sum_{\sigma\in\mathcal{E}_{\rm ext}} v_\sigma((u(\sigma_e) - u_\sigma) - (u(\sigma_b) - u_{\sigma^-}))\\ &\qquad = \sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{E}_K}(v_K - v_\sigma)\left(F_{K,\sigma}(\mathcal{P}_{\mathcal{D}} u) + \int_\sigma \nabla u(\mathbf{x})\cdot\mathbf{n}_{K,\sigma}(\mathbf{x})d\gamma(\mathbf{x})\right),\end{aligned} \tag{19}$$

where $F_{K,\sigma}$ are the discrete fluxes defined in [4, (2.22)–(2.25), p. 1018] (this stems from the identification [4, (2.15), p.1017]).
Let us define the error by

$$e_{\mathcal{D}} = \mathcal{P}_{\mathcal{D}} u - u_{\mathcal{D}}. \tag{20}$$

We set

$$\mathcal{R}_{K,\sigma}(u) = F_{K,\sigma}(\mathcal{P}_{\mathcal{D}} u) + \int_\sigma \nabla u(\mathbf{x})\cdot\mathbf{n}_{K,\sigma}(\mathbf{x})d\gamma(\mathbf{x}) \quad \text{and} \quad r_\sigma = u(\mathbf{x}_\sigma) - u(\sigma_e).$$

Thanks to [4, (4.27), p. 1033], the following estimate holds

$$\left(\sum_{K\in\mathscr{M}}\sum_{\sigma\in\mathscr{E}_K}\frac{d_{K,\sigma}}{\mathrm{m}(\sigma)}\left(\mathscr{R}_{K,\sigma}(u)\right)^2\right)^{\frac{1}{2}} \leq C_2 h_{\mathscr{D}}\|u\|_{\mathscr{C}^2(\overline{\Omega})}. \tag{21}$$

Using a Taylor expansion, we have

$$|r_\sigma| \leq C_3\mathrm{m}(\sigma)\|u\|_{\mathscr{C}^1(\overline{\Omega})}. \tag{22}$$

We write equation (19) as

$$\begin{aligned}\langle e_{\mathscr{D}}, v\rangle_F + \alpha\sum_{\sigma\in\mathscr{E}_{\mathrm{ext}}} v_\sigma\,(e_\sigma - e_{\sigma^-}) &= \sum_{K\in\mathscr{M}}\sum_{\sigma\in\mathscr{E}_K}(v_K - v_\sigma)\mathscr{R}_{K,\sigma}(u)\\ &\quad + \alpha\sum_{\sigma\in\mathscr{E}_{\mathrm{ext}}} v_\sigma\,(r_\sigma - r_{\sigma^-}).\end{aligned} \tag{23}$$

Re-ordering the sum in the last term on the right hand side of (23) implies that

$$\begin{aligned}\langle e_{\mathscr{D}}, v\rangle_F + \alpha\sum_{\sigma\in\mathscr{E}_{\mathrm{ext}}} v_\sigma\,(e_\sigma - e_{\sigma^-}) &= \sum_{K\in\mathscr{M}}\sum_{\sigma\in\mathscr{E}_K}(v_K - v_\sigma)\mathscr{R}_{K,\sigma}(u)\\ &\quad - \alpha\sum_{\sigma\in\mathscr{E}_{\mathrm{ext}}} r_\sigma\,(v_{\sigma^+} - v_\sigma).\end{aligned} \tag{24}$$

We use the following stability result provided in [4, Lemma 4.2]

$$C_4|v|_{\mathscr{X}} \leq \|\nabla_{\mathscr{D}} v\|_{\mathbb{L}^2(\Omega)} \leq C_5|v|_{\mathscr{X}},\ \forall v\in\mathscr{X}_{\mathscr{D}}. \tag{25}$$

Taking $v = e_{\mathscr{D}}$ in (24) and using (25) with the rule (15) to get

$$\begin{aligned}C_4^2|e_{\mathscr{D}}|_{\mathscr{X}}^2 + \frac{\alpha}{2}\sum_{\sigma\in\mathscr{E}_{\mathrm{ext}}}(e_\sigma - e_{\sigma^-})^2 &\leq \sum_{K\in\mathscr{M}}\sum_{\sigma\in\mathscr{E}_K}(e_K - e_\sigma)\mathscr{R}_{K,\sigma}(u)\\ &\quad - \alpha\sum_{\sigma\in\mathscr{E}_{\mathrm{ext}}} r_\sigma\,(e_{\sigma^+} - e_\sigma).\end{aligned} \tag{26}$$

This with the Cauchy Schwarz inequality and estimates (21)–(22) implies that

$$\begin{aligned}C_4^2|e_{\mathscr{D}}|_{\mathscr{X}}^2 + \frac{\alpha}{2}\sum_{\sigma\in\mathscr{E}_{\mathrm{ext}}}(e_\sigma - e_{\sigma^-})^2 &\leq C_2 h_{\mathscr{D}}\|u\|_{\mathscr{C}^2(\overline{\Omega})}|e_{\mathscr{D}}|_{\mathscr{X}}\\ &\quad + C_3\alpha\|u\|_{\mathscr{C}^1(\overline{\Omega})}\left(\sum_{\sigma\in\mathscr{E}_{\mathrm{ext}}}(\mathrm{m}(\sigma))^2\right)^{\frac{1}{2}}\left(\sum_{\sigma\in\mathscr{E}_{\mathrm{ext}}}(e_\sigma - e_{\sigma^-})^2\right)^{\frac{1}{2}}.\end{aligned} \tag{27}$$

Thanks to the use of inequality $a \leq \sqrt{a^2+b^2}$ for all real a and b, and the facts that $h_{\mathscr{D}} \leq \operatorname{diam}(\Omega)$, $\sum_{\sigma\in\mathscr{E}_{\rm ext}} \mathrm{m}(\sigma) = \mathrm{m}(\partial\Omega)$ and $\mathrm{m}(\sigma) \leq h_{\mathscr{D}}$, we are able to justify that the right hand side of (27) is bounded above by $C_6\sqrt{h_{\mathscr{D}}}\,|u|_{\mathscr{C}^2(\overline{\Omega})}$ $\left(C_4^2|e_{\mathscr{D}}|_{\mathscr{X}}^2 + \frac{\alpha}{2}\sum_{\sigma\in\mathscr{E}_{\rm ext}}(e_\sigma - e_{\sigma^-})^2\right)^{\frac{1}{2}}$ which together with (27) implies that

$$C_4|e_{\mathscr{D}}|_{\mathscr{X}} \leq C_6\sqrt{h_{\mathscr{D}}}\,|u|_{\mathscr{C}^2(\overline{\Omega})}. \tag{28}$$

This with the stability (25), the triangle inequality, and the consistency result [4, Lemma 4.4, p. 1029] yields the desired estimate (14) of Theorem 2. ■

As consequences of estimate (14) of Theorem 2, we derive the following $\mathbb{L}^2$–error estimate:

Corollary 1 ($\mathbb{L}^2$–error estimate for scheme (12)–(13)) *Under the same assumptions of Theorem 2, let u be the solution of problem (1)–(2) in the sense of the weak formulation (6)–(7) and $u_{\mathscr{D}}$ be the finite solution of scheme (12)–(13). Then, the following $\mathbb{L}^2$–error estimate holds:*

$$\|\mathscr{P}_{\mathscr{M}}u - \Pi_{\mathscr{M}}u_{\mathscr{D}}\|_{\mathbb{L}^2(\Omega)} \leq C_7\sqrt{h_{\mathscr{D}}}\,|u|_{\mathscr{C}^2(\overline{\Omega})}. \tag{29}$$

Proof Let $C_{\mathscr{M}} \in \mathbb{R}$ such that

$$\sum_{K\in\mathscr{M}} \mathrm{m}(K)\bar{u}(\mathbf{x}_K) = 0, \tag{30}$$

where $\bar{u} = u + C_{\mathscr{M}}$.
Error estimate (28) leads to (recall that $e_{\mathscr{D}}$ is given by (20))

$$C_4|\mathscr{P}_{\mathscr{D}}\bar{u} - u_{\mathscr{D}}|_{\mathscr{X}} \leq C_6\sqrt{h_{\mathscr{D}}}\,|u|_{\mathscr{C}^2(\overline{\Omega})}. \tag{31}$$

We need to use the following discrete H^1-seminorm, see [3, Definition 10.2]: For $u \in \mathscr{X}_{\mathscr{D}}$

$$|\Pi_{\mathscr{M}}u|_{1,2,\mathscr{M})} = \left(\sum_{\mathscr{M}_\sigma=\{K,L\}} \frac{\mathrm{m}(\sigma)}{d_\sigma}(u_K - u_L)^2\right)^{\frac{1}{2}}. \tag{32}$$

Using the Cauchy Schwarz inequality, we get, for all $\sigma \in \mathscr{E}_{\rm int}$ with $\mathscr{M}_\sigma = \{K, L\}$

$$\frac{(u_K-u_L)^2}{d_\sigma} \leq \frac{(u_K-u_\sigma)^2}{d_{K,\sigma}} + \frac{(u_\sigma-u_K)^2}{d_{L,\sigma}}, \ \forall u \in \mathscr{X}_{\mathscr{D}}. \tag{33}$$

Therefore $|\Pi_{\mathscr{M}}u|_{1,2,\mathscr{M}} \leq |u|_{\mathscr{X}}$, $\forall u \in \mathscr{X}_{\mathscr{D}}$. This with error estimate (31) leads to

$$C_4|\mathscr{P}_{\mathscr{M}}\bar{u} - \Pi_{\mathscr{M}}u_{\mathscr{D}}|_{1,2,\mathscr{M}} \leq C_6\sqrt{h_{\mathscr{D}}}\,|u|_{\mathscr{C}^2(\overline{\Omega})}. \tag{34}$$

This with the discrete mean Poincaré inequality of [3, Lemma 10.2, p. 796] (its proof remains valid here since the expressions of the norms involved in [3, Lemma 10.2] remain the same when considering the meshes of Definition 1) and the fact that $\int_{\Omega} (\mathscr{P}_{\mathscr{M}} \bar{u} - \Pi_{\mathscr{M}} u_{\mathscr{D}})(\mathbf{x}) d\mathbf{x} = 0$ implies that

$$\| \mathscr{P}_{\mathscr{M}} \bar{u} - \Pi_{\mathscr{M}} u_{\mathscr{D}} \|_{\mathbb{L}^2(\Omega)} \leq C_8 \sqrt{h_{\mathscr{D}}} \| u \|_{\mathscr{C}^2(\overline{\Omega})}. \tag{35}$$

On another hand, using the fact that $\int_{\Omega} u(\mathbf{x}) d\mathbf{x} = 0$, Eq. (30), and a Taylor expansion yields that

$$\begin{aligned} \mathrm{m}(\Omega) C_{\mathscr{M}} &= \int_{\Omega} (\bar{u}(\mathbf{x}) - u(\mathbf{x}))\, d\mathbf{x} = \int_{\Omega} \bar{u}(\mathbf{x})\, d\mathbf{x} \\ &= \sum_{K \in \mathscr{M}} \int_{K} (\bar{u}(\mathbf{x}_K) + \nabla \bar{u}(\psi(x)) \cdot (\mathbf{x} - \mathbf{x}_K))\, d\mathbf{x} = \sum_{K \in \mathscr{M}} \int_{K} \nabla \bar{u}(\psi(x)) \cdot (\mathbf{x} - \mathbf{x}_K)\, d\mathbf{x}. \end{aligned}$$

This with the fact that $\sum_{K \in \mathscr{M}} \mathrm{m}(K) = \mathrm{m}(\Omega)$ implies that $C_{\mathscr{M}} \leq C_9 h_{\mathscr{D}} \| u \|_{\mathscr{C}^1(\overline{\Omega})}$. This with the fact that $\bar{u} = u + C_{\mathscr{M}}$, estimate (35), and the triangle inequality yields the desired estimate (29) of Corollary 1. ■

References

1. Bradji, A., Gallouët, T.: Error estimate for finite volume approximate solutions of some oblique derivative boundary problems. Int. J. Finite **3**(2), 35 (electronic) (2006)
2. Bradji, A.: Improved convergence order in finite volume and finite element methods. Thesis, University of Marseille (2005)
3. Eymard, R., Gallouët, T., Herbin, R.: Finite volume methods. In: Ciarlet P.G., Lions J.L. (eds.) Handbook of Numerical Analysis, vol. VII, pp. 723–1020 (2000)
4. Eymard, R., Gallouët, T., Herbin, R.: Discretization of heterogeneous and anisotropic diffusion problems on general nonconforming meshes SUSHI: a scheme using stabilization and hybrid interfaces. IMA J. Numer. Anal. **30**(4), 1009–1043 (2010)
5. Mehats, F.: Convergence of a numerical scheme for a nonlinear oblique derivative boundary value problem. M2AN Math. Model. Numer. Anal. **36**(6), 1111–1132 (2002)

Optimal and Pressure-Independent L^2 Velocity Error Estimates for a Modified Crouzeix-Raviart Element with BDM Reconstructions

Christian Brennecke, Alexander Linke, Christian Merdon and Joachim Schöberl

Abstract Nearly all inf-sup stable mixed finite elements for the incompressible Stokes equations relax the divergence constraint. The price to pay is that a-priori estimates for the velocity error become pressure-dependent, while *divergence-free* mixed finite elements deliver *pressure-independent* estimates. A recently introduced new variational crime using lowest-order Raviart-Thomas velocity reconstructions delivers a much more robust modified Crouzeix-Raviart element, obeying an optimal *pressure-independent* discrete H^1 velocity estimate. Refining this approach, a more sophisticated variational crime employing the lowest-order BDM element is proposed, which also allows proving an optimal pressure-independent L^2 *velocity error*. Numerical examples confirm the analytical results.

1 Introduction

The success of classical mixed finite elements for the incompressible Navier-Stokes equations relies heavily on the relaxation of the divergence constraint, enabling the construction of large classes of inf-sup stable finite element pairs (X_h, Q_h) for the

C. Brennecke
Eidgenössische Technische Hochschule Zürich, Departement Mathematik, Rämistr. 101, 8092 Zürich, Switzerland
e-mail: cbrenne@student.ethz.ch

A. Linke (✉) · C. Merdon
Weierstrass Institute for Applied Analysis and Stochastics, Mohrenstr. 39, 10117 Berlin, Germany
e-mail: alexander.linke@wias-berlin.de

C. Merdon
e-mail: christian.merdon@wias-berlin.de

J. Schöberl
TU Wien, Institut für Analysis und Scientific Computing, Wiedner Hauptstr. 8-10/101, 1040 Wien, Austria
e-mail: joachim.schoeberl@tuwien.ac.at

J. Fuhrmann et al. (eds.), *Finite Volumes for Complex Applications VII - Methods and Theoretical Aspects,* Springer Proceedings in Mathematics & Statistics 77,
DOI: 10.1007/978-3-319-05684-5_14,

approximation of velocity and pressure [3]. Unfortunately, this relaxation is not for free. Looking at the simplest case, the incompressible Stokes equations

$$-\nu\Delta\mathbf{u} + \nabla p = \mathbf{f}, \quad \nabla\cdot\mathbf{u} = 0, \tag{1}$$

the classical a-priori error estimate for the velocity error [3] in the discrete H^1 norm reads (for homogeneous Dirichlet boundary conditions) as

$$\|\mathbf{u}-\mathbf{u}_h\|_{1,h} \leq C_1 \inf_{\mathbf{w}_h\in X_h} \|\mathbf{u}-\mathbf{w}_h\|_{1,h} + \frac{C_2}{\nu} \inf_{q_h\in Q_h} \|p-q_h\|_0. \tag{2}$$

But *divergence-free* mixed finite elements like the Scott-Vogelius finite element deliver the *pressure-independent* velocity error estimate [3]

$$\|\mathbf{u}-\mathbf{u}_h\|_{1,h} \leq C_3 \inf_{\mathbf{w}_h\in X_h} \|\mathbf{u}-\mathbf{w}_h\|_{1,h}. \tag{3}$$

In many physical situations, where the pressure is comparably small w.r.t. the velocity, the appearance of the pressure in the estimate (2) is indeed negligible. But in some situations, the situation is different and mixed methods suffer from so-called *poor mass conservation*. The easiest example, where mixed methods reveal their lack of robustness, is the *no-flow example*, where one prescribes $\mathbf{f} = \nabla\phi$ as the forcing in (1). Assuming homogeneous Dirichlet boundary conditions, $(\mathbf{u}, p) = (\mathbf{0}, \phi)$ uniquely solves (1). Obviously, in this example the pressure $p = \phi$ is *not small* compared to the velocity $\mathbf{u} = \mathbf{0}$. According to (3), divergence-free methods deliver indeed a discrete velocity $\mathbf{u}_h = \mathbf{0}$, while mixed methods with a relaxed divergence constraint will have a velocity error, which can be arbitrarily large, only dependent on ϕ, ν, and the applied mixed method. Since the continuous velocity $\mathbf{u} = \mathbf{0}$ lies in the approximation space of the discrete method, mixed methods indeed suffer from a *stability problem*.

Recently in [4] a new approach has been proposed, in order to avoid *poor mass conservation* completely. The approach is based on the observation that the proper source of the numerical instability is a *poor momentum balance*, where irrotational and divergence-free forces interact in a non-physical manner. Due to their L^2-orthogonality, divergence-free and irrotational forces are balanced separately in the continuous equations. But due to the relaxation of the divergence constraint in mixed methods, this separation fails in mixed methods, in general.

In [4] it is shown how to reestablish L^2-orthogonality between discretely divergence-free and irrotational vector fields modifying the nonconforming Crouzeix-Raviart element [2] by a variational crime. Here, a velocity reconstruction operator maps *discretely divergence-free test functions* onto *divergence-free vector fields*, by employing the lowest-order H(div)-conforming Raviart-Thomas element in the right-hand side of the incompressible Stokes equations. Replacing the test functions by reconstructions, introduces an additional consistency error, but improves the robustness of the Crouzeix-Raviart element, since one can prove a *pressure-independent*,

a-priori discrete H^1 velocity error estimate [4]. Unfortunately, in [4] the author did not succeed in proving also an optimal a-priori L^2 error estimate for the velocity, although numerical experiments show that such an estimate probably holds.

This contribution introduces a more sophisticated velocity reconstruction operator onto lowest-order BDM finite elements [1] with better interpolation properties.

2 Continuous and Discrete Setting

This section explains the continuous and the discrete setting for the model problem under consideration.

2.1 Continuous Setting

Given the Sobolev spaces $V := H_0^1(\Omega)^d$, $H(\mathrm{div};\Omega)$, and $Q := L_0^2(\Omega)$, the weak solution $(\mathbf{u}, p) \in V \times Q$ of the continuous steady incompressible Stokes equations satisfies

$$a(\mathbf{u}, \mathbf{v}) = l(\mathbf{v}) \quad \text{and} \quad b(\mathbf{u}, q) = 0 \qquad \text{for all } (\mathbf{v}, q) \in V \times Q \tag{4}$$

with the multilinear forms defined by

$$\begin{aligned} a : V \times V \to \mathbb{R}, \quad a(\mathbf{u}, \mathbf{v}) &:= \nu \int_\Omega \nabla \mathbf{u} : \nabla \mathbf{v} \, d\mathbf{x}, \\ b : V \times Q \to \mathbb{R}, \quad b(\mathbf{u}, q) &:= -\int_\Omega q \nabla \cdot \mathbf{u} \, d\mathbf{x}, \\ l : V \to \mathbb{R}, \quad l(\mathbf{v}) &:= \int_\Omega \mathbf{f} \cdot \mathbf{v} \, d\mathbf{x}. \end{aligned}$$

Within the set of weakly differentiable, divergence-free functions

$$V_0 := \{\mathbf{v} \in V : \nabla \cdot \mathbf{v} = 0\}, \tag{5}$$

the saddle point problem (4) reduces to the elliptic problem: seek $\mathbf{u} \in V_0$ such that

$$a(\mathbf{u}, \mathbf{v}) = l(\mathbf{v}) \qquad \text{for all } \mathbf{v} \in V_0. \tag{6}$$

2.2 Notation

In the following, $\mathscr{T}$ denotes a regular triangulation of the domain Ω into triangles for $d = 2$ or tetrahedra for $d = 3$. For any element $T \in \mathscr{T}$, $\mathrm{mid}(T)$ denotes the

barycenter of T. The set of all simplex faces, i.e., edges of triangles for $d = 2$ and faces of tetrahedra for $d = 3$, is denoted by $\mathscr{F}$. The subset $\mathscr{F}(\Omega)$ denotes the set of interior faces, while $\mathscr{F}(\partial\Omega)$ denotes the set of boundary faces along $\partial\Omega$. For any $F \in \mathscr{F}$, $\mathrm{mid}(F)$ denotes the barycenter of F and $\mathbf{n}_F$ abbreviates a face unit normal vector. The orientations of these normal vectors for the interior faces $F \in \mathscr{F}(\Omega)$ are arbitrary, but fixed. For boundary faces $F \in \mathscr{F}(\partial\Omega)$, the normal vectors $\mathbf{n}_F$ point outward of the domain Ω. For every simplex $T \in \mathscr{T}$, $\mathscr{F}(T)$ denotes the set of faces of this simplex and $\mathbf{n}_T$ denotes the outer normal of the simplex $T \in \mathscr{T}$. The function space of $P_k(\mathscr{T})$ contains piecewise polynomials of degree k with respect to $\mathscr{T}$. For a piecewise Sobolev function $\mathbf{v} \in H^1(\mathscr{T})^d$ and some face $F \in \mathscr{F}(\Omega)$, the notion $[\mathbf{v} \cdot \mathbf{n}_F]$ denotes the jump of the normal flux over F, while $\{\{\mathbf{v} \cdot \mathbf{n}_F\}\}$ denotes the average value of the normal flux over F. The space of Crouzeix-Raviart velocity trial functions is given by

$$\mathrm{CR}(\mathscr{T}) := \big\{\mathbf{v}_h \in P_1(\mathscr{T})^d : \text{for all } T \in \mathscr{T},\ [\mathbf{v}_h](\mathrm{mid}(F)) = \mathbf{0} \text{ for all } F \in \mathscr{F}(\Omega) \\ \& \ \mathbf{v}_h(\mathrm{mid}(F)) = \mathbf{0} \text{ for all } F \in \mathscr{F}(\partial\Omega)\big\}.$$

The pressure trial function space reads

$$Q(\mathscr{T}) := \left\{q_h \in P_0(\mathscr{T}) : \int_\Omega q_h \, d\mathbf{x} = 0\right\}.$$

The space of Brezzi-Douglas-Marini finite element functions reads

$$\mathrm{BDM}(\mathscr{T}) := \big\{\mathbf{v}_h \in P_1(\mathscr{T})^d : [\mathbf{v}_h \cdot \mathbf{n}_F] = 0 \text{ along all } F \in \mathscr{F}\big\}.$$

Furthermore, consider its subspace of lowest order Raviart-Thomas functions

$$\mathrm{RT}(\mathscr{T}) := \Big\{\mathbf{v}_h \in \mathrm{BDM}(\mathscr{T}) : \forall T \in \mathscr{T}\ \exists \mathbf{a}_T \in \mathbb{R}^d, b_T \in \mathbb{R},\ \mathbf{v}_h|_T(\mathbf{x}) = \mathbf{a}_T + b_T\mathbf{x}\Big\}.$$

The space $\mathrm{RT}(\mathscr{T})$ contains exactly the subset of functions with constant normal fluxes $\mathbf{v}_h \cdot \mathbf{n}_F \in P_0(F)$ on every face $F \in \mathscr{F}$. By that, any Raviart-Thomas function is uniquely defined by its face normal fluxes at the face barycenters. Note, that a Crouzeix-Raviart function $\mathbf{v}_h \in \mathrm{CR}(\mathscr{T})$ is, in general, discontinuous along element faces $F \in \mathscr{F}$ except at the face barycenters. Therefore, $\mathrm{CR}(\mathscr{T}) \not\subset H(\mathrm{div}; \Omega)$ and $\mathrm{CR}(\mathscr{T}) \not\subset V_0$. On the contrary, $\mathrm{RT}(\mathscr{T}) \subset \mathrm{BDM}(\mathscr{T}) \subset H(\mathrm{div}; \Omega)$, because the normal components of any $\mathbf{v}_h \in \mathrm{RT}(\mathscr{T})$ or $\mathbf{v}_h \in \mathrm{BDM}(\mathscr{T})$ are continuous.

The discrete setting employs the *broken gradient* $\nabla_h : V \oplus \mathrm{CR}(\mathscr{T}) \to L^2(\Omega)^{d\times d}$ and the *broken divergence* $\nabla_h \cdot (\cdot) : V \oplus \mathrm{CR}(\mathscr{T}) \to L^2(\Omega)$ in the sense that

$$(\nabla_h \mathbf{v}_h)|_T := \nabla(\mathbf{v}_h|_T), \quad (\nabla_h \cdot \mathbf{v}_h)|_T := \nabla \cdot (\mathbf{v}_h|_T) \quad \text{for all } T \in \mathscr{T}.$$

The discrete energy norm for $V \oplus \mathrm{CR}(\mathscr{T})$ reads $\|\mathbf{v}\|_{1,h} := \big(\int_\Omega \nabla_h \mathbf{v} : \nabla_h \mathbf{v} \, d\mathbf{x}\big)^{1/2}$.

2.3 Interpolation Operators

The usual Crouzeix-Raviart interpolation operator $\boldsymbol{\pi}^{\mathrm{CR}} : V \to \mathrm{CR}(\mathscr{T})$ is defined by

$$(\boldsymbol{\pi}^{\mathrm{CR}}\,\mathbf{v})(\mathrm{mid}(F)) = \frac{1}{|F|}\int_F \mathbf{v} ds \quad \text{for all } F \in \mathscr{F}\,.$$

The Raviart-Thomas interpolation operator $\boldsymbol{\pi}^{\mathrm{RT}} : V \oplus \mathrm{CR}(\mathscr{T}) \to \mathrm{RT}(\mathscr{T})$ is defined by

$$\mathbf{n}_F \cdot (\boldsymbol{\pi}^{\mathrm{RT}}\,\mathbf{v})(\mathrm{mid}(F)) = \frac{1}{|F|}\int_F \mathbf{v}\cdot\mathbf{n}_F ds \quad \text{for all } F \in \mathscr{F}\,.$$

Note that, due to continuity in the face barycenters, this is well-defined also for $\mathbf{v} \in \mathrm{CR}(\mathscr{T})$. Moreover, it holds the identity $\boldsymbol{\pi}^{\mathrm{RT}}\boldsymbol{\pi}^{\mathrm{CR}}\,\mathbf{v} = \boldsymbol{\pi}^{\mathrm{RT}}\,\mathbf{v}$ for any $\mathbf{v} \in V$.

The BDM interpolation operator $\boldsymbol{\pi}^{\mathrm{BDM}} : V \oplus \mathrm{CR}(\mathscr{T}) \to \mathrm{BDM}(\mathscr{T})$ on a face $F \in \mathscr{F}$ is defined such that, for all $p_h \in P_1(F)$,

$$\int_F (\boldsymbol{\pi}^{\mathrm{BDM}}\,\mathbf{v})\cdot\mathbf{n}_F\; p_h ds = \begin{cases} \int_F \{\{\mathbf{v}\cdot\mathbf{n}_\mathbf{F}\}\}\; p_h ds & \text{for all } F \in \mathscr{F}(\Omega) \\ \int_F (\boldsymbol{\pi}^{\mathrm{RT}}\,\mathbf{v})\cdot\mathbf{n}_F\; p_h ds & \text{for all } F \in \mathscr{F}(\partial\Omega). \end{cases}$$

However, at the domain boundary $\partial\Omega$ the BDM interpolation equals the RT interpolation to ensure that the BDM interpolation $\boldsymbol{\pi}^{\mathrm{BDM}}\,\mathbf{v}_h$ of functions $\mathbf{v}_h \in \mathrm{CR}(\mathscr{T})$ vanishes along the complete boundary $\partial\Omega$. With this, the boundary integral in the integration by parts formula,

$$\int_\Omega (\boldsymbol{\pi}^{\mathrm{BDM}}\,\mathbf{v}_h)\nabla p\, d\mathbf{x} = \int_\Omega \nabla\cdot(\boldsymbol{\pi}^{\mathrm{BDM}}\,\mathbf{v}_h)\, p\, d\mathbf{x} + \int_{\partial\Omega} (\boldsymbol{\pi}^{\mathrm{BDM}}\,\mathbf{v}_h)\cdot\mathbf{n}\; p ds,$$

vanishes and enables L^2-orthogonality of $\boldsymbol{\pi}^{\mathrm{BDM}}\,\mathbf{v}_h$ on gradients of all functions $p \in H^1(\Omega)$ for any discretely divergence-free $\mathbf{v}_h \in \mathrm{CR}(\mathscr{T})$. For any $\mathbf{v} \in V_0$, it immediately follows $\nabla\cdot\boldsymbol{\pi}^{\mathrm{BDM}}\,\mathbf{v} = 0$, $\nabla\cdot\boldsymbol{\pi}^{\mathrm{RT}}\,\mathbf{v} = 0$, and $\nabla_h\cdot\boldsymbol{\pi}^{\mathrm{CR}}\,\mathbf{v} = 0$ by Gauss' theorem. Furthermore, there are the well-known stability and approximation properties

$$\|\boldsymbol{\pi}^{\mathrm{CR}}\,\mathbf{v}\|_{1,h} \le \|\nabla\mathbf{v}\|_0, \qquad \text{for all } v \in V, \tag{7}$$

$$\|\mathbf{v} - \boldsymbol{\pi}^{\mathrm{CR}}\,\mathbf{v}\| \le Ch\|\mathbf{v} - \boldsymbol{\pi}^{\mathrm{CR}}\,\mathbf{v}\|_{1,h} \quad \text{for all } v \in V, \tag{8}$$

$$\|\mathbf{v} - \boldsymbol{\pi}^{\mathrm{CR}}\,\mathbf{v}\|_{1,h} \le Ch|\mathbf{v}|_2 \qquad \text{for all } v \in V \cap H^2(\Omega)^d, \tag{9}$$

$$\|\mathbf{v} - \boldsymbol{\pi}^{\mathrm{RT}}\,\mathbf{v}\|_0 \le Ch\|\mathbf{v}\|_{1,h} \qquad \text{for all } v \in V \oplus \mathrm{CR}(\mathscr{T}), \tag{10}$$

$$\|\mathbf{v} - \boldsymbol{\pi}^{\mathrm{BDM}}\,\mathbf{v}\|_0 \le Ch\|\mathbf{v}\|_{1,h} \qquad \text{for all } v \in V \oplus \mathrm{CR}(\mathscr{T}), \tag{11}$$

$$\|\mathbf{v} - \boldsymbol{\pi}^{\mathrm{BDM}}\,\mathbf{v}\|_0 \le Ch^2|\mathbf{v}|_2 \qquad \text{for all } v \in H^2(\Omega)^d \cap H_0^1(\Omega)^d, \tag{12}$$

where the generic constants C depend only on the shape of the simplices in the triangulation $\mathscr{T}$ but not on their size. Note, that the proofs of the estimates (10) and (11) are extendable to functions $\mathbf{v} \in \mathrm{CR}(\mathscr{T})$. Moreover, we emphasize that the interpolation operator $\boldsymbol{\pi}^{\mathrm{BDM}}$ does not suffer from a loss of convergence at the boundary, since we only apply it to H_0^1-functions in property (12).

2.4 The Finite Element Scheme with and Without Divergence-Free Reconstruction

The discrete weak formulation of the model problem employs the multilinear forms

$$\begin{aligned} a_h : (V \oplus \mathrm{CR}(\mathscr{T})) \times (V \oplus \mathrm{CR}(\mathscr{T})) \to \mathbb{R}, \quad a_h(\mathbf{u}_h, \mathbf{v}_h) &:= \nu \int_\Omega \nabla_h \mathbf{u}_h : \nabla_h \mathbf{v}_h \, d\mathbf{x}, \\ b_h : (V \oplus \mathrm{CR}(\mathscr{T})) \times Q \to \mathbb{R}, \quad b_h(\mathbf{u}_h, q_h) &:= -\int_\Omega q_h \nabla_h \cdot \mathbf{u}_h \, d\mathbf{x}, \\ l_h : (V \oplus \mathrm{CR}(\mathscr{T})) \to \mathbb{R}, \quad l_h(\mathbf{v}_h) &:= \int_\Omega \mathbf{f} \cdot \mathbf{v}_h \, d\mathbf{x}. \end{aligned}$$

Given one of the three interpolation operators above $\boldsymbol{\pi}^{\mathrm{div}} \in \{\boldsymbol{\pi}^{\mathrm{CR}}, \boldsymbol{\pi}^{\mathrm{RT}}, \boldsymbol{\pi}^{\mathrm{BDM}}\}$, the discrete Stokes problem seeks $(\mathbf{u}_h, p_h) \in \mathrm{CR}(\mathscr{T}) \times Q(\mathscr{T})$ such that

$$\begin{aligned} a_h(\mathbf{u}_h, \mathbf{v}_h) + b_h(\mathbf{v}_h, p_h) &= l_h(\boldsymbol{\pi}^{\mathrm{div}} \mathbf{v}_h), \\ b_h(\mathbf{u}_h, q_h) &= 0 \qquad \text{for all } (\mathbf{v}_h, q_h) \in \mathrm{CR}(\mathscr{T}) \times Q(\mathscr{T}). \end{aligned} \tag{13}$$

The choice $\boldsymbol{\pi}^{\mathrm{div}} = \boldsymbol{\pi}^{\mathrm{CR}}$ leads to the classical Crouzeix-Raviart nonconforming finite element method in the spirit of [2], while the other two choices $\boldsymbol{\pi}^{\mathrm{div}} = \boldsymbol{\pi}^{\mathrm{RT}}$ or $\boldsymbol{\pi}^{\mathrm{div}} = \boldsymbol{\pi}^{\mathrm{BDM}}$ constitute another variational crime that maps discretely divergence-free test functions to divergence-free functions in $H(\mathrm{div}; \Omega)$. The benefits of these divergence-free reconstructions are discussed below. Like the continuous incompressible Stokes equations, also the discretization (13) can be formulated as an elliptic problem [3] within the space of discretely divergence-free functions

$$V_{0,h} := \{\mathbf{v}_h \in \mathrm{CR}(\mathscr{T}) : b_h(\mathbf{v}_h, q_h) = 0 \text{ for all } q_h \in Q(\mathscr{T})\}. \tag{14}$$

Then, $\mathbf{u}_h \in V_{0,h}$ is uniquely defined by

$$a_h(\mathbf{u}_h, \mathbf{v}_h) = l_h(\boldsymbol{\pi}^{\mathrm{div}} \mathbf{v}_h) \quad \text{for all } \mathbf{v}_h \in V_{0,h}. \tag{15}$$

3 Error Estimates

This section presents a-priori finite element error estimates for the modified Crouzeix-Raviart discretization of the incompressible Stokes equations (13). Due to the divergence-conforming reconstruction the scheme (13) allows for an error estimate of the discrete velocity that is independent of the pressure regularity.

Lemma 1 *For* $\boldsymbol{\pi}^{\mathrm{div}} \in \{\boldsymbol{\pi}^{\mathrm{RT}}, \boldsymbol{\pi}^{\mathrm{BDM}}\}$ *and any* $\mathbf{v} \in V \cap H^2(\Omega)$, $\mathbf{w}_h \in V \oplus \mathrm{CR}(\mathscr{T})$, *it holds*

$$\left| \int_\Omega \nabla \mathbf{v} : \nabla_h \mathbf{w}_h + \Delta \mathbf{v} \cdot \boldsymbol{\pi}^{\mathrm{div}} \mathbf{w}_h \, d\mathbf{x} \right| \leq Ch|\mathbf{v}|_2 \|\mathbf{w}_h\|_{1,h}.$$

The estimate of the consistency error is a corollary to Lemma 1.

Lemma 2 (Consistency error estimate) *Given the solution* $(\mathbf{u}, p) \in H^2(\Omega)^d \times H^1(\Omega)$ *of the continuous Stokes equations* (4) *and* $\boldsymbol{\pi}^{\mathrm{div}} \in \{\boldsymbol{\pi}^{\mathrm{RT}}, \boldsymbol{\pi}^{\mathrm{BDM}}\}$, *it holds*

$$\sup_{\mathbf{w}_h \in V_0 + V_{0,h}} \left| a_h(\mathbf{u}, \mathbf{w}_h) - l_h(\boldsymbol{\pi}^{\mathrm{div}} \mathbf{w}_h) \right| / \|\mathbf{w}_h\|_{1,h} \leq Ch\nu|\mathbf{u}|_2.$$

Theorem 1 *For a solution* $(\mathbf{u}, p) \in H^2(\Omega)^d \times H^1(\Omega)$ *of the continuous Stokes equations* (4) *and the discrete solution* $(\mathbf{u}_h, p_h)$ *of the scheme* (13) *with* $\boldsymbol{\pi}^{\mathrm{div}} \in \{\boldsymbol{\pi}^{\mathrm{RT}}, \boldsymbol{\pi}^{\mathrm{BDM}}\}$, *the following error estimates hold*

$$\begin{aligned} &(i) \quad \|\mathbf{u} - \mathbf{u}_h\|_{1,h} \leq Ch\,|\mathbf{u}|_2\,, \\ &(ii) \quad \|p - p_h\|_0 \leq C\,h\,(\nu\,|\mathbf{u}|_2 + |p|_1)\,. \end{aligned}$$

Lemma 3 *For a right-hand side* $\mathbf{g} \in L^2(\Omega)^d$, *consider* $\mathbf{u}_g \in V_0$ *and* $\mathbf{u}_{g,h} \in V_{0,h}$ *with*

$$\begin{aligned} a(\mathbf{u}_g, \mathbf{v}) &= (\mathbf{g}, \mathbf{v}) \quad \textit{for all } \mathbf{v} \in V_0, \\ a_h(\mathbf{u}_{g,h}, \mathbf{v}_h) &= (\mathbf{g}, \boldsymbol{\pi}^{\mathrm{div}} \mathbf{v}_h) \quad \textit{for all } \mathbf{v}_h \in V_{0,h}. \end{aligned}$$

Then, for the the error $\mathbf{e} := \mathbf{u} - \mathbf{u}_h$ *between* $\mathbf{u}$ *from* (4) *and* $\mathbf{u}_h$ *from* (13), *it holds*

$$\begin{aligned} \|\mathbf{e}\|_0 \leq \sup_{\mathbf{g} \in L^2(\Omega)^d, \|\mathbf{g}\|_0 = 1} \Big\{ &\nu \|\mathbf{e}\|_{1,h} \|\mathbf{u}_g - \mathbf{u}_{g,h}\|_{1,h} + \left| a_h(\mathbf{e}, \mathbf{u}_g) - (\mathbf{g}, \boldsymbol{\pi}^{\mathrm{div}} \mathbf{e}) \right| \\ &+ \left| a_h(\mathbf{u}, \mathbf{u}_g - \mathbf{u}_{g,h}) - (\mathbf{f}, \boldsymbol{\pi}^{\mathrm{div}}(\mathbf{u}_g - \mathbf{u}_{g,h})) \right| \\ &+ \left| (\mathbf{g}, \mathbf{e} - \boldsymbol{\pi}^{\mathrm{div}} \mathbf{e}) \right| + \left| (\mathbf{f}, \mathbf{u}_g - \boldsymbol{\pi}^{\mathrm{div}} \mathbf{u}_g) \right| \Big\}. \end{aligned}$$

Table 1 Energy error and L^2 error for the velocity and pressure

ndof→	10176	40488	162152	646376	2585272
$\|\mathbf{u}-\mathbf{u}_h\|_0$ $(\boldsymbol{\pi}^{\mathrm{CR}})$	1.462715e-02	3.714616e-03	9.311043e-04	2.346116e-04	5.889322e-05
$\|\mathbf{u}-\mathbf{u}_h\|_0$ $(\boldsymbol{\pi}^{\mathrm{RT}})$	5.738088e-05	1.468924e-05	3.655164e-06	9.201573e-07	2.299916e-07
$\|\mathbf{u}-\mathbf{u}_h\|_0$ $(\boldsymbol{\pi}^{\mathrm{BDM}})$	6.475907e-05	1.651350e-05	4.117682e-06	1.036546e-06	2.589664e-07
$\|\mathbf{u}-\mathbf{u}_h\|_{1,h}$ $(\boldsymbol{\pi}^{\mathrm{CR}})$	1.333391	6.688239e-01	3.349285e-01	1.682897e-01	8.432380e-02
$\|\mathbf{u}-\mathbf{u}_h\|_{1,h}$ $(\boldsymbol{\pi}^{\mathrm{RT}})$	6.189144e-03	3.115982e-03	1.556097e-03	7.801799e-04	3.899851e-04
$\|\mathbf{u}-\mathbf{u}_h\|_{1,h}$ $(\boldsymbol{\pi}^{\mathrm{BDM}})$	6.184352e-03	3.115428e-03	1.556023e-03	7.801701e-04	3.899841e-04
$\|p-p_h\|_0$ $(\boldsymbol{\pi}^{\mathrm{CR}})$	1.293413e-02	6.371610e-03	3.174234e-03	1.590767e-03	7.960010e-04
$\|p-p_h\|_0$ $(\boldsymbol{\pi}^{\mathrm{RT}})$	1.270086e-02	6.297825e-03	3.147287e-03	1.579164e-03	7.904408e-04
$\|p-p_h\|_0$ $(\boldsymbol{\pi}^{\mathrm{BDM}})$	1.270086e-02	6.297825e-03	3.147287e-03	1.579164e-03	7.904408e-04

For smooth data, all terms on the right-hand side except the last one are of quadratic order. The last term is only of quadratic order for $\boldsymbol{\pi}^{\mathrm{div}} = \boldsymbol{\pi}^{\mathrm{BDM}}$ which leads to the final theorem.

Theorem 2 *For a convex domain Ω the exact solution $(\mathbf{u}, p) \in H^2(\Omega)^d \times H^1(\Omega)$ of the continuous Stokes equations (4) and the discrete solution $(\mathbf{u}_h, p_h)$ of (13) for $\boldsymbol{\pi}^{\mathrm{div}} = \boldsymbol{\pi}^{\mathrm{BDM}}$ satisfy an optimal L^2 error estimate for the discrete velocity, i.e.,*

$$\|\mathbf{u}-\mathbf{u}_h\|_0 \leq C\, h^2\, |\mathbf{u}|_2 .$$

4 Numerical Results

The first benchmark prescribes the stream function $\xi = x^2(1-x)^2y^2(1-y)^2$ with $\mathbf{u} = \mathrm{rot}\xi \in P_7(\Omega)^2 \cap V$ and $p = x^3 + y^3 - 1/2$ on the unit square $\Omega = (0,1)^2$. For given viscosity ν, the volume force equals $\mathbf{f} := -\nu\Delta\mathbf{u} + \nabla p$.

Table 1 compares the results of the three methods under consideration for $\nu = 10^{-2}$. While the error in the pressure is only slightly smaller, the H^1 error in the velocity is more than two magnitudes smaller for the methods with a divergence-free reconstruction due to the influence of the $1/\nu\,|p|_1$ contribution in the classical velocity error estimate (2). In this smooth example, the convergence speed of the L^2-error in the velocity is optimal also for $\boldsymbol{\pi}^{\mathrm{RT}}$ and leads to similar results as $\boldsymbol{\pi}^{\mathrm{BDM}}$.

The second benchmark considers the Stokes problem for $p \equiv 0$ and

$$\mathbf{u}(x,y) = \mathrm{rot}(x^s \log(x) + y^s \log(y))/5 \in H^{s-1}(\Omega) \setminus H^s(\Omega)$$

on $\Omega = (0,1)^2$ with right-hand side $f \equiv -\Delta\mathbf{u}$ and $\nu = 1$.

Figure 1 shows the convergence history of the L^2 error for $s = 2$ for all three methods. The reconstruction with $\boldsymbol{\pi}^{\mathrm{BDM}}$ leads to better results and, more importantly, to a better convergence rate than the reconstruction with $\boldsymbol{\pi}^{\mathrm{RT}}$. Since $p \equiv 0$, the results

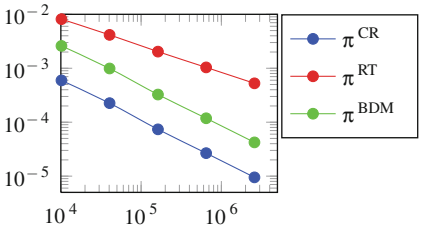


Fig. 1 Convergence history of the L^2 velocity error

of the unmodified Crouzeix-Raviart method for $\boldsymbol{\pi}^{\mathrm{div}} = \boldsymbol{\pi}^{\mathrm{CR}}$ are the best. The benefits of the reconstructions in case of nonzero pressure can be seen in the first example above.

Acknowledgments This research has been partially funded by the project "Macroscopic Modeling of Transport and Reaction Processes in Magnesium-Air-Batteries" (Grant 03EK3027D) under the research initiative "Energy storage" of the German Federal government.

References

1. Brezzi, F., Fortin, M.: Mixed and Hybrid Finite Element Methods. Springer, New York (1991)
2. Crouzeix, M., Raviart, P.A.: Conforming and nonconforming finite element methods for solving the stationary Stokes equations. I. Rev. Française Automat. Informat. Recherche Opérationnelle Sér. Rouge **7**(R-3), 33–75 (1973)
3. Girault, V., Raviart, P.A.: Finite Element Methods for Navier-Stokes Equations. Springer Series in Computational Mathematics, vol. 5. Springer, Berlin (1986)
4. Linke, A.: On the role of the Helmholtz decomposition in mixed methods for incompressible flows and a new variational crime. Comput. Methods Appl. Mech. Engrg. **268**, 782–800 (2014)

Conservative Finite Differences as an Alternative to Finite Volume for Compressible Flows

Jens Brouwer, Julius Reiss and Jörn Sesterhenn

Abstract Finite Volume schemes are the natural choice when simulating flows with shocks, since conservation is essential in the physics and as such in the simulation of this phenomenon. But finite difference schemes can be conservative as well. Conservation requires in such schemes a high internal consistency of the spatial and the temporal discretization. We present a skew-symmetric finite difference scheme, which is fully conservative due to its consistency, still easy to implement and numerically efficient. A variety of different flow configurations containing shocks and turbulence are presented.

1 Introduction

Finite volume (FV) schemes and conservative schemes are so strongly connected, that often these two terms are used synonymously. Indeed all FV schemes conserve the quantities of the underlying discretized flux-equations. The inverse statement however is not true: A set of discrete equation can be conservative, even if a form other than the flux form is the starting point of the scheme. This extra freedom can be beneficial to fulfill additional requirements. In the case presented here this requirement is a low dissipation of the scheme. This allows the direct numerical simulation (DNS) of turbulence and acoustics. Further, a high discretization order in space and time is important to keep the computational cost of simulations of large physical systems as low as possible.

J. Brouwer (✉) · J. Reiss · J. Sesterhenn
TU Berlin, ISTA, Mueller-Breslau Str. 8, 10623 Berlin, Germany
e-mail: jens.brouwer@tnt.tu-berlin.de

J. Reiss
e-mail: julius.reiss@tnt.tu-berlin.de

J. Sesterhenn
e-mail: joern.sesterhenn@tu-berlin.de

J. Fuhrmann et al. (eds.), *Finite Volumes for Complex Applications VII - Methods and Theoretical Aspects*, Springer Proceedings in Mathematics & Statistics 77,
DOI: 10.1007/978-3-319-05684-5_15,

We use a skew-symmetric finite difference scheme which meets the afore mentioned properties, which builds on similar concepts as [4, 12]. As computational variables for the equations of compressible flow the quantities $(\sqrt{\rho}, \sqrt{\rho}u_\alpha, p)$ are used. The conserved quantities of mass, $\sum \sqrt{\rho}^2$, momentum, $\sum \sqrt{\rho}(\sqrt{\rho}u_\alpha)$, and total energy, $\sum p/(\gamma-1) + (\sqrt{\rho}u_\alpha)^2/2$, are a consequence of the consistency of the discrete equations *and* a proper time stepping. They are not enforced by formulating the balance of these terms as in FV methods. Arbitrary order in space and time can be achieved. The scheme is computationally efficient in space as it builds on (non-upwind) finite-differences and point-wise multiplication. However, for full conservation an implicit time stepping scheme is needed. The resulting nonlinear system is solved by fix point iterations which is found to converge satisfactory well for time steps similar to those of an explicit scheme.

We present resolved calculations of a turbulent and transonic boundary layers, as well as a Richtmyer-Meshkov instability and one-dimensional shock problems.

2 Numerical Scheme

Here we provide a short overview of the skew-symmetric finite-difference scheme. A detailed derivation and discussion is out of scope of this paper but can be found in [10] for the spatial discretisation and [2] for the time stepping procedure.

Compressible flow is described by the Navier-Stokes equations which state the evolution of mass, momentum and energy:

$$\partial_t \rho + \partial_{x_\beta} \rho u_\beta = 0 \tag{1}$$

$$\partial_t \rho u_\alpha + \partial_{x_\beta}(\rho u_\beta u_\alpha) + \partial_{x_\alpha} p = \partial_{x_\beta} \tau_{\alpha\beta} \tag{2}$$

$$\partial_t \left(\rho \left[e + \frac{u_\alpha u_\alpha}{2}\right]\right) + \partial_{x_\beta} \left(\rho u_\beta \left[e + \frac{u_\alpha u_\alpha}{2} + \frac{p}{\rho}\right]\right) = \partial_{x_\alpha} u_\beta \tau_{\alpha\beta} + \partial_{x_\alpha} \phi_\alpha. \tag{3}$$

The ρ is the density, u_α is the $\alpha^{th} = 1, 2, 3$ velocity component. Pressure is p and $\tau_{\alpha\beta} = \mu(\partial_{x_\alpha} u_\beta + \partial_{x_\beta} u_\alpha) + (\mu_d - \mu 2/3)\delta_{\alpha\beta}\partial_{x_\gamma} u_\gamma$ is the Newtonian friction. The heat flux is given by $\phi_\alpha = \lambda \partial_{x_\alpha} T$ with the heat conductivity λ. Ideal gas with the internal energy $e = (p/\rho)/(\gamma-1)$ and adiabatic exponent γ is assumed in the following. Summing convention is assumed.

A pure rewriting of the momentum equations leads to the equations in skew-symmetric form

$$\partial_t \rho + \partial_{x_\beta} \rho u_\beta = 0 \tag{4}$$

$$\frac{1}{2}\left(\partial_t \rho \cdot + \rho \partial_t \cdot\right) u_\alpha + \frac{1}{2}\left(\partial_{x_\beta} u_\beta \rho \cdot + u_\beta \rho \partial_{x_\beta} \cdot\right) u_\alpha + \partial_{x_\alpha} p = \partial_{x_\beta} \tau_{\alpha\beta} \tag{5}$$

$$\begin{aligned} &\frac{1}{\gamma-1}\partial_t p + \frac{\gamma}{\gamma-1}\partial_{x_\beta}\left(u_\beta p\right) - u_\alpha \partial_{x_\alpha} p \\ &\quad = -u_\alpha \partial_{x_\beta} \tau_{\alpha\beta} \partial_{x_\beta} u_\alpha \tau_{\alpha\beta} + \partial_{x_\alpha} \phi_\alpha \end{aligned} \tag{6}$$

It is understood that the space and time derivatives in the first two terms of (5) act also on u right of the parentheses, which is marked by a „·". The momentum equation is called skew-symmetric, because the resulting spatial and temporal differentiation operators are skew-symmetric. The skew-symmetry of these operators leads to the analytical conservation of mass, momentum and energy. The kinetic energy was split from the total energy equation to arrive at an equation for the internal energy. To preserve the skew-symmetry in the discretization, skew symmetric derivative matrices $D^T = -D$ are used.

Morinishi's rewriting, [7], transforms the time derivative in the momentum equations (5) to $\frac{1}{2}\left(\partial_t \rho \cdot + \rho \partial_t \cdot\right) u_\alpha = \sqrt{\rho}\partial_t(\sqrt{\rho}u_\alpha)$. and leads to:

$$\sqrt{\rho}\partial_t(\sqrt{\rho}u_\alpha) + \frac{1}{2}\left(\partial_{x_\beta} u_\beta \rho \cdot + u_\beta \rho \partial_{x_\beta} \cdot\right) u_\alpha + \partial_{x_\alpha} p = \partial_{x_\beta}\tau_{\alpha\beta}. \tag{7}$$

The convective term $\frac{1}{2}\left(\partial_{x_\beta} u_\beta \rho \cdot + u_\beta \rho \partial_{x_\beta} \cdot\right)$ becomes a skew symmetric matrix $D^{\mathbf{u}}$, if discretized appropriately. By multiplying it by u_α^T

$$u_\alpha^T \sqrt{\rho}\partial_t(\sqrt{\rho}u_\alpha) + u_\alpha^T D^{\mathbf{u}} u_\alpha + u_\alpha^T \partial_{x_\alpha} p = u_\alpha^T \partial_{x_\beta}\tau_{\alpha\beta} \tag{8}$$

the change of kinetic energy is derived. Skew-symmetry implies $u_\alpha^T D^{\mathbf{u}} u_\alpha = 0$, thus

$$\frac{1}{2}\partial_t(\sqrt{\rho}u_\alpha)^2 = -u_\alpha^T \partial_{x_\alpha} p + u_\alpha^T \partial_{x_\beta}\tau_{\alpha\beta}.$$

The transport term conserves the kinetic energy; the kinetic energy is changed by pressure work and friction alone, as in the analytical theory, but in contrast to standard schemes. Now, also the unusual appearance of $\sqrt{\rho}$ instead of ρ in the momentum equation can be understood. It is the quadratic splitting of the kinetic energy. The terms $u_\alpha \partial_{x_\alpha} p - u_\alpha \partial_{x_\beta}\tau_{\alpha\beta}$ in Eq. (6) balance the change of kinetic energy by an according change of the internal energy, so that total energy is conserved. This structure carries over to the discrete case. Momentum conservation can be derived in a similar manner.

The method can be easily applied to transformed, structured grids, meaning grids generated by C^1 mappings of the unit cube. The conservation properties are strictly fulfilled as before. The resulting equations are

$$\begin{aligned}
&J 2\sqrt{\rho}\partial_t\sqrt{\rho} + \partial_{\xi^\beta}\tilde{u}_\beta \rho = 0\\
&J\sqrt{\rho}\partial_t(\sqrt{\rho}u_\alpha)u_\alpha + \frac{1}{2}\left(\partial_{\xi^\beta}\tilde{u}_\beta \rho \cdot + \tilde{u}_\beta \rho \partial_{\xi^\beta}\cdot\right) u_\alpha + J\partial_{x_\alpha} p = \partial_{\xi_\beta}\tilde{\tau}_{\alpha\beta}\\
&J\frac{1}{\gamma-1}\partial_t p + \frac{\gamma}{\gamma-1}\partial_{\xi^\beta}\left(\tilde{u}_\beta p\right) - J u_\alpha \partial_{x_\alpha} p\\
&\quad = -u_\beta \partial_{\xi^\alpha}\tilde{\tau}_{\alpha\beta} + \partial_{\xi^\alpha} u_\beta \tilde{\tau}_{\alpha\beta} + \partial_{\xi^\alpha}\tilde{\phi}_\alpha.
\end{aligned}$$

The effective velocities are defined to include the metric factors $\tilde{u}_{\gamma_1} = (\mathbf{e}_{\gamma_2} \times \mathbf{e}_{\gamma_3})\mathbf{u}$, γ_1 cyclic. The local basis vectors are defined as $\mathbf{e}_\alpha = \partial_{\xi^\alpha}\mathbf{r}$, with $r = (x, y, z)^T$. The Jacobian is $J = (\mathbf{e}_1 \times \mathbf{e}_2) \cdot \mathbf{e}_3$.

The discretization is done in a straight forward manner with the variables $(\sqrt{\rho}, \sqrt{\rho}u_\alpha, p)$. All derivatives are replaced by skew-symmetric derivative matrices (i.e. symmetric stencils). At boundaries the summation by parts property is assumed for which explicit derivatives constructed by Strand are used [11]. Details on the boundary treatment can be found in [10]. In addition, the use of SBP matrices allows the implementaion of an effevtive multiblock decomposition of the domain.

Time integration

The conserved quantities are (partly) quadratic forms of the discretization variables. Quadratic quantities are in general not conserved. Runge-Kutta schemes conserve quadratic invariants when the coefficients of their Butcher table fulfill the condition

$$b_i a_{ij} + b_j a_{ji} = b_i b_j. \tag{9}$$

This restrictive requirement is fulfilled by all Gauss-collocation methods, a family of s-stage implicit RK schemes of order $2s$. Time integration is done by the two stage, fourth order method:

$$\begin{array}{c|cc} \frac{1}{2} - \frac{\sqrt{3}}{6} & \frac{1}{4} & \frac{1}{4} - \frac{\sqrt{3}}{6} \\ \frac{1}{2} + \frac{\sqrt{3}}{6} & \frac{1}{4} + \frac{\sqrt{3}}{6} & \frac{1}{4} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}$$

These methods lead to full conservation, which is discussed in [3]. It is found that a fix point iteration works surprisingly well for moderate Δt. It is further observed, that the conservation converges quicker than one might estimate form the convergence of the full solution.

3 Numerical Examples

In this section we present four numerical examples to show the applicability of our method to physical flow situations containing small scale turbulence and shocks. Therefore we show computations of a classical shock-tube test case, a turbulent boundary layer, a developing Richtmyer-Meshkov instability and an instationary shock-wave/boundary-layer interaction (*SBLI*). All simulations use the previously described skew-symmetric finite difference scheme. The implementation is in FORTRAN and parallelized using MPI directives. Spatial discretization is done using 6th order central differences with SBP properties. Temporal discretizations is achieved by the implicit 4th order Gauss collocation method with one small exception. Due to the higher computational effort of the implicit scheme, the SBLI simulation is advanced in time until initial transients are gone using an explicit Runge-Kutta scheme of

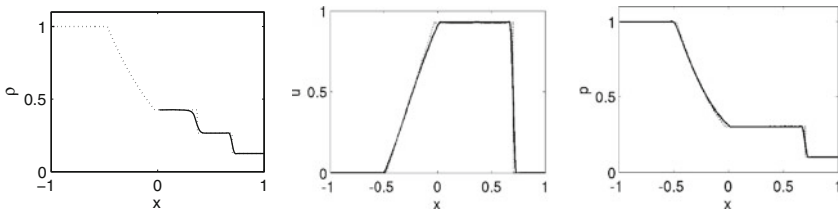


Fig. 1 Sod's test case at $t = 0.4$ using the Shock filter of Bogey et al. [1]. Good agreement is found. The Shock speed matches the analytical solution (*dotted line*)

fourth order. Once a statistically steady regime is reached, integration using the fully conservative Gauss-collocation scheme is resumed.

Sod's test case

The first test case is the classical Sod's shock-tube problem for the Euler equations. Starting conditions are $q^l = (1, 0, 1)$ and $q^l = (0.125, 0, 0.1)$ where $q = (\rho, u, p)$. The problem is discretized using 201 points and Bogeys conservative shockfilter is applied, see [1] for details. Figure 1 shows a comparison between the numerical and analytical solution, displaying agreement of propagation speed with the analytical solution, as expected for a conservative scheme. Only the contact discontinuity shows slightly higher damping than needed. The filter method is independent of the base scheme and can be easily modified to improve this.

DNS of a turbulent boundary layer at $Re \approx 5000$

To show the validity of the skew-symmetric finite-difference approach to small scale turbulence a direct numerical simulation of a turbulent boundary layer is shown. Reynolds number $Re_{\delta_{in}} = 4736$ and free-stream Mach number $M = 0.8$, where δ_{in} is the 99 % boundary layer thickness. The computational domain of dimensions $[106\delta_{in} \times 8\delta_{in} \times 9\delta_{in}]$ is resolved using roughly 80 million grid points. This resolution is chosen so that the grid spacing at the wall satisfies a dimensionless wall distance of $\Delta y^+ < 1$. Throughout the domain and the average Δy^+ at the boundary layer edge is not larger than 7. The turbulent inlet conditions are enforced using a recycling/rescaling method as introduced by Lund [6], and modified by Pirozzoli [9]. Results are in good agreement with reference computations of Pirozzoli et al. [8]. Figure 2 depicts streamwise velocity in a wall-parallel plane where the formation of characteristic streak structures is visible.

2-Dimensional Richtmyer-Meshkov instability

A Richtmyer-Meshkov instability is an instability mechanism that develops when an interface between fluids is impulsivly accelerated by a passing shockwave. In our simulation the fluid-fluid interface is modeled by a discontinous jump in density. The shock Mach number of the accelerating shockwave is $M_s = 1.5$. The two-dimensional domain is discretized with $[4096 \times 2048]$ gridpoints and the afore mentioned conservative shock-filter by Bogey et al. is used. Figure 3 shows the

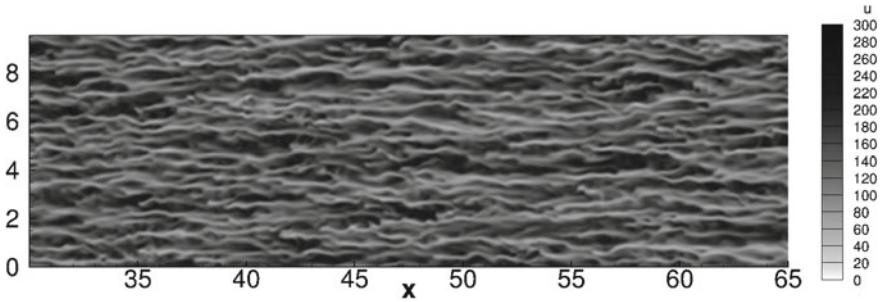


Fig. 2 Contours of instantaneous streamwise velocity in a xz-plane located at $y^+ \approx 10$

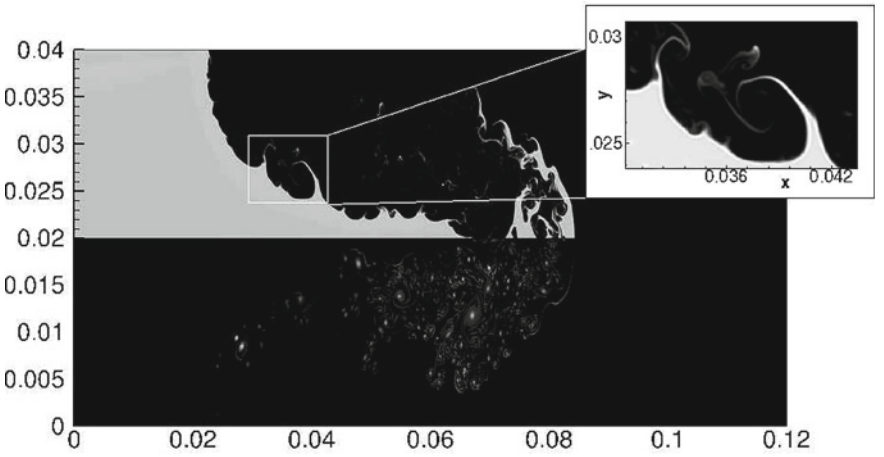


Fig. 3 Contour plot of a 2-dimensional Richtmyer-Meshkov instability; the *top* half of the domain depicts density while vorticity is shown in the *lower* half

mushroom-like growth out of the fluid interface. In the top half of the figure density contours are plotted while the lower half depicts vorticity. Due to the minimal dissipation of the skew-symmetric scheme, many secondary and even tertiary instabilities can be observed. The prime examples being the Kelvin-Helmholtz instabilities that form at the shear-layer between the two fluids. The vorticity plot reveals the complex turbulent flow field in the vicinity of the large scale structure that drives the creation of many of the smaller instabilities.

Shock-wave/boundary-layer Interaction

Shock-wave/boundary-layer interactions can occur in many important engineering applications. A prominent example is transonic flow over an airfoil, as the flow is accelerated over the airfoil, a super-sonic pocket forms that is terminated by a shock. The strong pressure gradient leads to the separation of the boundary-layer behind the shock and a recirculation bubble forms. Under certain conditions the shock can

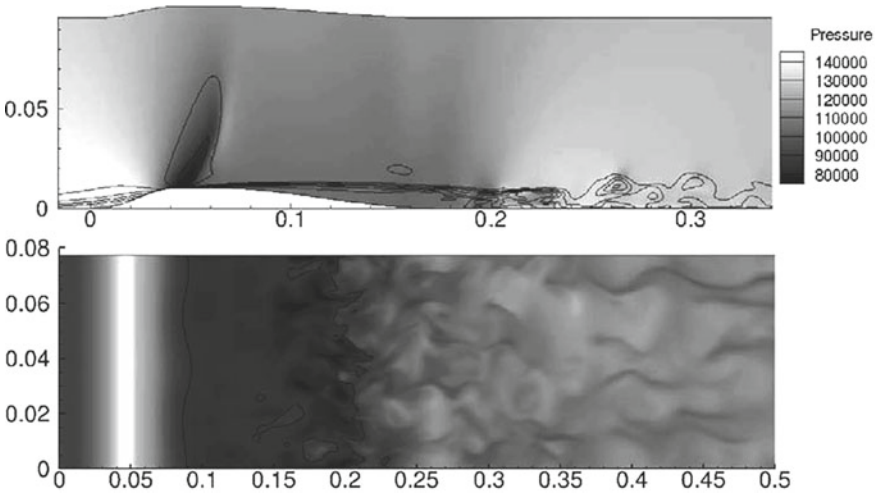


Fig. 4 *Top*: Snapshot of the transonic flow over a bump. Pressure contours and iso-lines of stream-wise velocity behind the bump are displayed. *Bottom*: Instantaneous streamwise velocity contours in wall-parallel plane at $y^+ \approx 25$. The current location of the recirculation bubble is marked by the $u = 0$ iso-line

exhibit large scale movements which has disastrous influence on all aerodynamic quantities. For a comprehensive review of the different forms of transsonic SBLI see e.g. [5].

The phenomenon of Shock-wave/boundary-layer interaction was one of the main motivations for the development of the conservative finite-difference scheme, as the simulation has to resolve the small turbulent scales in the boundary layer as well as to handle the shock movements. The simulations shown below are preliminary studies of SBLI occurring due to transonic flow over a bump. A laminar boundary layer is impinging on a small bump, a shock forms over the bump and a recirculation bubble that exhibits small breathing motions forms behind the interaction. Past the interaction zone the boundary layer begins its transition to turbulence. The size of the computational domain is $[65 \times 20 \times 12]$ measured in inlet boundary layer thicknesses δ_{in} and is resolved using approximately 20 million grid points. The maximum height of the bump is $1.2\delta_{in}$ while its length is $21.9\delta_{in}$. Shown below are snapshots of the instantaneous velocity and pressure fields. In Fig. 4 the geometry of the case is visible. The shock is visible both in the pressure fields and the contourlines of the stream-wise velocity. The recirculation bubble in the snapshot can be seen extending up to $x \approx 0.2$. In the lower panel of the figure the stream-wise velocity in a wall-parallel plane is shown. Again, the position of the shock at $x = 0.05$ and the length of the recirculation bubble can be seen. In addition the transition of the laminar flow field to turbulence through the interaction zone is visible.

4 Conclusions

We presented a fully conservative finite-difference scheme for the compressible Navier-Stokes equations on arbitrarily distorted, structured grids. In addition to its energy preserving nature the scheme introduces no artificial dissipation. The scheme is shown to be applicable to physical situations containing shocks and small scale turbulence while being easy to implement. This makes the skew-symmetric finite difference discretization a worthy alternative to Finite Volume methods in the context of large and small scale simulations of compressible flow.

References

1. Bogey, C., De Cacqueray, N., Bailly, C.: A shock-capturing methodology based on adaptative spatial filtering for high-order non-linear computations. JCP **228**(5), 1447–1465 (2009)
2. Brouwer, J., Reiss, J., Sesterhenn, J.: Fully conservative finite-difference schemes of arbitrary order for compressible flow. AIP Conf. Proc. **1479**(1), 2290–2293 (2012)
3. Brouwer, J., Reiss, J., Sesterhenn, J.: Conservative time integrators of arbitrary order for finite-difference discretization of compressible flow (2013). Submitted to Computers & Fluids
4. Kok, J.: A high-order low-dispersion symmetry-preserving finite-volume method for compressible flow on curvilinear grids. JCP **228**(18), 6811 (2009)
5. Lee, B.: Self-sustained shock oscillations on airfoils at transonic speeds. Prog. Aerosp. Sci. **37**(2), 147–196 (2001)
6. Lund, T.: Generation of turbulent inflow data for spatially-developing boundary layer simulations. JCP **140**, 233–258 (1998)
7. Morinishi, Y.: Forms of convecction and quadratic conservative finite difference schemes for low mach number compressible flow simulations. Trans. Jap. Soc. Mech. Eng. B (2007)
8. Pirozzoli, S., Bernardini, M.: Turbulence in supersonic boundary layers at moderate reynolds number. J. Fluid Mech. **688**, 120 (2011)
9. Pirozzoli, S., Bernardini, M., Grasso, F.: Direct numerical simulation of transonic shock/ boundary layer interaction under conditions of incipient separation. JFM **657**, 361 (2010)
10. Reiss, J., Sesterhenn, J.: A conservative, skew-symmetric finite difference scheme for the compressible navier-stokes equations. Accepted by Computers and Fluids
11. Strand, B.: Summation by parts for finite difference approximations for d/dx. JCP **110**, 47 (1994)
12. Verstappen, R., Veldman, A.: Symmetry-preserving discretization of turbulent flow. JCP **187**(1), 343 (2003)

FV Upwind Stabilization of FE Discretizations for Advection–Diffusion Problems

Fabian Brunner, Florian Frank and Peter Knabner

Abstract We apply a novel upwind stabilization of a mixed hybrid finite element method of lowest order to advection–diffusion problems with dominant advection and compare it with a finite element scheme stabilized by finite volume upwinding. Both schemes are locally mass conservative and employ an upwind-weighting formula in the discretization of the advective term. Numerical experiments indicate that the upwind-mixed method is competitive with the finite volume method. It prevents the appearance of spurious oscillations and produces nonnegative solutions for strongly advection-dominated problems, while the amount of artificial diffusion is lower than that of the finite volume method. This makes the method attractive for applications in which too much numerical diffusion is critical and may lead to false predictions; e.g., if highly nonlinear reactive processes take place only in thin interaction regions.

1 Introduction

In this article, we consider the linear advection–diffusion equation

$$\partial_t u - \nabla \cdot (D\nabla u - \mathbf{Q}u) = 0 \quad \text{in } J \times \Omega \tag{1}$$

(and semilinear system variants thereof) on a finite time interval $J =]0, t_{\text{end}}]$ and a polygonally bounded convex domain $\Omega \subset \mathbb{R}^2$. Equation (1) serves as a model for

F. Brunner (✉) · F. Frank · P. Knabner
Department of Mathematics, University of Erlangen–Nuremberg, Cauerstr. 11,
91058 Erlangen, Germany
e-mail: brunner@math.fau.de

F. Frank
e-mail: frank@math.fau.de

P. Knabner
e-mail: knabner@math.fau.de

J. Fuhrmann et al. (eds.), *Finite Volumes for Complex Applications VII - Methods and Theoretical Aspects*, Springer Proceedings in Mathematics & Statistics 77,
DOI: 10.1007/978-3-319-05684-5_16,

many natural processes, e.g., heat transfer or mass transport in porous media. The physical principle underlying this equation is conservation of mass, which should be reflected by any numerical method that is used for discretization.

The numerical simulation of (1) becomes particularly challenging if the advective term dominates the diffusive term, i.e., when the Péclet number is large. Then, sharp fronts in the solution cannot be resolved properly by conventional numerical schemes, which typically produce solutions that are polluted by spurious oscillations. To circumvent this, various approaches with different strengths and weaknesses were proposed in the literature. One of the most widely used techniques to handle advection dominance is upwinding, which is easy to implement and which preserves monotonicity well at the cost of introducing additional diffusion to the problem. It relies on the simple idea of discretizing the advection term as a function of the flow direction.

In this work, we compare a novel upwind stabilization of a mixed hybrid finite element scheme, which was studied numerically in [8] and analytically in [4] with a linear finite element scheme that uses an upwind finite volume approximation of the advective term. The latter was used in [5] to incorporate upwinding into an existing linear finite element code and thus to recover the discrete maximum principle, which is violated if linear finite elements are applied to advection-dominated transport problems.

By means of two test scenarios, we demonstrate that the upwind-mixed hybrid method is competitive with the finite volume upwind method with respect to robustness, monotonicity properties, and the amount of artificial numerical diffusion introduced by the schemes.

The rest of the work is organized as follows. In Sect. 2, the basic notation and the most important functional spaces are introduced. In Sect. 3, the discretization of problem {(1), (2)} with the two schemes under consideration is briefly sketched. Finally, Sect. 4 contains the description and the results of the test scenarios.

2 Notation and Problem Statement

Let the boundary $\partial\Omega$ decompose into a Dirichlet part $\partial\Omega_{\mathrm{D}}$, a Neumann part $\partial\Omega_{\mathrm{N}}$, and a flux part $\partial\Omega_{\mathrm{flux}}$ with outward unit normal $\mathbf{n}$. In order to obtain a well-posed problem, Eq. (1) is supplemented by the following initial and boundary conditions:

$$u = u_{\mathrm{D}} \qquad \text{on } J \times \partial\Omega_{\mathrm{D}}, \tag{2a}$$

$$-D\nabla u \cdot \mathbf{n} = 0 \qquad \text{on } J \times \partial\Omega_{\mathrm{N}}, \tag{2b}$$

$$-D\nabla u \cdot \mathbf{n} + u\,\mathbf{Q} \cdot \mathbf{n} = 0 \qquad \text{on } J \times \partial\Omega_{\mathrm{flux}}, \tag{2c}$$

$$u = u^0 \qquad \text{on } \{0\} \times \Omega \tag{2d}$$

with u_{D} and u_0 given. All coefficients are assumed to be sufficiently smooth.

Let the time interval J be decomposed into N subintervals of equal length and let $\Delta t := t_{\mathrm{end}}/N$ denote the time step size. Let $\mathscr{T}_h$ be a regular family of decomposi-

tions into closed triangles T of characteristic size h such that $\overline{\Omega} = \cup T$. We denote by $\mathbb{P}_k(T)$ the space of polynomials of degree at most k on a triangle $T \in \mathscr{T}_h$ and define by $\mathbb{RT}_0(T) := \{\mathbf{v} : T \to \mathbb{R}^2 \mid \mathbf{v}(\mathbf{x}) = a\mathbf{x} + \mathbf{b},\ a \in \mathbb{R},\ \mathbf{b} \in \mathbb{R}^2\}$ the local Raviart–Thomas space. Moreover, let $\mathbb{P}_k(\mathscr{T}_h) := \{w_h : \overline{\Omega} \to \mathbb{R} \mid \forall T \in \mathscr{T}_h, w_h|_T \in \mathbb{P}_k(T)\}$ denote the (discontinuous) global polynomial space on the triangulation $\mathscr{T}_h$ and let $\mathbb{P}_1^c(\mathscr{T}_h) := C(\overline{\Omega}) \cap \mathbb{P}_1(\mathscr{T}_h)$. The set of edges of $\mathscr{T}_h$ is denoted by $\mathscr{E}$ and that of $T \in \mathscr{T}_h$ by $\mathscr{E}(T)$, where we omit the index $_h$ here. Finally, let the space $H^1(\Omega)$ contain those functions of $L^2(\Omega)$ which have a weak derivative in $L^2(\Omega)$, and let $H^1_{0,\mathrm{D}}(\Omega)$ denote the subspace of $H^1(\Omega)$ consisting of functions with vanishing trace on $\partial\Omega_{\mathrm{D}}$.

3 Numerical Schemes

The two numerical schemes under consideration are outlined in the following. The first one is a linear finite element scheme that uses an upwind finite volume approximation of the advective term (LFEMstab) as presented in [5]. The second one is a mixed hybrid finite element scheme combined with an upwind-weighting formula based on the Lagrange multipliers (MHFEMstab), which are introduced into the formulation by hybridization; cf. [2, 4]. For ease of presentation, we assume that $u_{\mathrm{D}} = 0$ and we use full upwinding in the sequel. However, the schemes can be easily extended to inhomogeneous Dirichlet data and to more sophisticated upwind formulas, e.g., partial upwinding [6].

3.1 Scheme: LFEMstab

The discretization of {(1), (2)} with piecewise linear, globally continuous finite elements in space and with the implicit Euler method in time yields the following discrete problem.

Let $n \in \{1, \ldots, N\}$ and let $u_h^{n-1} \in \mathbb{P}_1^c(\mathscr{T}_h) \cap H^1_{0,\mathrm{D}}(\Omega)$ be given. Find $u_h^n \in \mathbb{P}_1^c(\mathscr{T}_h) \cap H^1_{0,\mathrm{D}}(\Omega)$ such that

$$\frac{1}{\Delta t}\int_\Omega (u_h^n - u_h^{n-1})z_h + \int_\Omega D\nabla u_h^n \cdot \nabla z_h + \int_\Omega z_h \nabla \cdot (\mathbf{Q}u_h^n) = 0$$

for all $z_h \in \mathbb{P}_1^c(\mathscr{T}_h) \cap H^1_{0,\mathrm{D}}(\Omega)$.

Let φ_k be the basis function of $\mathbb{P}_1^c(\mathscr{T}_h)$ that is associated with node $\mathbf{a}_k$, i.e., $\varphi_k(\mathbf{a}_j) = \delta_{kj}$ holds. Since φ_k has a local support on the triangles around $\mathbf{a}_k$, the advection term can be approximated by

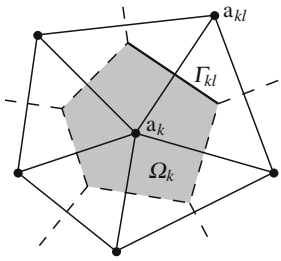


Fig. 1 Control volume Ω_k associated with the node $\mathbf{a}_k$ according to the Voronoi diagram of $\mathscr{T}_h$. The support of $\varphi_k \in \mathbb{P}_1^c(\mathscr{T}_h)$ is the union of all triangles containing the vertex $\mathbf{a}_k$

$$\int_{\Omega} \varphi_k \nabla \cdot (\mathbf{Q} u_h^n) \approx \int_{\Omega_k} \nabla \cdot (\mathbf{Q} u_h^n) = \int_{\partial \Omega_k} u_h^n \, \mathbf{Q} \cdot \mathbf{n} = \sum_j \int_{\Gamma_{kj}} u_h^n \, \mathbf{Q} \cdot \mathbf{n} \,,$$

where Ω_k is the Voronoi cell around $\mathbf{a}_k$, the boundary of which decomposes into line segments Γ_{kl}, $l \in \{1, 2, \ldots\}$, cf. Fig. 1. The boundary integral on Γ_{kl} can now be treated with a finite volume upwind scheme:

$$\int_{\Gamma_{kl}} u_h^n \, \mathbf{Q} \cdot \mathbf{n} \approx |\Gamma_{kl}| \, \alpha_{kl}(u_h^n) \, \mathbf{Q} \left(\frac{\mathbf{a}_k + \mathbf{a}_{kl}}{2} \right) \cdot \mathbf{n}$$

with $\mathbf{n}$ still denoting the unit normal pointing outward of Ω_k and with $\mathbf{a}_{kl}$ denoting the reflection of node $\mathbf{a}_k$ across Γ_{kl}. The function α for a full upwind scheme reads

$$\alpha_{kl}(u_h^n) := \begin{cases} u_h^n(\mathbf{a}_k) & \text{if } \mathbf{Q} \cdot \mathbf{n} \geq 0 \text{ at } (\mathbf{a}_k + \mathbf{a}_{kl})/2 \quad \text{(outflow of } \Omega_k) \,, \\ u_h^n(\mathbf{a}_{kl}) & \text{otherwise} \quad \text{(inflow into } \Omega_k) \,. \end{cases}$$

Using Voronoi cells as control volumes, for nonobtuse triangular meshes, LFEMstab is equivalent to the classical cell centered finite volume method if diffusion is cellwise constant, cf. [6]. Therefore, LFEMstab is locally mass conservative and provides formally first order accurate approximations of the scalar unknown u in $L^2(\Omega)$.

3.2 Scheme: MHFEMstab

In this section, the discretization of {(1), (2)} using the upwind-stabilized mixed hybrid finite element scheme of [4] is sketched. It relies on an Euler-implicit time stepping scheme and lowest order Raviart–Thomas finite elements for the spatial discretization. In contrast to non-hybrid schemes, the continuity constraints on the normal fluxes are not incorporated into the function space, but are ensured by introducing additional variables—the Lagrange multipliers—along with additional equations. More precisely, the space

$$\mathbf{V}_h := \{\mathbf{v} \in (L^2(\Omega))^2 \mid \forall T \in \mathscr{T}_h,\ \mathbf{v}|_T \in \mathbb{RT}_0(T)\}$$

is used as the ansatz space for the discrete approximation of the mass flux $\mathbf{q} := -D\nabla u + \mathbf{Q}u$. The Lagrange multipliers are taken from the space

$$\Lambda_h := \{\lambda \in L^2(\mathscr{E}) \mid \forall E \in \mathscr{E},\ \lambda|_E \in \mathbb{P}_0(E);\ \forall E \in \mathscr{E}_\mathrm{D},\ \lambda|_E = 0\}\,,$$

where $\mathscr{E}$ denotes the set of interior edges and $\mathscr{E}_\mathrm{D}$ the set of Dirichlet edges. Finally, the scalar unknown is approximated in the space $W_h := \mathbb{P}_0(\mathscr{T}_h)$.

The definition of the upwind-mixed hybrid scheme involves the discrete velocity field $\mathbf{Q}_h^n := \Pi_h \mathbf{Q}^n$, where Π_h denotes the usual Raviart–Thomas projection operator. We assume that $\mathbf{Q}_h^n$ has the representation $\mathbf{Q}_h^n = \sum_{T\in\mathscr{T}_h} \sum_{E\in\mathscr{E}(T)} Q_{TE}^n \mathbf{v}_{TE}$ in a basis $\{\mathbf{v}_{TE}\}_{T\in\mathscr{T}_h, E\in\mathscr{E}(T)}$ of $\mathbf{V}_h$. Basis functions of W_h and Λ_h are given by characteristic functions $\{\chi_T\}_{T\in\mathscr{T}_h}$ and $\{\mu_E\}_{E\in\mathscr{E}}$, respectively. The scheme MHFEMstab reads as follows.

Let $n \in \{1, \ldots, N\}$ and let $u_h^{n-1} \in W_h$ be given. Find $(\mathbf{q}_h^n, u_h^n, \lambda_h^n) \in \mathbf{V}_h \times W_h \times \Lambda_h$ with $\mathbf{q}_h^n = \sum_{T\in\mathscr{T}_h} \sum_{E\in\mathscr{E}(T)} q_{TE}^n \mathbf{v}_{TE}, u_h^n = \sum_{T\in\mathscr{T}_h} u_T^n \chi_T, \lambda_h^n = \sum_{E\in\mathscr{E}_\Omega} \lambda_E^n \mu_E$ such that

$$\int_\Omega D^{-1}\,\mathbf{v}_h \cdot \mathbf{q}_h^n - \int_\Omega u_h^n \nabla\cdot \mathbf{v}_h$$

$$- \sum_{T\in\mathscr{T}_h} \sum_{E\in\mathscr{E}(T)} Q_{TE}^n \alpha_{TE}(u_T^n, \lambda_E^n) \int_T D^{-1}\,\mathbf{v}_h \cdot \mathbf{v}_{TE} = -\sum_{T\in\mathscr{T}_h} \int_{\partial T} \lambda_h^n\, \mathbf{v}_h \cdot \mathbf{n}\,, \tag{3a}$$

$$\frac{1}{\Delta t} \int_\Omega (u_h^n - u_h^{n-1}) w_h + \int_\Omega w_h\, \nabla\cdot \mathbf{q}_h^n = 0\,, \tag{3b}$$

$$\sum_{T\in\mathscr{T}_h} \int_{\partial T} \mu_h\, \mathbf{q}_h^n \cdot \mathbf{n} = 0 \tag{3c}$$

for all $(\mathbf{v}_h, w_h, \mu_h) \in \mathbf{V}_h \times W_h \times \Lambda_h$, where the upwind weights are defined as

$$\alpha_{TE}(u_T^n, \lambda_E^n) = \begin{cases} u_T^n & \text{if } Q_{TE}^n \geq 0\,, \\ \lambda_E^n & \text{otherwise}\,. \end{cases} \tag{4}$$

The function α_{TE} takes the flow direction into account: If $Q_{TE}^n \geq 0$, i.e., if there is an outflow across the edge E, the value u_T^n on the current triangle is used to discretize the advective term. Otherwise, the Lagrange multiplier λ_E^n—which represents an approximation of u^n on E—is used. Note that the definition (4) of α_{TE} is slightly different than that in [4, 8], where $\alpha_{TE}(u_T^n, \lambda_E^n) := 2\lambda_E^n - u_T^n$ was used if $Q_{TE}^n < 0$. This is because less numerical diffusion was observed using (4). The proof of convergence in [4], however, applies to either choice.

Since the basis functions of $\mathbf{V}_h$ can be chosen to have support only on a single mesh element, static condensation is usually employed in standard cell-centered mixed hybrid schemes in order to reduce the number of global unknowns by local elimination. With the specific choice of α_{TE} in the above scheme, Eq. (3a) remains fully local and static condensation may be applied further on. Therefore, the upwind-mixed hybrid scheme can be implemented more efficiently than standard upwind-mixed schemes that use information from neighbor elements to discretize the advection term, cf. [4].

4 Robustness of the Schemes

In the following, the schemes LFEMstab and MHFEMstab presented in Sect. 3 and their non-stabilized versions LFEM and MHFEM are compared with respect to numerical attributes that are essential for reliable simulation of advection-dominated flows, e.g., monotonicity and the amount of artificial diffusion they introduce.

4.1 Scenario: Pulse

We consider a time interval $J :=]0, 1]$ using a time step size of $\Delta t := 5\text{E}{-3}$ and a rectangular domain $\Omega :=]0, 2[\times]0, 1[$ with $\partial\Omega_\text{N} = \{2\} \times [0, 1]$ and $\partial\Omega_\text{flux} = \partial\Omega \setminus \partial\Omega_\text{N}$, which is triangulated by an unstructured grid containing 2, 704 triangles. We choose the following data in {(1), (2)}: $D := 2\text{E}{-4}$, $\mathbf{Q} := (1, 0)^\text{T}$, and $u^0 := 1$ on $[1/4, 3/4]^2$ and zero elsewhere.

The center of mass of the initial (quadratic) distribution u^0 should be transported to $\mathbf{x} = (1.5, 0.5)^\text{T}$ by advection and be slightly smeared by diffusion. Figure 2 shows the distribution of u_h at t_end for the four schemes under investigation. The non-stabilized schemes LFEM and MHFEM produce oscillations that reach negative values. Although the MHFEM solution at t_end is closer to the expected one, the oscillations are stronger and lead to non-convergence shortly after t_end, which is not the case with LFEM. Both of the stabilized schemes LFEMstab and MHFEMstab conserve the nonnegativity of u^0, however, MHFEMstab adds less artificial diffusion to the solution.

4.2 Scenario: Contaminant Biodegradation

As a second example, we consider the simulation of contaminant biodegradation according to a simplified Monod model. This nonlinear test problem was used by several authors, cf. [1, 3, 7], to compare different numerical schemes with respect to the numerical diffusion they introduce. It illustrates that prognoses of methods with

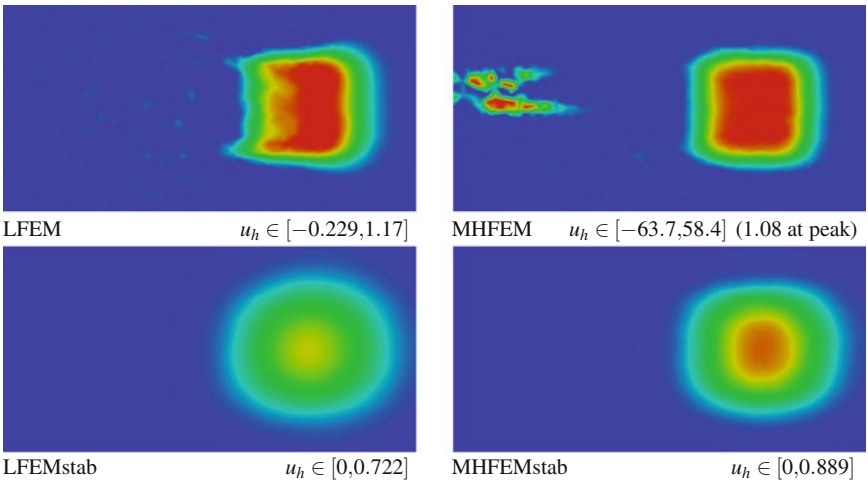


Fig. 2 The distribution of u_h at t_{end} for the different schemes under investigation. The color scaling is fixed from zero *(blue)* to one *(red)*; the global minima and maxima are listed below each plot

large artificial diffusion can be completely wrong. Precisely, a degradation reaction between an electron donor u_{don} (e.g. a contaminant) and an electron acceptor u_{acc} (e.g. oxygen) is considered, which is catalyzed by a bio species u_{bio}. As a simplification, biomass growth is neglected and the process is modeled by the equations

$$\partial_t(\theta u_i) - \nabla \cdot (\theta D \nabla u_i - \mathbf{Q} u_i) = \alpha_i \, \mu \ , \quad i \in \{\text{don}, \text{acc}\}$$

with the Monod reaction rate $\mu = -u_{\text{bio}} u_{\text{don}} (K_{\text{don}} + u_{\text{don}})^{-1} u_{\text{acc}} (K_{\text{acc}} + u_{\text{acc}})^{-1}$. For the simulation the following data are used: $\Omega =]0, 0.5[\times]0, 1[$, $\theta = 0.2$, $D = 10\text{E}{-4}$, $\mathbf{Q} = (0, -1)^{\text{T}}$, $\alpha_{\text{don}} = 5$, $\alpha_{\text{acc}} = 0.5$, $K_{\text{don}} = 0.1$, $K_{\text{acc}} = 0.1$, $u_{\text{bio}} = 1$. As initial conditions, $u^0_{\text{don}} = 0$ and $u^0_{\text{acc}} = 0.1$ are chosen in Ω. The electron donor is injected at the middle part of the upper boundary and transported toward the lower boundary by advection. The stationary Dirichlet boundary conditions are given by $u_{\text{don}} = 1$ and $u_{\text{acc}} = 0$ on $]0.225, 0.275[\times \{1\}$ and $u_{\text{don}} = 0$ and $u_{\text{acc}} = 0.1$ elsewhere on the upper boundary, respectively. The degradation reaction takes place only in those parts of the domain where the concentrations of both species are sufficiently large, which is the case at the interface between the electron donor and the surrounding area, where still enough electron acceptor is available. Thus, numerical methods introducing much artificial diffusion lead to an overestimation of the mixing zone of the two species, and the contaminant is degraded too fast in this case.

Figure 3 shows the predicted contaminant concentrations using LFEMstab and MHFEMstab on a locally preadapted unstructured grid with 1, 988 elements at t_{end}, where a steady state has been reached. For both schemes, Newton's method was used for the linearization of the nonlinear reaction terms. We observe that both methods

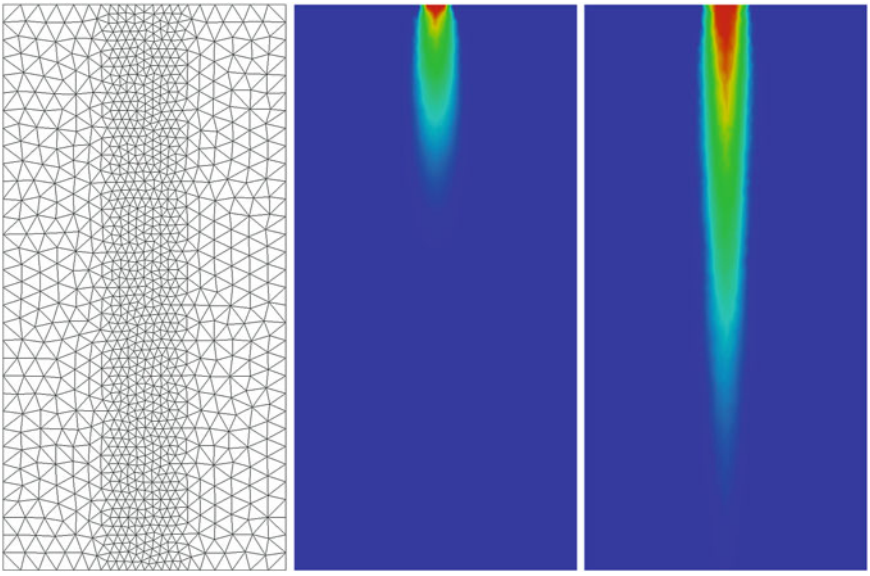


Fig. 3 Locally preadapted grid *(left)* and computed concentration profiles for $u_{\mathrm{don},h}$ using LFEMstab *(center)* and MHFEMstab *(right)*. The values of $u_{\mathrm{don},h}$ are in [0, 1] for both methods

produce nonnegative solutions. On this relatively coarse grid, LFEMstab predicts a complete degradation of the contaminant within the computational domain, which is incorrect and may have serious consequences in practice. The contaminant plume computed by MHFEMstab, however, covers the full length of the domain and reaches the outflow boundary. This is in accordance with a reference solution we computed on a grid with 250,000 elements.

5 Conclusion

We conclude that the upwind-mixed hybrid method provides a suitable scheme for simulating advection-driven transport problems. Compared to the finite volume method the amount of artificial diffusion appears to be lower, which is important in applications where the dominating processes take place in small interaction regions. Moreover, similarly to the classical cell-centered mixed method, it is fully hybridizable, and the incorporation of upwinding does not increase the computational costs in contrast to standard upwind-mixed methods.

References

1. Bause, M., Knabner, P.: Numerical simulation of contaminant biodegradation by higher order methods and adaptive time stepping. Comput. Vis. Sci. **7**, 61–78 (2004)
2. Boffi, D., Brezzi, F., Fortin, M.: Mixed Finite Element Methods and Applications. Springer, Berlin (2013)
3. Brunner, F., Radu, F.A., Bause, M., Knabner, P.: Optimal order convergence of a modified BDM_1 mixed finite element scheme for reactive transport in porous media. Adv. Water Resour. **35**, 163–171 (2012)
4. Brunner, F., Radu, F.A., Knabner, P.: Analysis of an upwind-mixed hybrid finite element method for transport problems. SIAM J. Num. Anal. **52**(1), 83–102 (2014)
5. Hoffmann, J.: Reactive transport and mineral dissolution/precipitation in porous media: efficient solution algorithms, benchmark computations and existence of global solutions. Ph.D. thesis, Friedrich–Alexander University of Erlangen–Nuremberg (2010)
6. Knabner, P., Angermann, L.: Numerical Methods for Elliptic and Parabolic Partial Differential Equations. Springer, New York (2003)
7. Ohlberger, M., Rohde, C.: Adaptive finite volume approximations of weakly coupled convection dominated parabolic systems. IMA J. Numer. Anal. **22**, 253–280 (2002)
8. Radu, F.A., Suciu, N., Hoffmann, J., Vogel, A., Kolditz, O., Park, C.H., Attinger, S.: Accuracy of numerical simulations of contaminant transport in heterogeneous aquifers: a comparative study. Adv. Water Resour. **34**(1), 47–61 (2011)

Entropy-Diminishing CVFE Scheme for Solving Anisotropic Degenerate Diffusion Equations

Clément Cancès and Cindy Guichard

Abstract We consider a Control Volume Finite Elements (CVFE) scheme for solving possibly degenerated parabolic equations. This scheme does not require the introduction of the so-called Kirchhoff transform in its definition. The discrete solution obtained *via* the scheme remains in the physical range whatever the anisotropy of the problem, while the natural entropy of the problem decreases with time. Moreover, the discrete solution converges towards the unique weak solution of the continuous problem. Numerical results are provided and discussed.

1 The Continuous Problem and Objectives

Let Ω be a polygonal open bounded and connected subset of $\mathbb{R}^2$, and let $t_\mathrm{f} > 0$ be a finite time horizon. We aim to approximate the solution of the (possibly) degenerate parabolic equation

$$\begin{cases} \partial_t u - \nabla \cdot (\eta(u) \Lambda \nabla p(u)) = 0 & \text{in } Q_{t_\mathrm{f}} := \Omega \times (0, t_\mathrm{f}), \\ \eta(u) \Lambda \nabla p(u) \cdot \mathbf{n} = 0 & \text{on } \partial\Omega \times (0, t_\mathrm{f}), \\ u_{|_{t=0}} = u_0 & \text{in } \Omega. \end{cases} \tag{1}$$

C. Cancès (✉) · C. Guichard
Laboratoire Jacques-Louis Lions, Sorbonne Universités, UPMC University Paris 06, UMR 7598, 75005 Paris, France
e-mail: cances@ljll.math.upmc.fr

C. Guichard
e-mail: guichard@ljll.math.upmc.fr

C. Cancès · C. Guichard
Laboratoire Jacques-Louis Lions, CNRS, UMR 7598, 75005 Paris, France

J. Fuhrmann et al. (eds.), *Finite Volumes for Complex Applications VII - Methods and Theoretical Aspects,* Springer Proceedings in Mathematics & Statistics 77,
DOI: 10.1007/978-3-319-05684-5_17,

In (1), $\Lambda : \Omega \to \mathscr{M}_2(\mathbb{R})$ should be a measurable function, and we assume that there exists $0 < \underline{\lambda} \leq \overline{\lambda}$ such that

$$\Lambda(\mathbf{x}) = \Lambda(\mathbf{x})^T, \qquad \underline{\lambda}|\mathbf{v}|^2 \leq \Lambda(\mathbf{x})\mathbf{v} \cdot \mathbf{v} \leq \overline{\lambda}|\mathbf{v}|^2 \quad \text{for all } \mathbf{v} \in \mathbb{R}^2 \text{ and a.a.} \mathbf{x} \in \Omega.$$

The function η is assumed to be continuous, to be such that $\eta(s) > 0$ if $s \in (0, 1)$ and $\eta(s) = 0$ otherwise. The function p belongs to $C^1((0, 1); \mathbb{R}) \cap L^1(0, 1)$, is supposed to be increasing, and to be such that $\lim_{s \to \{0,1\}} \eta(s)p(s) = 0$. Note that p is not necessarily bounded in the vicinity of 0 and 1. We also assume that $\sqrt{\eta}p'$ belongs to $L^1(0, 1)$, so that the *Kirchhoff transforms*

$$\phi : u \mapsto \int_0^u \eta(s)p'(s)\mathrm{d}s \quad \text{and} \quad \xi : u \mapsto \int_0^u \sqrt{\eta(s)}p'(s)\mathrm{d}s$$

are continuous and increasing on $[0, 1]$. The initial data u_0 is assumed to belong to $L^\infty(\Omega)$, and to be such that $0 \leq u_0(\mathbf{x}) \leq 1$ for a.a. $\mathbf{x} \in \Omega$.

Using the Kirchhoff transform ϕ, the problem (1) can be rewritten as

$$\begin{cases} \partial_t u - \nabla \cdot (\Lambda \nabla \phi(u)) = 0 & \text{in } Q_{t_\mathrm{f}}, \\ \Lambda \nabla \phi(u) \cdot \mathbf{n} = 0 & \text{on } \partial\Omega \times (0, t_\mathrm{f}), \\ u_{|_{t=0}} = u_0 & \text{in } \Omega. \end{cases} \tag{2}$$

Following [1, 10], there exists a unique weak solution to the problem (2). Moreover, the monotonicity of the problem ensures that

$$0 \leq u(\mathbf{x}, t) \leq 1 \quad \text{for a.a. } (\mathbf{x}, t) \in Q_{t_\mathrm{f}}. \tag{3}$$

Considering formally $p(u) - p(1/2)$ as a test function in (2) yields, for all $t \in [0, t_\mathrm{f}]$,

$$\int_\Omega H(u(\mathbf{x}, t))\mathrm{d}\mathbf{x} + \iint_{Q_t} \Lambda \nabla \xi(u) \cdot \nabla \xi(u)\mathrm{d}\mathbf{x}\mathrm{d}\tau = \int_\Omega H(u_0(\mathbf{x}))\mathrm{d}\mathbf{x} < \infty, \tag{4}$$

where $H(u) = \int_{1/2}^u (p(s) - p(1/2))\mathrm{d}s$ is a nonnegative convex entropy to the problem (1), which is supposed to physically meaningful. As a consequence of (4), the function $\xi(u)$ belongs to $L^2((0, T); H^1(\Omega))$.

Since in many configurations like for example porous media flows, the physical meaning of the Kirchhoff transform ϕ is unclear (see e.g. [3, 11]), we aim to discretize the problem in its form (1) rather than in its form (2). Moreover, we aim to derive a method such that the L^∞-estimate (3) remains true at the discrete level despite the anisotropy of the problem, such that the discrete counterpart of the entropy $\int_\Omega H(u(\mathbf{x}, t))\mathrm{d}\mathbf{x}$ decreases with time as prescribed by (4) in the continuous setting, and such that the discrete solution converges towards the unique weak solution as the discretization steps tend to 0. Let us stress that the decay of the entropy $\int_\Omega H(u(\mathbf{x}, t))\mathrm{d}\mathbf{x}$ plays an important role in the long-time behavior of the continuous and discrete solutions [5, 6].

2 The Implicit Nonlinear CVFE Scheme

Let $\mathscr{T}$ be a conforming triangulation of Ω with size $h_{\mathscr{T}} = \max_{T\in\mathscr{T}} h_T$, where h_T is the diameter of T, and regularity $\theta_{\mathscr{T}} = \max_{T\in\mathscr{T}} \frac{h_T}{\rho_T}$ where ρ_T is the diameter of the incircle of the triangle T. We denote by $\mathscr{V}$ the set of the vertices and by $\mathscr{E}$ the set of the edges of $\mathscr{T}$. For all $K \in \mathscr{V}$ (located at $\mathbf{x}_K$), we denote by $\mathscr{T}_K$ the set of the triangles of $\mathscr{T}$ admitting K as a vertex, by $\mathscr{V}_K$ the subset of $\mathscr{V}$ made of the vertices connected to K *via* an edge, and by $\mathscr{E}_K$ the set of the edges having $\mathbf{x}_K$ as an endpoint. The edge joining two vertices K and L is denoted by σ_{KL}. For all $K \in \mathscr{V}$, the star-shaped open subset ω_K of Ω is delimited by the centers of gravity $\mathbf{x}_T$ of the triangles $T \in \mathscr{T}_K$ and $\mathbf{x}_\sigma$ of the edges $\sigma \in \mathscr{E}_K$, yielding the dual barycentric mesh $\mathscr{M}$. We denote by

$$V_{\mathscr{T}} = \{f \in C(\Omega) \mid f_{|T} \in \mathbb{P}_1(T),\ \forall T \in \mathscr{T}\}$$

the usual $\mathbb{P}_1$-finite element space, and by $(e_K)_{K\in\mathscr{V}}$ the canonical basis of $V_{\mathscr{T}}$. We also introduce the set

$$X_{\mathscr{M}} = \{f \in L^\infty(\Omega) \mid f_{|\omega_K} \in \mathbb{P}_0(K),\ \forall K \in \mathscr{V}\}$$

of the piecewise constant functions on the dual cells. For the ease of notations, we restrict our study to the case of uniform time discretizations with step $\Delta t = t_{\mathrm{f}}/N$. Setting $t_n = n\Delta t$ for $0 \le n \le N$, we introduce the discrete functional sets

$$\begin{aligned} V_{\mathscr{T},\Delta t} &= \{f \in L^\infty(Q_{t_{\mathrm{f}}}) \mid f(\cdot,t) = f(\cdot,n\Delta t) \in V_{\mathscr{T}},\ \forall t \in (t_{n-1},t_n],\ 1 \le n \le N\},\\ X_{\mathscr{M},\Delta t} &= \{f \in L^\infty(Q_{t_{\mathrm{f}}}) \mid f(\cdot,t) = f(\cdot,n\Delta t) \in X_{\mathscr{M}},\ \forall t \in (t_{n-1},t_n],\ 1 \le n \le N\}. \end{aligned}$$

Hence, given $\left(v_K^n\right)_{K\in\mathscr{V},1\le 0\le N}$, there exist two reconstructions $v_{\mathscr{T},\Delta t} \in v_{\mathscr{T},\Delta t}$ and $v_{\mathscr{M},\Delta t} \in X_{\mathscr{M},\Delta t}$ such that

$$v_{\mathscr{T},\Delta t}(\mathbf{x}_K,t_n) = v_{\mathscr{M},\Delta t}(\mathbf{x}_K,t_n) = v_K^n, \text{ for all } K \in \mathscr{V} \text{and } n \in \{1,\dots,N\}.$$

For $K \in \mathscr{V}$, we set $m_K = \int_{\omega_K} \mathrm{d}\mathbf{x} = \int_\Omega e_K(\mathbf{x})\mathrm{d}\mathbf{x}$. The initial data u_0 is discretized by $u_{\mathscr{M}}^0 \in X_{\mathscr{M}}$, where

$$u_K^0 = \frac{1}{m_K}\int_{\omega_K} u_0(\mathbf{x})\mathrm{d}\mathbf{x}, \qquad \forall K \in \mathscr{V}. \tag{5}$$

We introduce now the so-called implicit *nonlinear CVFE* scheme [2], which is closely related to $\mathbb{P}_1$-finite elements with mass lumping. Let $n \ge 1$, then for $\left(u_K^{n-1}\right)_{K\in\mathscr{V}}$ in $[0,1]^{\#\mathscr{V}}$, we look for $\left(u_K^n\right)_{K\in\mathscr{V}}$ such that

$$\frac{u_K^n - u_K^{n-1}}{\Delta t} m_K + \sum_{L \in \mathscr{V}_K} a_{KL} \eta_{KL}^n \left(p(u_K^n) - p(u_L^n) \right) = 0, \qquad \forall K \in \mathscr{V}. \tag{6}$$

In (6), we have set $a_{KL} = -\int_\Omega \Lambda \nabla e_K \cdot \nabla e_L \mathrm{d}\mathbf{x} = a_{LK}$ and

$$\eta_{KL}^n = \begin{cases} \max_{s \in I_{KL}^n} \eta(s) & \text{if } a_{KL} > 0, \\ \min_{s \in I_{KL}^n} \eta(s) & \text{if } a_{KL} \leq 0, \end{cases} \qquad \text{with } I_{KL}^n = \left[\min(u_K^n, u_L^n), \max(u_K^n, u_L^n) \right].$$

Setting $F_{KL}^n = a_{KL} \eta_{KL}^n \left(p(u_K^n) - p(u_L^n) \right)$, it is easy to check that for all $n \geq 1$,

$$\begin{cases} F_{KL}^n + F_{LK}^n = 0, & \forall \sigma_{KL} \in \mathscr{E}, \\ \frac{u_K^n - u_K^{n-1}}{\Delta t} m_K + \sum_{L \in \mathscr{V}_K} F_{KL}^n = 0, & \forall K \in \mathscr{V}, \end{cases}$$

leading naturally to the following statement.

Proposition 1 *The scheme* (6) *is locally conservative on the dual mesh* $\mathscr{M}$.

As usually in CVFE schemes, the accumulation term is obtained thanks to mass-lumping. The originality here comes from the treatment of the diffusion term. The flux F_{KL}^n is obtained as the flux across the interface $x = 0$ corresponding to the simili Riemann problem

$$\begin{cases} \partial_t v + \partial_x \left(\eta(v) q_{KL}^n \right) = 0, & \text{in } \mathbb{R} \times \mathbb{R}_+^*, \\ v_{|t=0} = u_K^n \mathbf{1}_{x<0} + u_L^n \mathbf{1}_{x>0} & \text{in } \mathbb{R}, \\ q_{KL}^n = a_{KL}(p(u_K^n) - p(u_L^n)). \end{cases}$$

3 Discrete Estimates and Convergence of the Scheme

All the numerical analysis results stated in this section are thoroughly justified in the forthcoming paper [4]. First, let us give some *a priori* estimates.

Proposition 2 *For all* $n \geq 1$, *one has*

$$\begin{aligned} &\sum_{K \in \mathscr{V}} H(u_K^n) m_K + \Delta t \sum_{\sigma_{KL} \in \mathscr{E}} a_{KL} \left(\xi(u_K^n) - \xi(u_L^n) \right)^2 \\ &\leq \sum_{K \in \mathscr{V}} H(u_K^n) m_K + \Delta t \sum_{\sigma_{KL} \in \mathscr{E}} a_{KL} \eta_{KL}^n \left(p(u_K^n) - p(u_L^n) \right)^2 \leq \sum_{K \in \mathscr{V}} H(u_K^{n-1}) m_K. \end{aligned} \tag{7}$$

Sketch of the proof. In order to prove the second inequality of (7), multiply the scheme (6) by $\Delta t \left(p(u_K^n) - p(1/2) \right)$ and to sum over $K \in \mathscr{V}$. The inequality

$$(u_K^n - u_K^{n-1})\left(p(u_K^n) - p(1/2)\right) \geq H(u_K^n) - H(u_K^{n-1})$$

stems from the definition and the convexity of H.

The first inequality of (7) is a consequence of the definitions of ξ and η_{KL}^n, which ensure that

$$a_{KL}\eta_{KL}^n\left(p(u_K^n) - p(u_L^n)\right)^2 \geq a_{KL}\left(\xi(u_K^n) - \xi(u_L^n)\right)^2,$$

for all $\sigma_{KL} \in \mathscr{E}$ and all $n \geq 1$.

Denoting by $\xi_{\mathscr{T}}^n$ the function of $V_{\mathscr{T}}$ with nodal values $\left(\xi(u_K^n)\right)_{K\in\mathscr{V}}$, and by $u_{\mathscr{M}}^n$ the function of $X_{\mathscr{M}}$ with nodal values $\left(u_K^n\right)_{K\in\mathscr{T}}$, then Proposition 2 implies that

$$\int_\Omega H(u_{\mathscr{M}}^n)\mathrm{d}\mathbf{x} + \int_{t_{n-1}}^{t_n}\int_\Omega \Lambda\nabla\xi_{\mathscr{T}}^n\cdot\nabla\xi_{\mathscr{T}}^n\mathrm{d}\mathbf{x}\mathrm{d}t \leq \int_\Omega H(u_{\mathscr{M}}^{n-1})\mathrm{d}\mathbf{x},$$

ensuring that the entropy is dissipated at each time step.

The discrete diffusion operator appearing in the scheme (6) can be split into two parts:

- a monotone part

$$\left(u_K^n\right)_K \mapsto \left(\sum_{L\in\mathscr{V}_K}(a_{KL})^+\eta_{KL}^n\left(p(u_K^n) - p(u_L^n)\right)\right)_K$$

  whose contribution preserves the maximum principle. This contribution consists also in a diffusion operator, but it is not a consistent discretization of the continuous operator $u \mapsto -\nabla\cdot(\Lambda\eta(u)\nabla p(u))$;
- a correcting part

$$\left(u_K^n\right)_K \mapsto \left(-\sum_{L\in\mathscr{V}_K}(a_{KL})^-\eta_{KL}^n\left(p(u_K^n) - p(u_L^n)\right)\right)_K$$

  that ensures the consistency of the scheme. Due to the definition of η_{KL}^n, this contribution is continuous and vanishes where $\eta(u_K^n) = 0$, i.e.,

$$\sum_{L\in\mathscr{V}_K}(a_{KL})^-\eta_{KL}^n\left(p(u_K^n) - p(u_L^n)\right) = 0 \quad \text{if } u_K^n \in \{0,1\}.$$

Therefore, the scheme preserves the natural L^∞ bounds 0 and 1. This is the purpose of Proposition 3 (the proof is given in [4]), that moreover ensures that all the terms in (6) are finite.

Proposition 3 *For all $n \geq 1$ and for all $K \in \mathscr{V}$, $0 \leq u_K^n \leq 1$. Moreover, if there exist $K_\star$ (resp. $K^\star$) in $\mathscr{V}$ such that $u_{K_\star}^0 > 0$ (resp. $u_{K^\star}^0 < 1$), and if $\lim_{s\to 0} p(s) = -\infty$ (resp. $\lim_{s\to 1} p(s) = +\infty$), then for all $K \in \mathscr{V}$ and all $n \geq 1$, one has*

$$u_K^n \geq c(u_{K_\star}^0, \mathscr{T}, \Delta t, n, \Lambda) > 0 \qquad (\text{resp. } u_K^n \leq 1 - c(u_{K^\star}^0, \mathscr{T}, \Delta t, n, \Lambda) < 1). \tag{8}$$

All these *a priori* estimates allow us to prove the existence of a discrete solution $\left(u_K^n\right)_{K\in\mathscr{V}, n\geq 0}$ to the scheme (5)–(6).

Proposition 4 *For all $n \in \{1, \ldots, N\}$, there exists a solution $\left(u_K^n\right)_{K\in\mathscr{V}}$ to the scheme (5)–(6).*

The proof of Proposition 4 is inspired from the existence proof given in [7] and relies on a topological degree argument. Nevertheless, in the case where p is unbounded, the application $\left(u_K^n\right)_{K\in\mathscr{V}} \mapsto \sum_{L\in\mathscr{V}_K} a_{KL}\eta_{KL}^n \left(p(u_K^n) - p(u_L^n)\right)$ is not continuous on $[0, 1]^{\#\mathscr{V}}$. Therefore, the enhanced L^∞-estimates (8) are mandatory to restrict the study on a smaller domain $[\epsilon, 1-\epsilon]^{\#\mathscr{V}}$ on which the discrete operator is uniformly continuous.

In what follows, we denote by $u_{\mathscr{M},\Delta t}$ the unique element of $X_{\mathscr{M},\Delta t}$ such that

$$u_{\mathscr{M},\Delta t}(\mathbf{x}_K, t_n) = u_K^n, \qquad \forall K \in \mathscr{V},\ \forall n \in \{1, \ldots, N\}. \tag{9}$$

The convergence of the discrete solution $u_{\mathscr{M},\Delta t}$ as the space and time discretization steps tend to 0 towards the unique weak solution u of the continuous problem is the purpose of the following theorem, whose proof is contained in [4].

Theorem 1 *Let $(\mathscr{T}_m)_{m\geq 1}$ be a sequence of conforming triangulations of Ω such that $h_{\mathscr{T}_m} \to 0$ as $m \to \infty$, and such that $\theta_{\mathscr{T}_m} \leq \theta^\star < \infty$, and let $(\Delta t_m)_{m\geq 1}$ be a sequence of time steps such that $\Delta t_m \to 0$ as $m \to \infty$, then, for all $q \in [1, \infty)$, the discrete solution $u_{\mathscr{M}_m,\Delta t_m}$ converges in $L^q(Q_{t_f})$ towards the unique weak solution u of the problem (2) as $m \to \infty$.*

The proof of Theorem 1 follows (with some additional technical difficulties) the path proposed in §4.3 of [8], that consists in first proving some compactness on the family of discrete solutions $\left(u_{\mathscr{M}_m,\Delta t_m}\right)_{m\geq 1}$, and then to identify any limit value (up to a subsequence) $u = \lim_{m\to\infty} u_{\mathscr{M}_m,\Delta t_m}$ as the unique weak solution to the problem (2). The uniqueness of the limit ensures the convergence of the whole sequence.

Remark 1 The choice of homogeneous Neumann boundary condition is not mandatory. A similar convergence result can be obtained in the case where an inhomogeneous Dirichlet condition u_D is imposed on a part of the boundary as long as u_D and $p(u_D)$ are sufficiently regular. In this case, the entropy of the system if not necessarily decreasing (but it remains bounded) because of a contribution coming from the boundary.

4 Numerical Illustration

This section illustrates the numerical behavior of the scheme (6). In order to compare the numerical solution with an analytical solution, we apply our discretization strategy on a test case that does not fully fits with our assumptions. Indeed, Dirichlet boundary conditions are prescribed. But as mentioned in Remark 1, the convergence of the scheme can be proved also in this case. The meshes used for the discretization of the domain $\Omega = (0, 1)^2$ are issued from a 2D benchmark on anisotropic diffusion problem [9]. These triangle meshes show no symmetry which could artificially increase the convergence rate, and all angles of triangles are acute. This allows to compare situations where all coefficients a_{KL} defined previously are positive, with situations where some of them are negative by introducing anisotropic permeability tensors. This family of meshes is built through the same pattern, which is reproduced at different scales.

In the following numerical experiments, we consider a diagonal permeability tensor $\Lambda = \mathrm{diag}(l_x, l_y)$. A first constant time step, denoted by Δt_1, is associated to the coarsest mesh and then between two successive meshes, the time step is divided by four since the mesh size is divided by two, so that the error due to the implicit Euler-time discretization remains negligible compared to that issued from the space discretization. The nonlinear systems obtained at each time step are solved by a Newton-Raphson algorithm.

The test case deals with a degenerate parabolic equation with a Dirichlet boundary condition. The functions involved in (1) are defined by $\eta_{n\ell}(u) = 2\min(u, 1-u)$ and $p_{n\ell}(u) = u$. Since the continuous solution and the discrete one computed with the nonlinear scheme (6) remain bounded between 0 and 0.5 (cf. Tables 1 and 2), this amounts to consider the porous medium equation

$$\partial_t u - \nabla \cdot \left(\Lambda \nabla u^2 \right) = 0,$$

and we compare the results with the scheme obtained by taking the following functions $p_\ell(u) = u^2$ and $\eta_\ell(u) = 1$ where the subscript ℓ has been added for this formulation called the quasilinear one. The numerical convergence of both schemes has been studied through the following analytical solution,

$$\tilde{u}((x, y), t) = \max(2l_x t - x, 0),$$

for $(x, y) \in \Omega$, $t \in (0, t_\mathrm{f})$, and where the final time t_f has been fixed to 0.25 s and the first time step is given by $\Delta t_1 = 0.01024$ s. The values of $\tilde{u}$ on $\partial\Omega \times (0, t_\mathrm{f})$ are prescribed as Dirichlet boundary condition. Two permeability tensors have been tested : the isotropic one $l_x = l_y = 1$ (cf. Table 1) and an anisotropic one $l_x = 1$, $l_y = 10^2$ (cf. Table 2). For all tests we have computed the errors in the classical discrete $L^1(Q_{t_\mathrm{f}})$ and $L^\infty(Q_{t_\mathrm{f}})$ norms. Each table provides the mesh size h, the discrete errors and the associated convergence rate, and finally the minimum and maximum values of the discrete solutions.

Table 1 $l_x = l_y = 1$: the isotropic case

h	$\mathrm{err}_\ell - L^1$	Rate	$\mathrm{err}_{n\ell} - L^1$	Rate	$\mathrm{err}_\ell - L^\infty$	Rate	$\mathrm{err}_{n\ell} - L^\infty$	Rate	Min u_ℓ	Max u_ℓ	Min $u_{n\ell}$	Max $u_{n\ell}$
0.250	0.842E-03	-	0.215E-02	-	0.421E-01	-	0.810E-01	-	0.000	0.500	0.000	0.500
0.125	0.280E-03	1.589	0.115E-02	0.907	0.199E-01	1.080	0.564E-01	0.522	0.000	0.500	0.000	0.500
0.063	0.859E-04	1.704	0.611E-03	0.907	0.942E-02	1.081	0.365E-01	0.629	0.000	0.500	0.000	0.500
0.031	0.250E-04	1.782	0.322E-03	0.923	0.453E-02	1.056	0.229E-01	0.674	0.000	0.500	0.000	0.500
0.016	0.704E-05	1.827	0.168E-03	0.942	0.223E-02	1.024	0.140E-01	0.710	0.000	0.500	0.000	0.500

The subscript $n\ell$ stands for nonlinear scheme (6) while the subscript ℓ stands for the quasilinear scheme

Table 2 $l_x = 1, l_y = 10^2$: the anisotropic case

h	$\text{err}_\ell - L^1$	Rate	$\text{err}_{n\ell} - L^1$	Rate	$\text{err}_\ell - L^\infty$	Rate	$\text{err}_{n\ell} - L^\infty$	Rate	Min u_ℓ	Max u_ℓ	Min $u_{n\ell}$	Max $u_{n\ell}$
0.250	0.736E-02	-	0.469E-02	-	0.777E+00	-	0.110E+00	-	-0.777	0.500	0.000	0.500
0.125	0.254E-02	1.536	0.350E-02	0.423	0.373E+00	1.061	0.933E-01	0.243	-0.341	0.500	0.000	0.500
0.063	0.674E-03	1.913	0.243E-02	0.527	0.164E+00	1.184	0.751E-01	0.314	-0.149	0.500	0.000	0.500
0.031	0.188E-03	1.846	0.160E-02	0.604	0.812E-01	1.014	0.600E-01	0.325	-0.074	0.500	0.000	0.500
0.016	0.531E-04	1.820	0.101E-02	0.664	0.407E-01	0.998	0.472E-01	0.344	-0.037	0.500	0.000	0.500

The subscript $n\ell$ stands for nonlinear scheme (6) while the subscript ℓ stands for the quasilinear scheme

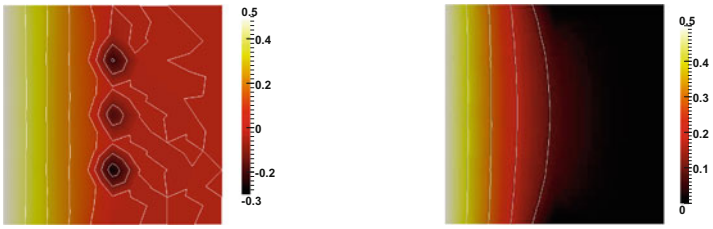


Fig. 1 2nd mesh. Discrete unknown u and its iso-values, for each scheme (*right*: quasilinear diffusion scheme, *left*: nonlinear scheme), with an anisotropic tensor at the end of the simulation

We observe that, as expected, the convergence rates of the linear implementation are better. Nevertheless, in the anisotropic case, the magnitude of the undershoots (illustrated by Fig. 1) is such that the absolute value of the observed error is lower in the nonlinear implementation than in the quasilinear one for the coarsest meshes, which are currently used in industrial applications.

Acknowledgments This work was supported by the French National Research Agency ANR (project GeoPor, grant ANR-13-JS01-0007-01).

References

1. Alt, H.W., Luckhaus, S.: Quasilinear elliptic-parabolic differential equations. Math. Z. **183**(3), 311–341 (1983)
2. Baliga, B.R., Patankar, S.V.: A control volume finite-element method for two-dimensional fluid flow and heat transfer. Numer. Heat Transfer **6**(3), 245–261 (1983)
3. Bear, J.: Dynamic of Fluids in Porous Media. American Elsevier, New York (1972)
4. Cancès, C., Guichard, C.: Convergence of a nonlinear entropy diminishing control volume finite element scheme for solving anisotropic degenerate parabolic equations (2014). HAL: hal-00955091
5. Chainais-Hillairet, C.: Entropy method and asymptotic behaviours of finite volume schemes. In: FVCA7 conference proceedings (2014).
6. Chainais-Hillairet, C., Jüngel, A.S.S.: Entropy-dissipative discretization of nonlinear diffusion equations and discrete Beckner inequalities (2014). HAL: hal-00924282
7. Eymard, R., Gallouët, T., Ghilani, M., Herbin, R.: Error estimates for the approximate solutions of a nonlinear hyperbolic equation given by finite volume schemes. IMA J. Numer. Anal. **18**(4), 563–594 (1998)
8. Eymard, R., Gallouët, T., Herbin, R.: Finite volume methods. In: Ciarlet, P.G. et al. (ed.) Handbook of numerical analysis, pp. 713–1020. North-Holland, Amsterdam (2000)
9. Herbin, R., Hubert, F.: Benchmark on discretization schemes for anisotropic diffusion problems on general grids. In: Eymard, R, Herard, J.M. (eds.) Finite volumes for complex applications V, pp. 659–692. Wiley (2008)
10. Otto, F.: L^1-contraction and uniqueness for quasilinear elliptic-parabolic equations. J. Diff. Equat. **131**, 20–38 (1996)
11. Otto, F.: The geometry of dissipative evolution equations: the porous medium equation. Commun. Partial Differ. Equ. **26**(1–2), 101–174 (2001)

A Finite Volume Scheme with the Discrete Maximum Principle for Diffusion Equations on Polyhedral Meshes

Alexey Chernyshenko and Yuri Vassilevski

Abstract We present a cell-centered finite volume (FV) scheme with the compact stencil formed mostly by the closest neighboring cells. The discrete solution satisfies the discrete maximum principle and approximates the exact solution with second-order accuracy. The coefficients in the FV stencil depend on the solution; therefore, the FV scheme is nonlinear. The scheme is applied to the steady state diffusion equation discretized on a general polyhedral mesh.

1 Introduction

We present a new monotone FV method for the 3D diffusion equation with anisotropic coefficients based on a nonlinear multi-point flux approximation scheme. It satisfies the discrete maximum principle (DMP), works for full anisotropic diffusion tensors and on polyhedral meshes, provides the second order accuracy and has a compact stencil. The basic idea of our approach belongs to LePotier [7] who proposed a monotone FV scheme with a nonlinear two-point flux approximation for the discretization of parabolic equations on triangular meshes. The method was extended to steady-state diffusion problems with full anisotropic tensors on general meshes [4, 8, 11]. For a comprehensive review of nonlinear FV methods we refer to [5]. Recently a new cell-centered minimal stencil FV method with DMP was proposed for full diffusion tensors and unstructured conformal polygonal 2D meshes [9]. The

A. Chernyshenko (✉) · Y. Vassilevski
Institute of Numerical Mathematics, Gubkina 8, Moscow, Russia
e-mail: chernyshenko.a@gmail.com

A. Chernyshenko
Institute of Nuclear Safety, B.Tulskaya 52, Moscow, Russia

Y. Vassilevski
Moscow Institute of Physics and Technology, Institutski 9, Dolgoprudny, M.R., Russia
e-mail: yuri.vassilevski@gmail.com

J. Fuhrmann et al. (eds.), *Finite Volumes for Complex Applications VII - Methods and Theoretical Aspects,* Springer Proceedings in Mathematics & Statistics 77,
DOI: 10.1007/978-3-319-05684-5_18,

3D extension of the method was proposed in [3], the similar algorithm was proposed independently in [6]. In this paper, we demonstrate the properties of the 3D method from [3] on the set of benchmark problems [1]. The FV scheme works on general polyhedral meshes and satisfies DMP in contrast to nonlinear two-point FV scheme from [4], which provides only non-negativity of the discrete solution.

2 Steady State Diffusion Equation

Let Ω be a three-dimensional polyhedral domain with boundary Γ.We consider a model diffusion problem for unknown concentration c:

$$\begin{aligned} -\mathrm{div}(\mathbb{K}\nabla c) &= g \quad \text{in} \quad \Omega \\ c &= g_D \quad \text{on} \quad \Gamma_D \\ -\mathbf{n}\cdot\mathbb{K}\nabla c &= g_N \quad \text{on} \quad \Gamma_N, \end{aligned} \tag{1}$$

where $\Gamma = \Gamma_D \cup \Gamma_N$, $\Gamma_D \neq \emptyset$, $\mathbb{K}(\mathbf{x}) = \mathbb{K}^T(\mathbf{x}) > 0$ is a diffusion tensor, g is a source term and $\mathbf{n}$ is the exterior normal vector.

We consider a conformal polyhedral mesh $\mathscr{T}$ composed of shape-regular cells with planar faces. We assume that each cell is a star-shaped 3D domain with respect to its barycenter. For simplicity, we assume that the diffusion tensor $\mathbb{K}(\mathbf{x})$ is constant inside each cell; however it may jump across mesh faces as well as may change orientation of principal directions.

We denote by $\mathscr{F}_I$, $\mathscr{F}_B$ disjoint sets of interior and boundary faces, respectively. The subset $\mathscr{F}_J \subset \mathscr{F}_I$ collects faces with jumping tensor. Let $\mathscr{F}_T$ denote the sets of faces of polyhedron T. The set $\mathscr{F}_B$ is further split into subsets $\mathscr{F}_B^D$ and $\mathscr{F}_B^N$ where the Dirichlet and Neumann boundary conditions, respectively, are imposed.

3 Nonlinear FV Scheme

The FV scheme uses one degree of freedom, C_T, per cell T collocated at $\mathbf{x}_T$, the barycenter of the cell. For every face $f \in \mathscr{F}_I \cup \mathscr{F}_B$, we denote the face barycenter by $\mathbf{x}_f$ and associate a collocation point with $\mathbf{x}_f$ for $f \in \mathscr{F}_B$.

We shall refer to collocation points on faces as the auxiliary collocation points. They are introduced for mathematical convenience and will not enter the final algebraic system although will affect system coefficients. In contrast, we shall refer to the other collocation points as the primary collocation points whose discrete concentrations form the unknown vector in the algebraic system.

For every cell T we define a set Σ_T of nearby collocation points. First, we add to Σ_T the collocation point $\mathbf{x}_T$. Then, for every face $f \in \mathscr{F}_T \setminus (\mathscr{F}_J \cup \mathscr{F}_B)$, we add the

collocation point $\mathbf{x}_{T'_f}$, where T'_f is the cell sharing f with T. Finally, for boundary faces $f \in \mathscr{F}_T \cap \mathscr{F}_B$, we add the collocation point $\mathbf{x}_f$.

Let $\mathbf{q} = -\mathbb{K}\nabla c$ denote the flux which satisfies the mass balance equation:

$$\operatorname{div} \mathbf{q} = g \quad \text{in} \quad \Omega. \tag{2}$$

A cell-centered FV scheme is derived by integrating Eq. (2) over a polyhedral cell T and using the Green's formula:

$$\int_{\partial T} \mathbf{q} \cdot \mathbf{n}_T \, \mathrm{d}s = \int_T g \, \mathrm{d}x, \tag{3}$$

where $\mathbf{n}_T$ denotes the external unit normal to ∂T. Let f denote a face of cell T and $\mathbf{n}_f$ be the corresponding normal vector. It will be convenient to assume that $|\mathbf{n}_f| = |f|$, where $|f|$ denotes the area of face f. The Eq. (3) becomes

$$\sum_{f \in \partial T} \mathbf{q}_f \cdot \mathbf{n}_f = \int_T g \, \mathrm{d}x, \tag{4}$$

where $\mathbf{q}_f$ is the average flux density for face f.

3.1 Diffusive Flux in Homogeneous Anisotropic Medium

Let us first consider a homogeneous medium. We assume that for every cell-face pair $T_i \in \mathscr{T}$, $f \in \mathscr{F}_{T_i}$, there exist three points $\mathbf{x}_{f,j}$, $\mathbf{x}_{f,k}$, and $\mathbf{x}_{f,l}$ in set Σ_{T_i} such that the following condition holds: the co-normal vector $\boldsymbol{\ell}_f = \mathbb{K}(\mathbf{x}_f)\mathbf{n}_f$ started from $\mathbf{x}_{T_i}$ belongs to the trihedral corner formed by vectors

$$\mathbf{t}_{ij} = \mathbf{x}_{f,j} - \mathbf{x}_{T_i}, \quad \mathbf{t}_{ik} = \mathbf{x}_{f,k} - \mathbf{x}_{T_i}, \quad \mathbf{t}_{il} = \mathbf{x}_{f,l} - \mathbf{x}_{T_i}, \tag{5}$$

and

$$\ell_f = \alpha_{ij}\mathbf{t}_{ij} + \alpha_{ik}\mathbf{t}_{ik} + \alpha_{il}\mathbf{t}_{il}, \tag{6}$$

where $\alpha_{ij} > 0$, $\alpha_{ik} \geq 0$, $\alpha_{il} \geq 0$. We assume that the first point, $\mathbf{x}_{f,j}$, belongs to the cell T_j which shares f with T_i. If Σ_{T_i} does not contain the desired points, one can extend Σ_{T_i} with other neighbors of T_i. This extension leads to increasing the minimal stencil. The algorithm of a search of such points is described in [4].

Recalling the definition of the diffusive flux and using finite differences to approximate directional derivatives, we obtain:

$$\begin{aligned} \mathbf{q} \cdot \mathbf{n}_f &= -\nabla c \cdot (\mathbb{K}_{T_i} \mathbf{n}_f) = -\alpha_{ij}\nabla c \cdot \mathbf{t}_{ij} - \alpha_{ik}\nabla c \cdot \mathbf{t}_{ik} - \alpha_{il}\nabla c \cdot \mathbf{t}_{il} \\ &= -\alpha_{ij}(C_{T_j} - C_{T_i}) - \alpha_{ik}(C_{T_k} - C_{T_i}) - \alpha_{il}(C_{T_l} - C_{T_i}) + O(|f|). \end{aligned} \tag{7}$$

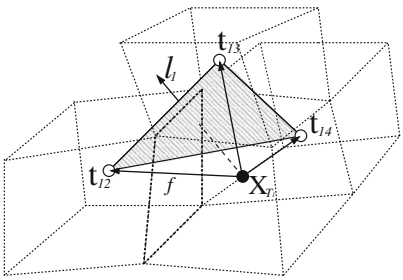


Fig. 1 Co-normal vector ℓ_f belongs to the trihedral corner formed by $\mathbf{t}_{12}, \mathbf{t}_{13}, \mathbf{t}_{14}$

The numerical diffusive flux is obtained by dropping out the term $O(|f|)$.

Setting $i = 1$, $j = 2$, $k = 3$, $l = 4$ in (7) (see Fig. 1) we obtain a numerical diffusive flux $q_f^{(1)}$ from cell T_1 to cell T_2 through their common face f. Similarly, setting $i = 2$, $j = 1$, $k = 5$, $l = 6$ in (7) and assuming that $-\ell_f$ started from $\mathbf{x}_{T_2}$ belongs to the trihedral corner formed by vectors $\mathbf{t}_{21}, \mathbf{t}_{25}, \mathbf{t}_{26}$, we obtain a different numerical flux, $q_f^{(2)}$, in the opposite direction. The final numerical flux is a linear combination of these two fluxes:

$$\begin{aligned} q_f &= \mu_1 q_f^{(1)} + \mu_2(-q_f^{(2)}) \\ &= \mu_1(\alpha_{12}(C_{T_1} - C_{T_2}) + \alpha_{13}(C_{T_1} - C_{T_3}) + \alpha_{14}(C_{T_1} - C_{T_4})) \\ &\quad - \mu_2(\alpha_{21}(C_{T_2} - C_{T_1}) + \alpha_{25}(C_{T_2} - C_{T_5}) + \alpha_{26}(C_{T_2} - C_{T_6})). \end{aligned} \tag{8}$$

In [4] the weights μ_1 and μ_2 are selected to obtain the two-point discretization. In this work they are selected to balance the relative contribution of the left and the right fluxes to the final flux. The second requirement is to approximate the true flux. These requirements lead us to the following system

$$\begin{aligned} q_f^{(1)}\mu_1 + q_f^{(2)}\mu_2 &= 0, \\ \mu_1 + \mu_2 &= 1. \end{aligned} \tag{9}$$

If $|q_f^{(1)}| + |q_f^{(2)}| = 0$, the solution of these two equations is not unique and we set $\mu_1 = \mu_2 = 1/2$. Otherwise, we have $|q_f^{(1)}| + |q_f^{(2)}| \neq 0$ and must consider two cases. In the first case $q_f^{(1)} q_f^{(2)} \leq 0$ and the solution is

$$\mu_1 = \frac{|q_f^{(2)}|}{|q_f^{(1)}| + |q_f^{(2)}|}, \quad \mu_2 = \frac{|q_f^{(1)}|}{|q_f^{(1)}| + |q_f^{(2)}|}. \tag{10}$$

Thus,

$$q_f = \frac{2q_f^{(1)}|q_f^{(2)}|}{|q_f^{(1)}| + |q_f^{(2)}|} = -\frac{2q_f^{(2)}|q_f^{(1)}|}{|q_f^{(1)}| + |q_f^{(2)}|} \tag{11}$$

and the diffusive flux has two equivalent algebraic representations:

$$\begin{aligned} q_f &= 2\mu_1(\alpha_{12}(C_{T_1} - C_{T_2}) - \alpha_{13}(C_{T_1} - C_{T_3}) - \alpha_{14}(C_{T_1} - C_{T_4})) \\ &= A_{12}(C_{T_1} - C_{T_2}) + A_{13}(C_{T_1} - C_{T_3}) + A_{14}(C_{T_1} - C_{T_4}) \end{aligned} \tag{12}$$

and

$$\begin{aligned} -q_f &= 2\mu_2(\alpha_{21}(C_{T_2} - C_{T_1}) - \alpha_{25}(C_{T_2} - C_{T_5}) - \alpha_{26}(C_{T_2} - C_{T_6})) \\ &= A_{21}(C_{T_2} - C_{T_1}) + A_{25}(C_{T_2} - C_{T_5}) + A_{26}(C_{T_2} - C_{T_6}) \end{aligned} \tag{13}$$

with non-negative coefficients A_{12}, A_{13}, A_{14}, A_{21}, A_{25} and A_{26}. Note that these coefficients depend on the fluxes and hence on the concentrations at neighboring cells. The second case $q_f^{(1)} q_f^{(2)} > 0$ leads to a potentially degenerate diffusive flux. In order to avoid this degeneracy, we re-group the terms in (8) following [11]

$$q_f = \mu_1 \tilde{q}_f^{(1)} + \mu_2(-\tilde{q}_f^{(2)}) + (\mu_1\alpha_{12} + \mu_2\alpha_{21})(C_{T_1} - C_{T_2}), \tag{14}$$

where $\tilde{q}_f^{(1)} = \alpha_{13}(C_{T_1} - C_{T_3}) + \alpha_{14}(C_{T_1} - C_{T_4})$, $\tilde{q}_f^{(2)} = \alpha_{25}(C_{T_2} - C_{T_5}) + \alpha_{26}(C_{T_2} - C_{T_6})$. The coefficients μ_1 and μ_2 are computed by balancing the modified numerical fluxes

$$\tilde{q}_f^{(1)}\mu_1 + \tilde{q}_f^{(2)}\mu_2 = 0$$

and using the convexity condition. Again, if the solution is not unique, we set $\mu_1 = \mu_2 = 1/2$. For the case $\tilde{q}_f^{(1)}\tilde{q}_f^{(2)} \leq 0$ we obtain

$$\begin{aligned} q_f &= 2\mu_1 \tilde{q}_f^{(1)} + (\mu_1\alpha_{12} + \mu_2\alpha_{21})(C_{T_1} - C_{T_2}) \\ &= A_{13}(C_{T_1} - C_{T_3}) + A_{14}(C_{T_1} - C_{T_4}) + A_{12}(C_{T_1} - C_{T_2}) \\ &= -2\mu_2 \tilde{q}_f^{(2)} - (\mu_1\alpha_{12} + \mu_2\alpha_{21})(C_{T_2} - C_{T_1}) \\ &= -A_{25}(C_{T_2} - C_{T_5}) - A_{26}(C_{T_2} - C_{T_6}) - A_{21}(C_{T_2} - C_{T_1}), \end{aligned} \tag{15}$$

where $A_{12} = A_{21} = \mu_1\alpha_{12} + \mu_2\alpha_{21}$. For the case $\tilde{q}_f^{(1)}\tilde{q}_f^{(2)} > 0$, we obtain

$$q_f = (\mu_1\alpha_{12} + \mu_2\alpha_{21})(C_{T_1} - C_{T_2}) = A_{12}(C_{T_1} - C_{T_2}). \tag{16}$$

The coefficients A_{12}, A_{13}, A_{14}, A_{21}, A_{25} and A_{26} in (15), (16) are non-negative by construction and depend on the concentrations.

We use the Dirichlet boundary data on faces $f \in \mathscr{F}_B^D$, $C_f = \int_f g_D ds/|f|$ as the known values of the concentration at points $\mathbf{x}_f$. For the Neumann boundary data on faces $f \in \mathscr{F}_B^N$ we calculate the diffusive flux as $q_f = \overline{g}_{N,f}$, where $\overline{g}_{N,f}$ is the average value of g_N on f.

3.2 Diffusive Flux in Heterogeneous Anisotropic Medium

Let us consider a heterogeneous medium. Let a face $f \in \mathscr{F}_J$ be shared by cells T_1 and T_2. We denote the plane containing f by p_f and consider a continuous piecewise linear function $\mathscr{R}(\mathbf{x})$ such that

$$\mathscr{R}(\mathbf{x}_{T_1}) = C_{T_1}, \quad \mathscr{R}(\mathbf{x}_{T_2}) = C_{T_2}, \tag{17}$$

and the diffusive flux of $\mathscr{R}(\mathbf{x})$ is continuous:

$$\mathbb{K}_{T_1} \nabla \mathscr{R}(\mathbf{x})|_{T_1} \cdot \mathbf{n}_f = \mathbb{K}_{T_2} \nabla \mathscr{R}(\mathbf{x})|_{T_2} \cdot \mathbf{n}_f. \tag{18}$$

Then, there exists a harmonic averaging point $\mathbf{y}_f \in p_f$ and a coefficient $0 \le \alpha_f \le 1$ independent of $\mathscr{R}$ such that [2]:

$$C_f \equiv \mathscr{R}(\mathbf{y}_f) = \alpha_f C_{T_1} + (1 - \alpha_f) C_{T_2}, \tag{19}$$

where

$$\alpha_f = \frac{d_{2,f} \mathbf{n}_f \cdot (\mathbb{K}_{T_1} \mathbf{n}_f)}{d_{2,f} \mathbf{n}_f \cdot (\mathbb{K}_{T_1} \mathbf{n}_f) + d_{1,f} \mathbf{n}_f \cdot (\mathbb{K}_{T_2} \mathbf{n}_f)}, \tag{20}$$

and $d_{i,f}$ is the distance from point $\mathbf{x}_{T_i}$ to plane p_f.

The scheme can be adjusted to discontinuous tensors by using harmonic averaging points. The approximation of the directional derivative $\nabla c \cdot \mathbf{t}_{ij}$ is accurate only inside each material. This limits significantly the number of admissible directions $\mathbf{t}_{ij}$ to the point that expansion (6) does not exist. The additional vectors from collocation points $\mathbf{x}_{T_i}$ and $\mathbf{x}_{T_j}$ to the harmonic point $\mathbf{y}_f$ can be used to find the expansion.

The formula for the final diffusive flux q_f involves both C_{T_i} and C_f, but the latter can be eliminated using the convex combination (19) without increasing the stencil size and preserving the DMP. For example, formula (12) is modified as follows:

$$\begin{aligned} q_f &= A_{12}(C_{T_1} - C_f) + A_{13}(C_{T_1} - C_{T_3}) + A_{14}(C_{T_1} - C_{T_4}) \\ &= A_{12}(1 - \alpha_f)(C_{T_1} - C_{T_2}) + A_{13}(C_{T_1} - C_{T_3}) + A_{14}(C_{T_1} - C_{T_4}). \end{aligned} \tag{21}$$

The other formulas are modified similarly.

3.3 Solution of the System

Let $\mathbf{C}$ be the vector of all cell-centered unknowns. Replacing the fluxes in Eq. (4) by their numerical approximations, we obtain a system of nonlinear equations

$$\mathbf{M}(\mathbf{C})\mathbf{C} = \mathbf{F}(\mathbf{C}). \tag{22}$$

with a square M-matrix **M** and a right hand side vector **F**. The entries of **M** are defined by formulas (12), (13), (15), (16) and (21) and depend on **C**. We note that coefficients from (12) and (13) are landed into the rows of **M** corresponding the cells T_1 and T_2, respectively. The matrix **M** has diagonal dominance in rows, which leads to the DMP. The system is solved by the Picard method or Anderson method [10]. The DMP holds for both the converged solution and each Picard iterate.

4 Numerical Experiments

We verify the convergence and monotonicity properties of the proposed nonlinear FV scheme with a few numerical experiments. We consider 3D benchmark problems from FVCA-6 [1] with corresponding notations.

- **Test 1: Mild anisotropy,** $c(x, y, z) = 1+\sin(\pi x)\sin\left(\pi\left(y+\frac{1}{2}\right)\right)\sin\left(\pi\left(z+\frac{1}{3}\right)\right)$, $\min = 0$, $\max = 2$, **Tetrahedral meshes (B),Voronoi meshes (C), Kershaw meshes(D), Checkerboard meshes (I)**

Mesh	i	nu	nmat	umin	umax	normg	erl2	ratiol2	ergrad	ratiograd
	2	3898	23300	0.003	1.986	1.764	6.23e−03	1.399	2.06e−01	0.872
	3	7711	44504	0.004	1.997	1.771	3.69e−03	2.303	1.67e−01	0.923
B	4	15266	86993	0.002	1.997	1.780	2.81e−03	1.197	1.31e−01	1.066
	5	30480	169809	1e−04	1.998	1.785	1.67e−03	2.258	1.06e−01	0.919
	6	61052	334864	3e−05	1.998	1.789	1.15e−03	1.611	8.66e−02	0.873
	2	66	1159	0.045	1.925	1.627	7.50e−02	−0.054	5.70e−01	1.695
	3	130	2241	0.020	1.967	1.608	3.92e−02	2.871	4.23e−01	1.320
C	4	228	3875	0.020	1.965	1.689	2.56e−02	2.275	3.13e−01	1.608
	5	356	6100	−0.002	1.991	1.689	2.05e−02	1.496	2.50e−01	1.513
	2	4096	33832	0.002	2.000	1.693	6.73e−02	0.398	5.55e−01	0.115
D	3	32768	250058	−0.002	1.996	1.723	4.95e−02	0.443	4.00e−01	0.472
	4	262144	1810432	0.003	1.997	1.761	3.02−03	0.713	2.31e−01	0.792
	2	288	3240	0.050	1.960	1.761	3.32e−02	1.060	3.16e−01	0.989
I	3	2304	23376	0.001	1.995	1.770	9.35e−03	1.828	1.37e−01	1.206
	4	18432	176544	0.002	1.998	1.789	2.79e−03	1.745	5.90e−02	1.215
	5	147456	1369920	1e−04	2.000	1.796	9.23e−04	1.596	2.72e−02	1.117

- **Test 2: Heterogeneous anisotropy,** $c(x, y, z) = x^3y^2z + x\sin(2\pi xz)\sin(2\pi xy)\sin(2\pi z)$, $\min = -0.862$, $\max = 1.0487$, **Prism meshes**

i	nu	nmat	umin	umax	normg	erl2	ratiol2	ergrad	ratiograd
1	1210	14275	−1.006	1.006	3.035	3.36e−02		3.14e−01	
2	8820	92696	−0.971	0.971	3.388	8.29e−03	2.114	1.17e−01	1.491
3	28830	289232	−1.000	1.000	3.492	3.68e−03	2.057	6.07e−02	1.662
4	67240	658039	−0.998	0.998	3.534	2.07e−03	2.038	3.67e−02	1.782

The differential problems of Test 1 and Test 2 do not satisfy the maximum principle since the exact solution has local extrema. Therefore, no numerical scheme can guarantee DMP.

- **Test 4: Flow around a well**, $\min = 0,\ \max = 5.415$, **Well meshes**

i	nu	nmat	umin	umax	normg	erl2	ratiol2	ergrad	ratiograd
3	5016	40585	0.172	5.329	1581.870	1.92e−03	2.765	7.22e−02	2.207
4	11220	87248	0.128	5.330	1603.979	1.18e−03	1.814	4.79e−02	1.529
5	23210	175975	0.097	5.339	1612.048	6.86e−04	2.239	3.20e−02	1.665
6	42633	318146	0.075	5.345	1615.236	4.78e−04	1.782	2.21e−02	1.826
7	74679	551433	0.058	5.361	1617.424	3.39e−04	1.839	1.69e−02	1.436

- **Test 5: Discontinuous permeability,** $c(x, y, z) = a_i \sin(2\pi x) \sin(2\pi y) \sin(2\pi z)$, $\min = -100,\ \max = 100$, **Locally refined meshes**

i	nu	nmat	umin	umax	normg	erl2	ratiol2	ergrad	ratiograd
1	22	124	−209.045	209.045	442.542	1.09e+00		1.00e+00	
2	176	1112	−43.618	43.618	58.442	2.23e−01	2.289	1.80e+00	−0.848
3	1408	9376	−83.042	83.042	89.814	5.76e−02	1.953	3.16e−01	2.510
4	11264	76928	−95.567	95.567	97.224	1.36e−02	2.082	1.53e−01	1.046

The proposed 3D nonlinear FV scheme for the diffusion equation satisfies the discrete maximum principle and has a compact stencil. The scheme provides asymptotic second order accuracy for concentrations except for extremely irregular Kershaw meshes.

Acknowledgments This work has been supported in part by RFBR grants 12-01-33084, 14-01-00830, Russian Presidential grant MK-7159.2013.1, Federal target programs of Russian Ministry of Education and Science, ExxonMobil Upstream Research Company, and project "Breakthrough" of Rosatom.

References

1. FVCA6 3D Benchmark. http://www.latp.univ-mrs.fr/latp_numerique/?q=node/4
2. Agelas, L., Eymard, R., Herbin, R.: A nine-point finite volume scheme for the simulation of diffusion in heterogeneous media. C. R. Acad. Sci. Paris Ser. I. **347**, 673–676 (2009)
3. Chernyshenko, A.: Generation of adaptive polyhedral meshes and numerical solution of 2nd order elliptic equations in 3D domains and on surfaces. Ph.D. thesis, INM RAS, Moscow (2013).
4. Danilov, A., Vassilevski, Y.: A monotone nonlinear finite volume method for diffusion equations on conformal polyhedral meshes. Russ. J. Numer. Anal. Math. Model. **24**(3), 207–227 (2009)
5. Droniou, J.: Finite volume schemes for diffusion equations: introduction to and review of modern methods. Math. Models Methods Appl. Sci. (2014). To appear
6. Gao, Z.M., Wu, J.M.: A small stencil and extremum-preserving scheme for anisotropic diffusion problems on arbitrary 2d and 3d meshes. J. Comp. Phys. **250**, 308–331 (2013)

7. LePotier, C.: Schema volumes finis monotone pour des operateurs de diffusion fortement anisotropes sur des maillages de triangle non structures. C. R. Acad. Sci. Paris Ser. I. **341**, 787–792 (2005)
8. Lipnikov, K., Svyatskiy, D., Shashkov, M., Vassilevski, Y.: Monotone finite volume schemes for diffusion equations on unstructured triangular and shape-regular polygonal meshes. J. Comp. Phys. **227**, 492–512 (2007)
9. Lipnikov, K., Svyatskiy, D., Vassilevski, Y.: Minimal stencil finite volume scheme with the discrete maximum principle. Russ. J. Numer. Anal. Math. Model. **27**(4), 369–385 (2012)
10. Lipnikov, K., Svyatskiy, D., Vassilevski, Y.: Anderson acceleration for nonlinear finite volume scheme for advection-diffusion problems. SIAM J. Sci. Comput. **35**(2), 1120–1136 (2013)
11. Yuan, G., Sheng, Z.: The finite volume scheme preserving extremum principle for diffusion equations on polygonal meshes. J. Comp. Phys. **230**(7), 2588–2604 (2011)

Continuous Finite-Elements on Non-Conforming Grids Using Discontinuous Galerkin Stabilization

Andreas Dedner, Robert Klöfkorn and Mirko Kränkel

Abstract In this paper we present a new idea to stabilize Continuous Galerkin Schemes on grids with hanging nodes by using Discontinuous Galerkin (DG) approximations. We derive an a posteriori error estimate for a class of DG schemes including the CDG and CDG2 methods and apply this to standard test cases for CG methods such a the reentrant corner.

1 Introduction

Adaptive conforming Finite Element (FE) discretizations have been successfully utilized for nearly 3 decades now [6]. For some applications or for parallelism it might be favorable to use non-conforming mesh refinement. This conflicts with the use of conforming FE and various techniques to overcome this issue have been developed. Conforming mesh refinement techniques can be used. For example bisection of simplicial grids [7] but especially in 3d this approach is extremely cumbersome. A further approach is red-green refinement (e.g. [3]) using non-conforming refinement in general and resolving hanging nodes by so called green closure. An issue is that starting with a grid containing only cube elements, the resulting grids become hybrid and in 3D will contain pyramid type elements which are difficult to handle

A. Dedner
University of Warwick, Coventry, UK
e-mail: a.s.dedner@warwick.ac.uk

R. Klöfkorn (✉)
National Center for Atmospheric Research, 1850 Table Mesa Drive,
Boulder, CO 80305, USA
e-mail: robertk@ucar.edu

M. Kränkel
University of Freiburg, Freiburg im Breisgau, Germany
e-mail: kraenkel@mathematik.uni-freiburg.de

J. Fuhrmann et al. (eds.), *Finite Volumes for Complex Applications VII - Methods and Theoretical Aspects*, Springer Proceedings in Mathematics & Statistics 77,
DOI: 10.1007/978-3-319-05684-5_19,

within an FE discretization. Also both approaches can create elements with small angles degrading the quality of the resulting FE scheme. Another strategy to resolve hanging nodes is to add constraints to the system that relate a hanging node to its surrounding nodes (e.g. [2]). A benefit of this technique is that it can also be used with hp-refinement where hanging nodes appear, even if the grid is conforming. But the implementation of this approach especially for higher order elements on general non-conforming meshes is quite challenging.

In this paper we present another strategy to deal with hanging nodes on non-conforming grids or hp-refinement. We will alter the standard FE discretization such that for hanging nodes we switch to a Discontinuous Galerkin (DG) discretization which works on general unstructured grids. We compare the conforming FE-DG method with a conforming FE method used on red-green refined grids as well as a regular DG method. One advantage of this approach is the ease with which existing FE discretizations can be extended to work on non-conforming grids while avoiding the increase in the number of DoFs associated with the DG method. In fact since we are avoiding closure our approach can even lead to fewer DoFs compared to a standard FE approach using for example red-green closure. We will demonstrate the effectiveness of the approach based on 2d and 3d Poisson problems and a Stokes problem. To construct the grid we first present an a-posteriori error estimate which can be applied to a very general class of DG discretizations. We conclude with numerical results based on the CDG2 method introduced in [4].

2 Discontinuous Galerkin Discretization

2.1 Elliptic problem

To derive the discontinuous Galerkin method we consider the following elliptic problem in $\mathbb{R}^d$, $d = 2$ of the form

$$-\nabla \cdot (K(x)\nabla u(x)) = f(x)\,, \quad x \in \Omega\,, \qquad u = g_D\,, \quad \text{on } \partial\Omega, \tag{1}$$

where $\Omega \subset \mathbb{R}^d$ is a bounded polygonal area, $K \in W^1_\infty(\Omega)$ a positive definite diffusion matrix, and $f \in L^2(\Omega)$. Given the weak formulation $a(u,\varphi) := \int_\Omega K\nabla u \cdot \nabla\varphi = \int_\Omega f\varphi$ we are interested in a discrete *primal formulation* of the form

$$B(u_h,\varphi) = \int_\Omega f\varphi \qquad \forall\,\varphi \in V_h\,. \tag{2}$$

The discrete solution u_h is in the piecewise polynomial space $V_h = V_h^1$ with (with $l \in \mathbb{N}$) $V_{\mathscr{G}}^l = \{\boldsymbol{v} \in L^2(\Omega,\mathbb{R}^l) : \boldsymbol{v}|_E \in [\mathscr{P}_k(E)]^l\}$ defined for a given partition $\mathscr{G} = \{E\}$ of Ω into polygons E. We also use the abbreviation $\Sigma_{\mathscr{G}} = V_h^d$ in the

following. For later a-posteriori analysis we also need the space $V(\mathscr{G})^l := [H^1(\Omega)]^l + V^l_{\mathscr{G}}$ where we use the abbreviations $V(\mathscr{G}) = V(\mathscr{G})^1$ and $\Sigma(\mathscr{G}) = V(\mathscr{G})^d$.

To derive the bilinear form B we need to introduce some standard notation (see [1]). By Γ_i we denote the family of all interior intersections e of grid elements $E_e^+, E_e^- \in \mathscr{G}$ where $e = E_e^- \cap E_e^+$ and positive measure in $\mathbb{R}^{d-1}$. Let Γ be the family of all edges $e \subset \partial E$, where $E \in \mathscr{G}$. For each intersection e we define the local mesh width $h_e = \frac{\min\{|E_e^-|,|E_e^+|\}}{|e|}$. Let $\phi \in V(\mathscr{G})$ and $\boldsymbol{\tau} \in \Sigma(\mathscr{G})$ then for $e \in \Gamma_i$ we introduce jump $[\![\cdot]\!]_e$, $[\![\![\cdot]\!]\!]_e$, and average operators $\{\cdot\}_e$, $\{\{\cdot\}\}_e$ in the usual way (see [1, 4]).

Following the derivation found in [1] we obtain, for given numerical fluxes $\widehat{u}$ and $\widehat{K}$, both mapping u_h to $[L^2(\Gamma)]^d$, the flux based bilinear form:

$$B(u_h,\varphi) := \sum_E \int_E (K\nabla u_h)\cdot\nabla\varphi - \sum_{e\in\Gamma}\int_e \left(\{\{K\nabla\varphi\}\}_e \cdot [\![\![u_h]\!]\!]_e + \widehat{K}(u_h,\nabla u_h)\cdot[\![\![\varphi]\!]\!]_e\right) + \sum_{e\in\Gamma_i}\int_e [K\nabla\varphi]_e\{\widehat{u}(u_h)-u_h\}_e. \tag{3}$$

The method is completely described once the physical parameter functions K and f are known and appropriate numerical fluxes have been chosen. To define the numerical diffusion fluxes, let us define two kinds of *lifting operators* $r_e : [L^2(e)]^d \to \Sigma_{\mathscr{G}}$ and $l_e : L^2(e) \to \Sigma_{\mathscr{G}}$, for every $e \in \Gamma$, with

$$\int_\Omega r_e(\boldsymbol{\xi})\cdot K\boldsymbol{\tau} = -\int_e \boldsymbol{\xi}\cdot\{\{K\boldsymbol{\tau}\}\}_e, \quad \int_\Omega l_e(\phi)\cdot K\boldsymbol{\tau} = -\int_e \phi[K\boldsymbol{\tau}]_e, \tag{4}$$

for all $\boldsymbol{\tau} \in \Sigma_{\mathscr{G}}$, $\boldsymbol{\xi} \in [L^2(e)]^d$, and $\phi \in L^2(e)$.

For our convenience we define $L_e(u) := r_e([\![\![u]\!]\!]_e) + l_e(\beta_e \cdot [\![\![u]\!]\!]_e)$ on $e \in \Gamma$. The parameter $\boldsymbol{\beta}$ (in the literature frequently $\mathbf{C}_{12}$) is called the *switch function*.

We consider the Compact Discontinuous Galerkin 2 (CDG2) method [4], where $\widehat{u}(u) = \{u\}_e$ and $\widehat{K}(u,\nabla_h u) = \{\{K\nabla_h u\}\}_e + 2\chi_e\big(Kr_e([\![\![u]\!]\!]_e)\big)|_{E_e^-}$. However, the following a-posteriori error estimate does not depend on the choice of the scheme and also works for the other methods presented in [4].

2.2 Stokes Problem

We consider the Stokes problem in $\mathbb{R}^d$, $d = 2$:

$$\begin{aligned} -\nabla\cdot(K(x)\nabla u(x)) + \nabla p(x) &= f(x) \quad x\in\Omega,\\ \nabla\cdot u(x) &= 0 \quad x\in\Omega,\\ u &= g_D \quad \text{on } \partial\Omega, \end{aligned}$$

and DG discretizations of the form

$$B(u_h, p_h; v_h, q_h) := B(u_h, v_h) + D(p_h, v_h) + \tilde{D}(q_h, u_h)$$
$$= \int_\Omega f(x) \cdot v_h \quad (v_h, q_h) \in V_h^d \times Q_h$$

with solutions $(u_h, p_h) \in V_h^d \times Q_h$ of $B(u_h, v_h)$ is defined as for the elliptic problem taking into account that u_h has values in $\mathbb{R}^d$ and the bilinear forms $D, \tilde{D} : V_h \times Q_h \to \mathbb{R}$ are chosen for example using IP fluxes [5].

3 A Posteriori Error Estimate

3.1 Elliptic Problem

Consider a bilinear form on $V_h \times V_h$ given by

$$B(u_h, \varphi) = \int_\Omega K\sigma_h(u_h) \cdot \nabla_h \varphi - \sum_{e \in \Gamma} \int_e \widehat{K}(u_h, \nabla_h u_h) \cdot [\![\varphi]\!]_e$$

where we used the abbreviation

$$\boldsymbol{\sigma}_h(u_h) := \nabla_h u_h + \sum_{e \in \Gamma} \left(r_e([\![u_h]\!]_e) + \beta_1 l_e(\beta_e \cdot [\![u_h]\!]_e) \right). \tag{5}$$

with suitable values for β_1, β_e depending on the scheme considered. Using the lifting operators this can be extended to the space $V(\mathscr{G}) \times V(\mathscr{G})$:

$$A(u, \varphi) := \int_\Omega K\sigma_h(u) \cdot \nabla_h \varphi + \sum_{e \in \Gamma} \int_\Omega \left(K \nabla_h u \right) \cdot \left(r_e([\![\varphi]\!]_e) \right.$$
$$\left. + \beta_1 l_e(\beta_e \cdot [\![\varphi]\!]_e) \right) - \sum_{e \in \Gamma} \int_e \boldsymbol{\delta}_e(u) \cdot [\![\varphi]\!]_e \tag{6}$$

where $\boldsymbol{\delta}_e(u)$ is the stabilization term, i.e. all terms from $\widehat{K}$ not involving $\nabla_h u$. For example for CDG2 $\boldsymbol{\delta}_e(u) := \chi_e\big(\{\{L_e(u)\}\}_e + \beta_e [L_e(u)]_e\big)$, [4].

The bilinear form A from (6) can be evaluated on the space $V(\mathscr{G}) \times V(\mathscr{G})$ and satisfies $A(v, w) = a(v, w)$ on $H^1 \times H^1$ since $r_e([\![v]\!]_e)$ and $l_e(\beta_e \cdot [\![v]\!]_e)$ vanish as does $[\![v]\!]_e$ which holds for $v, w \in H^1$, respectively. Furthermore, for $u \in V_{\mathscr{G}}$ and $\varphi \in V(\mathscr{G})$ we find $A(u, \varphi) = B(u, \varphi)$.

Using the perturbed bilinear form A we start with the dual problem with a given functional J that is defined for $v \in V(\mathscr{G})$ and $J|_{H^1} \in H^{-1}$ such that the dual problem

$$A(v, z) = J(v) \quad \text{for all } v \in H^1$$

has a solution $z \in H^1$ (note that also $a(v, z) = J(v)$).

Following [5] we decompose the error $e = u - u_h$ using $u_h = u_h^c + u_h^\perp$ with $u_h^c \in H^1$. Thus $\varepsilon^c = u - u_h^c \in H^1$ and with a few simple steps we arrive at

$$J(\varepsilon) = \int_\Omega f(z - z_h) - A(u_h, z - z_h) + A(u_h^\perp, z) - J(u_h^\perp) =: \langle R, z - z_h \rangle + \langle R^\perp, z \rangle \, .$$

Remark 1 Note that due to the definition of the dual solution and the decomposition of u_h we have $\langle R^\perp, z \rangle = A(u_h^\perp, z) - J(u_h^\perp) = A(u_h, z) - J(u_h)$.

Lemma 1 (Estimate for Error) *Define for all elements $E \in \mathscr{G}$ the local grid width $h_E := |E|^{\frac{1}{d}}$ and $R_{E,1}(u_h) := \sum_{e \subset \partial E} \| \big(\widehat{K}(u_h, \nabla_h u_h) - (K\sigma_h(u_h))|_E \big) \cdot n_e \|_e^2$ $R_{E,2}(u_h) := \| f + \nabla \cdot K\sigma_h(u_h) \|_E^2$, $R_{E,\perp}(u_h) := \sum_{e \subset \partial E} h_e^{-1} \| [\![u_h]\!]_e \|_e^2$.*

Then we can bound the error $\varepsilon := u - u_h$ $\|\varepsilon\|_{DG} \leq C \eta_h(u_h)$ with

$$\eta_h(u_h) = \Big\{ \sum_{E \in \mathscr{G}} \Big(h_E^2 R_{E,2}(u_h) + h_E R_{E,1}(u_h) + R_{E,\perp}(u_h) \Big) \Big\}^{\frac{1}{2}} \, .$$

The DG norm is given by

$$|[\![v]\!]|_\Gamma^2 := \sum_{e \in \Gamma} h_e^{-1} \| [\![v]\!]_e \|_e^2, \quad |v|_1^2 := \sum_{E \in \mathscr{G}} |v|_{1,E}^2, \quad \|v\|_{DG}^2 := |v|_1^2 + |[\![v]\!]|_\Gamma^2. \quad (7)$$

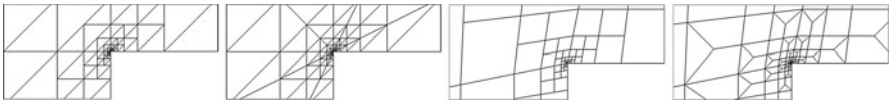
Proof First one has to bound the residual R for the perturbed bilinear form using the standard integration by parts formula on DG spaces []. $|\langle R, v \rangle| \leq \big| \sum_{E \in \mathscr{G}} R_E(v) \big|$ with

$$R_E(v) := \| f + \nabla \cdot K\sigma_h(u_h) \|_E \, \|v\|_E + \sum_{e \subset \partial E} \big\| \big(\widehat{K}(u_h, \nabla_h u_h) - (K\sigma_h(u_h))|_E \big) \cdot n_E \big\|_e \, \|v_{|E}\|_e$$

for all $v \in V(\mathscr{G})$. Next consider the functional

$$J(v) := \frac{1}{\|\varepsilon\|_{DG}} \left(\int_\Omega \nabla_h \varepsilon \cdot \nabla_h v + \sum_{e \in \Gamma} h_e^{-1} \int_e [\![\varepsilon]\!]_e \cdot [\![v]\!]_e \right). \quad (8)$$

Using boundedness estimates from [1] for the integral terms in (8), stability estimates for the dual solution z and interpolation estimates one arrives at the stated estimate.



(a) lag-red **(b)** lag-green **(c)** lag-red **(d)** lag-green

Fig. 1 Refined grids after 9 iterations algorithm for different versions of the discrete spaces using $k = 4$

3.2 The Stokes problem

Using the same ideas outlined in the previous section, we can derive an a-posteriori estimate for quite general DG methods for the Stokes equation, which we simply state in the following:

Lemma 2 (Estimate for Error) *Define for all* $E \in \mathscr{G}$

$$R_{E,2} := \|f + \nabla \cdot (K\sigma_h - p_h \mathbb{I})\|_E^2 . \qquad R_{E,1} := \sum_{e \in \partial E} \|(\widehat{K}_p - (K\sigma_h - p_h \mathbb{I})|_E) \cdot n_E\|_e^2,$$

$$R_{E,\perp} := \sum_{e \in \partial E} h_e^{-1} \|[\![u_h]\!]_e\|_e^2 . \qquad R_{E,div} := \|\nabla \cdot u_h\|_E^2 .$$

Then using $\eta_E^2 := h_E^2 R_{E,2} + h_E R_{E,1} + R_{E,\perp} + R_{E,div}$ *the error is bounded by*

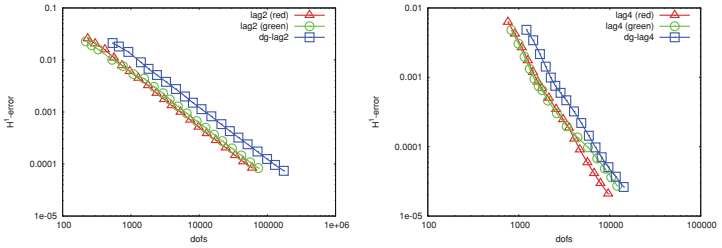
$$\|(\varepsilon_u, \varepsilon_p)\|_{DG} \leq C\eta_h \quad \textit{with} \quad \eta_h = \Big(\sum_E \eta_E^2\Big)^{\frac{1}{2}} .$$

4 Numerical Results

4.1 The L-Shaped Laplace Problem

We use the 270° reentrant corner problem to compare three approaches: full DG (dg-lag), DG with conforming Lagrange basis functions (lag-red), and continuous FE based on red-green closure (lag-green). Marking of elements for refinement is performed in the same way in all cases based on our a-posteriori estimator. Examples of the refinement near the corner on simplex and cube grids are shown in Fig. 1.

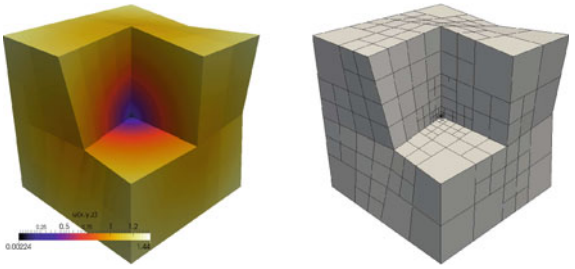
For a first comparison of the different methods we compare in this paper only the number of degrees of freedom (DOFs) versus the H^1-error for polynomial degrees $k = 2$ and $k = 4$ (see Fig. 2). It can be clearly seen that the method using a conforming Lagrange space with DG closure (lag-red) is the most efficient approach, especially in the case of quadrilaterals requiring about 60 % of the number of DOFs compared to using green closure to reach the same H^1-error.



(a) H^1-error per DOFs ($k = 2$) **(b)** H^1-error per DOFs ($k = 4$)

(c) H^1-error per DOFs ($k = 2$) **(d)** H^1-error per DOFs ($k = 4$)

Fig. 2 For polynomial orders $k = 2, 4$ the H^1-error per number of DOFs on triangular meshes (*top*) and perturbed quadrilateral meshes is shown

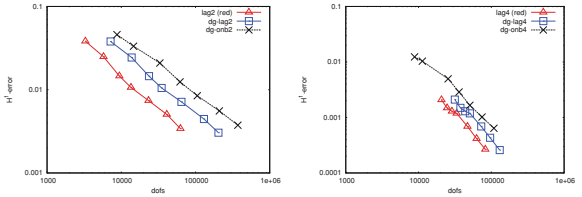


(a) u_h ($k = 4$) **(b)** grid

Fig. 3 *Left*, the computed solution u_h at iteration 11 of the adaptive algorithm. *Right*, the corresponding refined hexahedral grid with non-affine geometry mapping

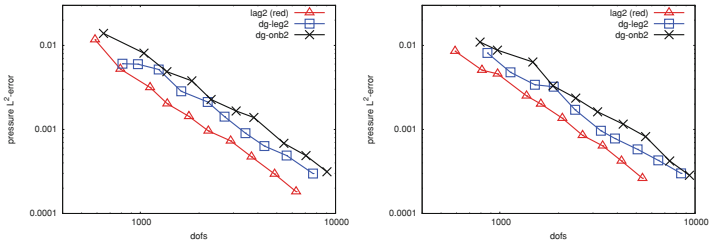
4.2 3d Problem: The Fichera Corner

We now study the Fichera corner problem in 3d comparing the dg and the lag-red method which both work on non conforming grids. We also added a DG approach using orthonormalized monomial basis functions reducing the number of DoFs per element required compared to the DG approach with Lagrange shape functions. Again the gradient of the solution has a singularity in $x = 0$. In Fig. 3 the numerical solution and the adaptive grid after 11 refinement cycles is presented. Here, the vertices of the domain have been perturbed resulting in hexahedrons with a non-affine geometry mapping.



(a) H^1 error ($k = 2$) **(b)** H^1 error ($k = 4$)

Fig. 4 For polynomial orders $k = 2, 4$ the H^1-error per number of DOFs is plotted



(a) L^2-error per DOFs ($k = 2$, affine) **(b)** L^2-error per DOFs ($k = 2$, nonaffine)

Fig. 5 For polynomial orders $k = 2, 4$ the L^2-error of the pressure per number of DOFs on quadrilateral (*left*) and perturbed quadrilateral meshes (*right*) is shown

In Fig. 4 we present the comparison of the different schemes for $k = 2, 4$ in terms of DOFs versus H^1-error. Using the orthonormal basis (ONB) does not lead to the desired improvement, since on non-affine grids the method does not converge optimaly. The DG approach using the conforming Lagrange basis functions (lag-red) again seems to be the best approach (Fig. 5).

4.3 The Stokes Problem

We present brief results for the Stokes problem in [5, Example 2] and compare the L^2-error for the pressure for three approaches: full DG with Legrendre basis (dg-leg), full DG with orthonormal basis (dg-onb), and DG with conforming Lagrange basis functions (lag (red)). The results confirm the tendencies we observed for the Poisson equation.

References

1. Arnold, D., Brezzi, F., Cockburn, B., Marini, L.: Unified analysis of discontinuous Galerkin methods for elliptic problems. SIAM J. Numer. Anal. **39**(5), 1749–1779 (2002)
2. Bangerth, W., Hartmann, R., Kanschat, G.: deal.II—a general purpose object oriented finite element library. ACM Trans. Math. Softw. **33**(4), 24/1–24/27 (2007)
3. Bastian, P., Birken, K., Johannsen, K., Lang, S., Neuss, N., Rentz-Reichert, H., Wieners, C.: UG—a flexible software toolbox for solving partial differential equations. Comput. Vis. Sci. **1**, 27–40 (1997)
4. Brdar, S., Dedner, A., Klöfkorn, R.: Compact and stable Discontinuous Galerkin methods for convection-diffusion problems. SIAM J. Sci. Comput. **34**(1), 263–282. http://dx.doi.org/10.1137/100817528 (2012)
5. Houston, P., Schötzau, D., Wihler, T.: Energy norm a posteriori error estimation for mixed discontinuous Galerkin approximations of the Stokes problem. J. Sci. Comput. **22–23**, 347–370 (2005)
6. Johnson, C.: Numerical solution of partial differential equations by the finite element method. Acta Applicandae Mathematica **18**(2), 184–186 (1990)
7. Schmidt, A., Siebert, K.: Design of Adaptive Finite Element Software—the Finite Element Toolbox ALBERTA. Springer, Heidelberg (2005)

A Well-Balanced Scheme for the Euler Equation with a Gravitational Potential

Vivien Desveaux , Markus Zenk , Christophe Berthon and Christian Klingenberg

Abstract The aim of this work is to derive a well-balanced numerical scheme to approximate the solutions of the Euler equations with a gravitational potential. This system admits an infinity of steady state solutions which are not all known in an explicit way. Among all these solutions, the hydrostatic atmosphere has a special physical interest. We develop an approximate Riemann solver using the formalism of Harten, Lax and van Leer, which takes into account the source term. The resulting numerical scheme is proven to be robust, to preserve exactly the hydrostatic atmosphere and to preserve an approximation of all the other steady state solutions.

1 Introduction

We consider the Euler equations with a gravity source term

V. Desveaux
Maison de la Simulation, CEA Saclay, USR 3441, bât. 565,
91191 Gif-sur-Yvette cedex, France
e-mail: vivien.desveaux@univ-nantes.fr

M. Zenk (✉) · C. Klingenberg
Universität Würzburg, Campus Hubland Nord, Emil-Fischer-Strasse 30,
97074 Würzburg, Germany
e-mail: markus.zenk@gmx.de

C. Klingenberg
e-mail: klingenberg@mathematik.uni-wuerzburg.de

C. Berthon
Laboratoire de Mathématiques Jean Leray, UMR 6629, 2 rue de la Houssinière,
BP 92208 - 44322 Nantes Cedex 3, France
e-mail: Christophe.Berthon@univ-nantes.fr

J. Fuhrmann et al. (eds.), *Finite Volumes for Complex Applications VII - Methods and Theoretical Aspects,* Springer Proceedings in Mathematics & Statistics 77,
DOI: 10.1007/978-3-319-05684-5_20,

$$\begin{cases} \partial_t \rho + \partial_x \rho u = 0, \\ \partial_t \rho u + \partial_x \left(\rho u^2 + p\right) = -\rho \partial_x \phi, \\ \partial_t E + \partial_x (u(E + p)) = -\rho u \partial_x \phi, \end{cases} \tag{1}$$

where $\rho > 0$ denotes the density, $u \in \mathbb{R}$ the velocity, $E > 0$ the total energy and $p > 0$ the pressure. We assume the system is closed by the ideal gas law

$$p = (\gamma - 1)(E - \rho u^2/2), \quad \text{with } \gamma \in (1, 3].$$

Concerning the gravity source term, we assume it derives from a gravitational potential $\phi(x)$, which is a given smooth function. The unknown vector $w = (\rho, \rho u, E)^T$ is assumed to belong to the set of physical admissible states

$$\Omega = \left\{ w \in \mathbb{R}^3 : \quad \rho > 0, \quad E - \rho u^2/2 > 0 \right\}.$$

Following the arguments stated in [9], when dealing with simulations of near equlibrium states of (1), well-balanced numerical schemes are expected to perform better than fractional splitting methods. It means that they should accurately capture the steady state solutions of the system, which is not neccesaraly true for general splitting methods. For the Euler equations with gravity, the steady state solutions at rest are characterized as follows:

$$u = 0, \qquad \partial_x p = -\rho \partial_x \phi. \tag{2}$$

We can exhibit a specific steady state solution of (1) which is of particular physical interest, namely the hydrostatic atmosphere defined for $\alpha > 0$ and $\beta > 0$ by

$$u(x) = 0, \qquad \rho(x) = \alpha e^{-\beta \phi(x)}, \qquad p(x) = \frac{\alpha}{\beta} e^{-\beta \phi(x)}. \tag{3}$$

In the well-known shallow-water model, the lake at rest is the unique steady state at rest (up to a constant) and it finds an explicit definition. In the last decade, numerous numerical schemes were developed to preserve the lake at rest in the shallow-water equations. The reader is referred for instance to [1, 5, 6].

For the Euler equations with gravity (1), the main discrepancy lies in the fact there are an infinity of solutions of (2) and we cannot explicit all of them. Therefore it is very difficult to derive numerical schemes which accurately capture all the solutions of (2). In a recent work [2], Chalons et al. succeeded to do so, but only in the case of a constant gravity field. We also mention the work of Käppeli and Mishra [8] where they manage to preserve all the isentropic solutions of (2).

Our aim is thus to derive a numerical scheme which captures exactly the hydrostatic atmosphere (3) and which preserves approximately all the solutions of (2). To address such an issue, we propose to build an approximate Riemann solver, following

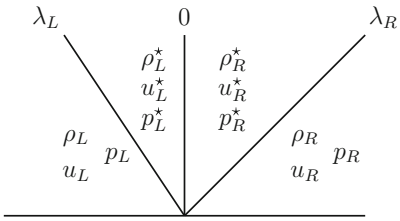


Fig. 1 Structure of the approximate Riemann solver $\widetilde{W}(x/t, w_L, w_R)$

the formalism of Harten, Lax and van Leer [7] and the extensions introduced by Gallice [4].

The paper is organized as follows. Section 2 is devoted to the derivation of a simple approximate Riemann solver which takes into account the definition of the steady states (2). In Sect. 3, we present the associated numerical scheme and we establish that it is positive preserving and well-balanced, since it preserves exactly the steady state (3). The relevance of this approach is illustrated in Section 4 with some numerical experiments.

2 The Approximate Riemann Solver

We now derive an approximate Riemann solver $\widetilde{W}(x/t, w_L, w_R)$ made of three waves with speeds $\lambda_L, 0$ and λ_R separating two intermediate states $w_L^\star$ and $w_R^\star$ (see Fig. 1). In order to enforce enough numerical viscosity, these speeds are assumed to satisfy $\lambda_L < 0 < \lambda_R$. As a consequence, this approximate Riemann solver will be fully characterized as soon as the intermediate values $\rho_{L,R}^\star$, $u_{L,R}^\star$ and $p_{L,R}^\star$ are given suitable definitions.

According to the work by Harten, Lax and van Leer [7], the approximate solver must satisfy the consistency relation

$$\frac{1}{\Delta x}\int_{-\Delta x/2}^{\Delta x/2} \widetilde{W}\left(\frac{x}{\Delta t}, w_L, w_R\right) dx = \frac{1}{\Delta x}\int_{-\Delta x/2}^{\Delta x/2} W_{\mathscr{R}}\left(\frac{x}{\Delta t}, w_L, w_R\right) dx, \qquad (4)$$

where $W_{\mathscr{R}}(x/t, w_L, w_R)$ denotes the exact solution of the Riemann problem for (1). If the CFL restriction $\frac{\Delta t}{\Delta x}\max(|\lambda_L|, |\lambda_R|) \leq \frac{1}{2}$ is satisfied, we can compute the average of the approximate Riemann solver $\widetilde{W}$ to get an equivalent formulation to (4):

$$\left(\frac{1}{2}+\lambda_L\frac{\Delta t}{\Delta x}\right)w_L-\lambda_L\frac{\Delta t}{\Delta x}w_L^\star+\lambda_R\frac{\Delta t}{\Delta x}w_R^\star+\left(\frac{1}{2}-\lambda_R\frac{\Delta t}{\Delta x}\right)w_R = \frac{1}{\Delta x}\int_{-\Delta x/2}^{\Delta x/2} W_{\mathscr{R}}\left(\frac{x}{\Delta t},w_L,w_R\right)dx. \tag{5}$$

First, we deal with the momentum equation by integrating the momentum component of the Riemann solution $W_{\mathscr{R}}^{\rho u}$ associated to (1). Provided that the wave velocities involved in the exact Riemann solution $W_{\mathscr{R}}$ stay within (λ_L,λ_R), we get

$$\frac{1}{\Delta x}\int_{-\Delta x/2}^{\Delta x/2}(\rho u)_{\mathscr{R}}\left(\frac{x}{\Delta t},w_L,w_R\right)dx = \frac{\rho_L u_L+\rho_R u_R}{2}-\frac{\Delta t}{\Delta x}(\rho_R u_R^2+p_R - \rho_L u_L^2-p_L)-\frac{1}{\Delta x}\int_{-\Delta x/2}^{\Delta x/2}\int_0^{\Delta t}\rho_{\mathscr{R}}\left(\frac{x}{t},w_L,w_R\right)\partial_x\phi dt dx. \tag{6}$$

For the sake of simplicity in the notations, we set

$$\widehat{q}=\frac{\lambda_R\rho_R u_R-\lambda_L\rho_L u_L}{\lambda_R-\lambda_L}-\frac{1}{\lambda_R-\lambda_L}(\rho_R u_R^2+p_R-\rho_L u_L^2-p_L).$$

Plugging (6) into relation (5) gives

$$\frac{\lambda_R\rho_R^\star u_R^\star-\lambda_L\rho_L^\star u_L^\star}{\lambda_R-\lambda_L}=\widehat{q}-\frac{1}{(\lambda_R-\lambda_L)\Delta t}\int_{-\Delta x/2}^{\Delta x/2}\int_0^{\Delta t}\rho_{\mathscr{R}}\left(\frac{x}{t},w_L,w_R\right)\partial_x\phi dt dx$$

The integral of the source term is usually difficult to compute exactly, so we propose the following approximation:

$$\frac{1}{\Delta t}\int_{-\Delta x/2}^{\Delta x/2}\int_0^{\Delta t}\rho_{\mathscr{R}}\left(\frac{x}{t},w_L,w_R\right)\partial_x\phi dt dx\approx\overline{\rho}(\phi_R-\phi_L). \tag{7}$$

Here, $\overline{\rho}$ represents an average between ρ_L and ρ_R that will be defined later in order to preserve the steady states. Finally, we get the equation

$$\frac{\lambda_R\rho_R^\star u_R^\star-\lambda_L\rho_L^\star u_L^\star}{\lambda_R-\lambda_L}=\widehat{q}-\frac{1}{(\lambda_R-\lambda_L)}\overline{\rho}(\phi_R-\phi_L). \tag{8}$$

We adopt the same strategy for the total energy. We introduce the intermediate total energy as follows:

$$\widehat{E} = \frac{\lambda_R E_R - \lambda_L E_L}{\lambda_R - \lambda_L} - \frac{1}{\lambda_R - \lambda_L}(u_R(E_R + p_R) - u_L(E_L + p_L)).$$

Then, an integration of the E-component of the Riemann solution associated to (1) leads to the following relation:

$$\frac{\lambda_R E_R^\star - \lambda_L E_L^\star}{\lambda_R - \lambda_L} = \widehat{E} - \frac{1}{(\lambda_R - \lambda_L)\Delta t} \int_{-\Delta x/2}^{\Delta x/2} \int_0^{\Delta t} (\rho u)_{\mathscr{R}} \left(\frac{x}{t}, w_L, w_R\right) \partial_x \phi dt dx.$$

According to (7), we approximate the integral of the source term by

$$\frac{1}{\Delta t} \int_{-\Delta x/2}^{\Delta x/2} \int_0^{\Delta t} (\rho u)_{\mathscr{R}} \left(\frac{x}{t}, w_L, w_R\right) \partial_x \phi dt dx \approx \overline{\rho} \frac{u_L + u_R}{2} (\phi_R - \phi_L). \tag{9}$$

It is worth noticing that we could replace $(u_L + u_R)/2$ by any consistent average between u_L and u_R, like we did for ρ. However this choice will not intervene into the preservation of the steady states, so for the sake of simplicity, we use the arithmetic mean value. We finally obtain the equation

$$\frac{\lambda_R E_R^\star - \lambda_L E_L^\star}{\lambda_R - \lambda_L} = \widehat{E} - \frac{1}{(\lambda_R - \lambda_L)} \overline{\rho} \frac{u_L + u_R}{2} (\phi_R - \phi_L). \tag{10}$$

Concerning the density, we suggest the three following Rankine-Hugoniot jump relations through the waves of speed λ_L, 0 and λ_R:

$$\rho_L^\star u_L^\star - \rho_L u_L = \lambda_L (\rho_L^\star - \rho_L), \tag{11}$$

$$\rho_R^\star u_R^\star = \rho_L^\star u_L^\star, \tag{12}$$

$$\rho_R u_R - \rho_R^\star u_R^\star = \lambda_R (\rho_R - \rho_R^\star). \tag{13}$$

Let us notice that the consistency relation (5) for the density component is automatically satisfied as soon as the three relations (11), (12) and (13) hold.

To complete the solver, there is one missing equation. We decide to choose a linearization of the equation (2) describing the steady states:

$$p_R^\star - p_L^\star = -\overline{\rho}(\phi_R - \phi_L). \tag{14}$$

The system formed by equations (8), (10), (11), (12), (13) and (14) is easily solved to find

$$\rho_{L,R}^\star = \rho_{L,R} + \frac{1}{\lambda_{L,R}}(q^\star - \rho_{L,R} u_{L,R}), \qquad u_{L,R}^\star = \frac{q^\star}{\rho_{L,R}^\star},$$

$$E_L^\star = \widehat{E} + \frac{\lambda_R}{\lambda_R - \lambda_L}\left(\frac{\rho_L^\star (u_L^\star)^2}{2} - \frac{\rho_R^\star (u_R^\star)^2}{2}\right) + \frac{\overline{\rho}(\phi_R - \phi_L)}{\lambda_R - \lambda_L}\left(\frac{\lambda_R}{\gamma - 1} - \frac{u_L + u_R}{2}\right),$$

$$E_R^\star = \widehat{E} + \frac{\lambda_L}{\lambda_R - \lambda_L}\left(\frac{\rho_L^\star (u_L^\star)^2}{2} - \frac{\rho_R^\star (u_R^\star)^2}{2}\right) + \frac{\overline{\rho}(\phi_R - \phi_L)}{\lambda_R - \lambda_L}\left(\frac{\lambda_L}{\gamma - 1} - \frac{u_L + u_R}{2}\right),$$

where we have set

$$q^\star = \widehat{q} - \frac{1}{\lambda_R - \lambda_L}\overline{\rho}(\phi_R - \phi_L).$$

The characterisation of the approximate Riemann solver will be achieved as soon as the density average $\overline{\rho}$ will be stated. The precise definition of $\overline{\rho}$ will be given in the next section, accordingly to the well-balanced property.

3 The Numerical Scheme

Now, we describe the numerical scheme associated with the approximate Riemann solver $\widetilde{W}$. We consider a mesh of $\mathbb{R}$ made of cells $[x_{i-1/2}, x_{i+1/2})$ for $i \in \mathbb{Z}$, with constant size $\Delta x = x_{i+1/2} - x_{i-1/2}$. We search an update w_i^{n+1} of the solution at time t^{n+1}, knowing an approximation w_i^n at time t^n and on the cell $[x_{i-1/2}, x_{i+1/2})$. We also introduce a discretization of the gravitational potential ϕ as follows:

$$\phi_i = \frac{1}{\Delta x}\int_{x_{i-1/2}}^{x_{i+1/2}} \phi(x)dx.$$

To evolve this approximate solution from t^n to $t^n + \Delta t$, we consider the juxtaposition of Riemann problems located at the interfaces $x_{i+1/2}$. We denote by $\lambda_{i+1/2}^{L,R}$ the left and right speed and by $\overline{\rho}_{i+1/2}^n$ the average value of the density which appear in the approximate Riemann solver $\widetilde{W}\left(\frac{x - x_{i+1/2}}{t - t^n}, w_i^n, w_{i+1}^n\right)$. To ensure that the approximate Riemann solvers do not interact, we enforce the CFL condition

$$\frac{\Delta t}{\Delta x}\max_{i\in\mathbb{Z}}\left|\lambda_{i+1/2}^{L,R}\right| \leq \frac{1}{2}.$$

Next, we follow the classical procedure for Godunov-type schemes to obtain a numerical scheme. It consists of a step of evolution using the approximate Riemann solver, followed by a step of projection on the space of piecewise constant functions. The update approximation at time $t^{n+1} = t^n + \Delta t$ is thus given by

$$w_i^{n+1} = \frac{1}{\Delta x}\int_{-\Delta x/2}^{0} \widetilde{W}\left(\frac{x}{\Delta t}, w_{i-1}^n, w_i^n\right)dx + \frac{1}{\Delta x}\int_{0}^{\Delta x/2} \widetilde{W}\left(\frac{x}{\Delta t}, w_i^n, w_{i+1}^n\right)dx.$$

After straightforward computations, the numerical scheme associated with the approximate Riemann solver developed in the Section 2 can be written as follows:

$$\begin{aligned}
\rho_i^{n+1} =& \rho_i^n - \frac{\Delta t}{\Delta x}\left(F^{\rho}_{i+1/2} - F^{\rho}_{i-1/2}\right),\\
\rho_i^{n+1}u_i^{n+1} =& \rho_i^n u_i^n - \frac{\Delta t}{\Delta x}\left(F^{\rho u}_{i+1/2} - F^{\rho u}_{i-1/2}\right)\\
&- \frac{\Delta t}{2}\left(\overline{\rho}_{i-1/2}\frac{\phi_i - \phi_{i-1}}{\Delta x} + \overline{\rho}_{i+1/2}\frac{\phi_{i+1} - \phi_i}{\Delta x}\right),\\
E_i^{n+1} =& E_i^n - \frac{\Delta t}{\Delta x}\left(F^{E}_{i+1/2} - F^{E}_{i-1/2}\right)\\
&- \frac{\Delta t}{2}\left(\overline{\rho}_{i-1/2}\frac{u_{i-1}^n + u_i^n}{2}\frac{\phi_i - \phi_{i-1}}{\Delta x} + \overline{\rho}_{i+1/2}\frac{u_i^n + u_{i+1}^n}{2}\frac{\phi_{i+1} - \phi_i}{\Delta x}\right),
\end{aligned} \tag{15}$$

where the numerical flux is defined by

$$\left(F^{\rho}_{i+1/2}, F^{\rho u}_{i+1/2}, F^{E}_{i+1/2}\right) = \left(F^{\rho}, F^{\rho u}, F^{E}\right)(w_i^n, w_{i+1}^n), \tag{16}$$

$$F^{\rho}(w_L, w_R) = \frac{\rho_L u_L + \rho_R u_R}{2} + \frac{\lambda_L}{2}(\rho_L^\star - \rho_L) + \frac{\lambda_R}{2}(\rho_R^\star - \rho_R), \tag{17}$$

$$F^{\rho u}(w_L, w_R) = \frac{\rho_L u_L^2 + p_L + \rho_R u_R^2 + p_R}{2} + \frac{\lambda_L}{2}(q^\star - \rho_L u_L) + \frac{\lambda_R}{2}(q^\star - \rho_R u_R), \tag{18}$$

$$F^{E}(w_L, w_R) = \frac{u_L(E_L + p_L) + u_R(E_R + p_R)}{2} + \frac{\lambda_L}{2}(E_L^\star - E_L) + \frac{\lambda_R}{2}(E_R^\star - E_R). \tag{19}$$

Now, we present the properties satisfied by the scheme (3). The first two results deal with the well-balanced properties and are straightforward according to the derivation of the scheme. The last result concerns the robustness of the scheme. The proof is more technical and the reader is referred to [3] for the details.

Theorem 1 *Assume there are positive constants α and β such that the initial data satisfies for all $i \in \mathbb{Z}$:*

$$u_i^0 = 0, \qquad \rho_i^0 = \alpha e^{-\beta\phi_i}, \qquad p_i^0 = \frac{\alpha}{\beta}e^{-\beta\phi_i}.$$

Assume the ρ-average is defined by $\overline{\rho} = \begin{cases} \frac{\rho_R - \rho_L}{\ln(\rho_R) - \ln(\rho_L)} & \text{if } \rho_L \neq \rho_R, \\ \rho_L & \text{if } \rho_L = \rho_R. \end{cases}$

Then the approximation given by (3) stays at rest: $w_i^n = w_i^0$, for all $n \in \mathbb{N}$ and $i \in \mathbb{Z}$.

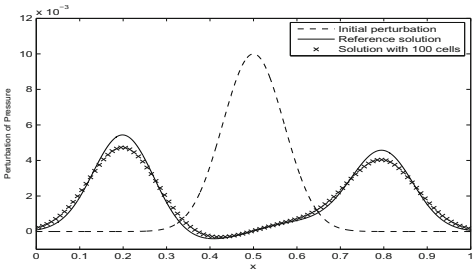


Fig. 2 Pressure perturbation for the hydrostatic atmosphere

Theorem 2 *Assume the initial data satisfies the following approximation of (2) for all $i \in \mathbb{Z}$:*

$$u_i^0 = 0, \qquad \frac{p_{i+1} - p_i}{\Delta x} + \overline{\rho}_{i+1/2} \frac{\phi_{i+1} - \phi_i}{\Delta x} = 0. \tag{20}$$

Then the approximation given by (3) stays at rest: $w_i^n = w_i^0$, for all $n \in \mathbb{N}$ and $i \in \mathbb{Z}$.

We underline that this result holds true independently of the definition of $\overline{\rho}$. In fact, Theorem 2 states a preservation of approximations of the solutions of (2), according to the discretization (20).

Finally, we establish the robustness of the scheme (3).

Theorem 3 *For all $i \in \mathbb{Z}$, assume $|\lambda^L_{i+1/2}|$ and $\lambda^R_{i+1/2}$ are large enough such that*

- $\left|\lambda^R_{i+1/2}/\lambda^L_{i+1/2}\right|$ *is large enough if* $\phi_{i+1} > \phi_i$*;*
- $\left|\lambda^L_{i+1/2}/\lambda^R_{i+1/2}\right|$ *is large enough if* $\phi_{i+1} < \phi_i$*.*

Then the scheme (3) preserves the set Ω: $\forall i \in \mathbb{Z},\ w_i^n \in \Omega \quad \Rightarrow \quad w_i^{n+1} \in \Omega$.

4 Numerical Results

We present now two numerical experiments to underline the relevance of the designed scheme.

The first experiment is taken from [10]. We consider here a constant gravity field given by the potential $\phi(x) = x$. We start with a hydrostatic atmosphere with a perturbation in pressure:

$$\rho_0(x) = e^{-x}, \qquad u_0(x) = 0, \qquad p_0(x) = e^{-x} + 0.01e^{-100(x-0.5)^2}.$$

This initial data is evolved on the computational domain [0, 1] using 100 cells until time $t = 0.25$. The obtained perturbation in pressure is presented in Fig. 2, where it is compared to a reference solution computed using 30.000 cells.

The second test is devoted to illustrate the behaviour of the scheme (3) around a non-hydrostatic steady state. Moreover, this experiment also emphasizes that the

Table 1 L^1 error and convergence rates for the density and the velocity

N	Density		Velocity	
100	2.68E-05	–	2.11E-05	–
200	6.05E-06	2.15	5.40E-06	1.97
400	1.09E-06	2.47	1.36E-06	1.99
800	2.20E-07	2.31	3.39E-07	2.00
1600	4.86E-08	2.18	8.46E-08	2.00
3200	1.14E-08	2.09	2.11E-08	2.00

scheme can deal with more complex gravitational fields than the constant one. Indeed, we consider a gravitational potential given by $\phi(x) = -\sin(2\pi x)$ on the domain $[0, 1]$ with periodic boundary conditions. We can easily check that the solution

$$\rho(x) = 3+2\sin(2\pi x), \quad u(x) = 0, \quad p(x) = 3+3\sin(2\pi x)-0.5\cos(4\pi x) \quad (21)$$

is a non-hydrostatic steady state of (1). We evolve the initial data given by (21) until time $t = 1$ for different values of the number of cells N. The L^1 errors in density and velocity are shown in Table 1 and we observe that although this steady state is not exactly preserved, a second-order convergence is achieved. Let us notice that the scheme (3) is first-order, but this particular steady state is captured up to second-order. This is due to the fact that equation (14) is a second-order approximation of (2).

Acknowledgments This work was partially supported by ANR-12-IS01-0004-01 GEONUM

References

1. Audusse, E., Bouchut, F., Bristeau, M.O., Klein, R., Perthame, B.: A fast and stable well-balanced scheme with hydrostatic reconstruction for shallow water flows. SIAM J. Sci. Comput. **25**(6), 2050–2065 (2004)
2. Chalons, C., Coquel, F., Godlewski, E., Raviart, P.A., Seguin, N.: Godunov-type schemes for hyperbolic systems with parameter-dependent source. the case of euler system with friction. Math. Models Methods Appl. Sci. **20**(11), 2109–2166 (2010)
3. Desveaux, V.: Contribution à l'approximation numérique des systèmes hyperboliques. Ph.D. thesis, Université de Nantes (2013)
4. Gallice, G.: Positive and entropy stable godunov-type schemes for gas dynamics and mhd equations in lagrangian or eulerian coordinates. Numer. Math. **94**(4), 673–713 (2003)
5. Gallouët, T., Hérard, J.M., Seguin, N.: Some approximate godunov schemes to compute shallow-water equations with topography. Comput. & Fluids **32**(4), 479–513 (2003)
6. Gosse, L.: A well-balanced flux-vector splitting scheme designed for hyperbolic systems of conservation laws with source terms. Comput. & Math. Appl. **39**(9), 135–159 (2000)
7. Harten, A., Lax, P., Van Leer, B.: On upstream differencing and godunov-type schemes for hyperbolic conservation laws. SIAM review **25**, 35–61 (1983)
8. Käppeli, R., Mishra, S.: Well-balanced schemes for the Euler equations with gravitation. Tech. rep, Seminar für Angewandte Mathematik Eidgenössische Technische Hochschule (2013)

9. LeVeque, R.: Balancing source terms and flux gradients in high-resolution godunov methods: the quasi-steady wave-propagation algorithm. J. Comput. Phys. **146**, 346–365 (1998)
10. Luo, J., Xu, K., Liu, N.: A well-balanced symplecticity-preserving gas-kinetic scheme for hydrodynamic equations under gravitational field. SIAM J. Sci. Comput. **33**(5), 2356–2381 (2011)

An Explicit Staggered Finite Volume Scheme for the Shallow Water Equations

D. Doyen and P. H. Gunawan

Abstract We propose an explicit finite volume scheme for the shallow water equations. The different unknowns of the system are approximated on staggered meshes. The numerical fluxes are computed with upwind and centered discretizations. We prove a number of properties of the scheme: positivity preserving, well-balanced, consistent with the global entropy inequality. We compare it with collocated schemes, using approximate Riemann solvers, on various problems.

1 Introduction

The shallow water equations are a nonlinear hyperbolic system of conservation laws with a source term due to the topography. In the one-dimensional case, they read

$$\partial_t h + \partial_x (hu) = 0, \tag{1}$$

$$\partial_t (hu) + \partial_x \left(hu^2 + \frac{1}{2} g h^2 \right) + gh \partial_x z = 0, \tag{2}$$

where t denotes the time variable, x denotes the space variable, h is the water height, u is the velocity, g is the gravitational constant, and z is the topography of the bottom.

For such a problem, where shocks can form in the solution, finite volume methods have proved to be very effective. Generally, all the unknowns of the system

D. Doyen
Université Paris-Est and CNRS, LAMA UMR8050, 77454 Marne-la-Vallée, France
e-mail: david.doyen@u-pem.fr

P. H. Gunawan (✉)
Université Paris-Est and Institut Teknologi Bandung, 10, Jalan Ganesha,
Bandung 40132, Indonesia
e-mail: putu-harry.gunawan@univ-paris-est.fr; harry.gunawan.putu@gmail.com

J. Fuhrmann et al. (eds.), *Finite Volumes for Complex Applications VII - Methods and Theoretical Aspects*, Springer Proceedings in Mathematics & Statistics 77,
DOI: 10.1007/978-3-319-05684-5_21,

are approximated on the same mesh and the numerical fluxes are computed with an approximate Riemann solver. We refer to [1, 6] for a thorough description and analysis of this approach. Staggered finite volume discretizations for solving nonlinear hyperbolic system of conservation laws have been investigated more recently (see, e.g., [3–5]). In contrast to the collocated discretization described above, the different unknowns of the system are approximated on staggered meshes. The numerical fluxes can then be computed simply componentwise, using upwind or centered approximations.

In the present paper, we propose an explicit staggered finite volume scheme for the shallow water equations. The scheme is identical to the one in [4] when the topography is flat. We prove a number of properties of the scheme: preservation of the water height positivity, preservation of some particular discrete steady states (well-balanced property), consistency with the entropy inequality. The preservation of the water height positivity is physically relevant and is crucial for the stability, the occurrence of negative quantities leading rapidly to the computation failure. We obtain this positivity for the staggered scheme under a CFL-like condition. The steady states of the shallow water equations are the states (h, u) such that $hu = cst$ and $\frac{1}{2}u^2 + g(h + z) = cst$. In particular, the steady states at rest, that is the steady states such that $u = 0$, satisfy $h + z = cst$. The preservation of these steady states at rest at the discrete level is important since many practical problems are perturbations of such states. With approximate Riemann solvers, this preservation is quite involved [1], whereas it is straightforward in the staggered framework (on uniform grids). The entropy η for the shallow water equations is the sum of the kinetic energy, the potential energy and a term stemming from the topography:

$$\eta(h, hu) := \frac{1}{2}hu^2 + \frac{1}{2}gh^2 + ghz. \tag{3}$$

The entropy inequality reads

$$\partial_t \eta(h, hu) + \partial_x G(h, hu) \leq 0, \tag{4}$$

where the entropy flux G is defined by

$$G(h, hu) := \left(\frac{1}{2}hu^2 + gh^2\right) u + ghzu \tag{5}$$

The staggered scheme does not satisfy a discrete entropy inequality. However, we prove that it is consistent with the global entropy inequalitiy, which guarantees that the discontinuities computed by the scheme are admissible discontinuities.

Finally, we test the staggered scheme on various problems with analytical solutions: a dam break on a wet bed, oscillations in a parabola and a transcritical flow with shock [2]. The results are compared with those of two approximate Riemann solvers (Suliciu/HLLC and kinetic solvers [1]).

The paper is organized as follows. Section 2 describes the staggered scheme and its first properties. Section 3 presents the entropy consistency theorem and Sect. 4 the numerical tests. A brief comparison between the staggered and collocated schemes is made in conclusion.

2 Description of the Scheme

For sake of brevity, we only describe the staggered scheme in the one-dimensional case. The extension to the two-dimensional case is discussed in Remark 2 below. We consider the time interval $(0, T)$ and the space domain $\Omega := (0, L)$ with solid wall boundary conditions (i.e., $u = 0$ at each end of the domain Ω). The time interval is divided into N_t time steps of length Δt and, for all $n \in \{0, \ldots, N_t\}$, $t^n := n\Delta t$. The domain Ω is divided into N_x cells of length Δx. The left end, the center and the right end of the i-th cell are denoted by $x_{i-\frac{1}{2}}$, x_i and $x_{i+\frac{1}{2}}$, respectively. We set $\mathcal{M} := \{1, \ldots, N_x\}$, $\mathcal{E}_{int} := \{1, \ldots, N_x - 1\}$, $\mathcal{E}_b := \{0, N_x\}$, and $\mathcal{E} := \mathcal{E}_{int} \cup \mathcal{E}_b$. The water height h and the topography z are discretized at the center of the cells. The approximation of h at point x_i and at time t^n is denoted by h_i^n. The approximation of z at point x_i is denoted by z_i. The velocity u is discretized at the interfaces between the cells. The approximation of u at point $x_{i+\frac{1}{2}}$ and at time t^n is denoted by $u^n_{i+\frac{1}{2}}$.

The mass conservation equation is discretized with an explicit upwind scheme:

$$h_i^{n+1} - h_i^n + \frac{\Delta t}{\Delta x}\left(q^n_{i+\frac{1}{2}} - q^n_{i-\frac{1}{2}}\right) = 0, \qquad \forall i \in \mathcal{M}, \tag{6}$$

where

$$q^n_{i+\frac{1}{2}} := \hat{h}^n_{i+\frac{1}{2}} u^n_{i+\frac{1}{2}}, \qquad \hat{h}^n_{i+\frac{1}{2}} := \begin{cases} h_i^n & \text{if } u^n_{i+\frac{1}{2}} \geq 0 \\ h^n_{i+1} & \text{if } u^n_{i+\frac{1}{2}} < 0 \end{cases}, \qquad \forall i \in \mathcal{E}.$$

The momentum balance equation is discretized with explicit upwind fluxes for the convection term and implicit centered fluxes for the pressure term and topography term:

$$h^{n+1}_{i+\frac{1}{2}} u^{n+1}_{i+\frac{1}{2}} - h^n_{i+\frac{1}{2}} u^n_{i+\frac{1}{2}} + \frac{\Delta t}{\Delta x}\Big[q^n_{i+1}\hat{u}^n_{i+1} - q^n_i \hat{u}^n_i + \frac{1}{2} g\left[(h^{n+1}_{i+1})^2 - (h^{n+1}_i)^2\right] + g h^{n+1}_{i+\frac{1}{2}}(z_{i+1} - z_i)\Big] = 0, \qquad \forall i \in \mathcal{E}_{int}, \tag{7}$$

where

$$h^n_{i+\frac{1}{2}} := \frac{1}{2}\left(h^n_i + h^n_{i+1}\right), \qquad \forall i \in \mathcal{E}_{int},$$

$$q_i^n := \frac{1}{2}\left(q_{i-\frac{1}{2}}^n + q_{i+\frac{1}{2}}^n\right), \qquad \hat{u}_i^n := \begin{cases} u_{i-\frac{1}{2}}^n & \text{if } q_i^n \geq 0 \\ u_{i+\frac{1}{2}}^n & \text{if } q_i^n < 0 \end{cases}, \qquad \forall i \in \mathscr{M}.$$

The discrete boundary conditions are

$$u_{i+\frac{1}{2}}^{n+1} = 0 \quad \forall i \in \mathscr{E}_b. \tag{8}$$

The computation of the discrete unknowns at each time step is completely explicit. First the discrete water heights $\{h_i^{n+1}\}$ are computed with (6), then the discrete velocities $\{u_{i+\frac{1}{2}}^{n+1}\}$ are computed with (7) (if $h_{i+\frac{1}{2}}^{n+1} = 0$, by convention, $u_{i+\frac{1}{2}}^{n+1}$ is set to zero). We see immediately that this scheme conserves the mass and, for a flat topography, the total momentum. It is easy to verify that the water height remains nonnegative at time t^{n+1} under the CFL-like condition

$$\Delta t \leq \frac{\Delta x}{(-u_{i-\frac{1}{2}}^n)^+ + (u_{i+\frac{1}{2}}^n)^+}, \quad \forall i \in \mathscr{M}, \tag{9}$$

where, for any $a \in \mathbb{R}$, $(a)^+ := \max(a, 0)$ (similarly, the notation $(a)^- := \min(a, 0)$ will be used in the next section). The steady states at rest are also preserved. Indeed, if $u_{i+\frac{1}{2}}^n = 0$ and $h_i^n + z_i = cst$ for all $i \in \{0, \ldots, N\}$, then $u_{i+\frac{1}{2}}^{n+1} = 0$ and $h_i^{n+1} + z_i = cst$. At each time step, the Courant number is defined by

$$\nu := \frac{\Delta t}{\Delta x} \max_{i \in \mathscr{M}} \left(\frac{\left| q_{i+\frac{1}{2}}^n + q_{i-\frac{1}{2}}^n \right|}{2h_i^n} + \sqrt{g h_i^n} \right). \tag{10}$$

The numerical simulations show that the staggered scheme is stable under the CFL condition $\nu < 1$.

Remark 1 It is essential to make an implicit discretization of the pressure term in (7). With an explicit discretization ($\left[(h_{i+1}^n)^2 - (h_i^n)^2\right]$ instead of $\left[(h_{i+1}^{n+1})^2 - (h_i^{n+1})^2\right]$), the scheme would not be consistent with the entropy inequality. As a consequence, non-entropic shocks might occur in the numerical simulations (see Fig. 2).

Remark 2 The extension of the scheme (6)–(8) to the two-dimensional case is direct with a Cartesian grid. The water height h and the topography z are discretized at the center of the cells. The velocity in the x-direction is discretized at the center of the edges normal to the x-direction, while the velocity in the y-direction is discretized at the center of the edges normal to the y-direction.

Remark 3 The upwind fluxes and the Euler-like time-integration limit the scheme (6)–(8) to first-order accuracy in space and time. Second-order accuracy can be achieved with usual techniques: MUSCL or ENO flux reconstruction, second-order time-integration scheme (for instance, a scheme mixing Heun and Crank-Nicolson discretizations).

Remark 4 Source terms are often added to the momentum equation to model friction phenomena. For instance, the Manning friction term is $Cu|u|/h^{1/3}$, where C is a given coefficient. In the staggered framework, such a term can be approximated with $Cu^{n+1}_{i+\frac{1}{2}}|u^n_{i+\frac{1}{2}}|/(h^{n+1}_{i+\frac{1}{2}})^{1/3}$.

3 Consistency with the Entropy Inequality

We first establish a balance on the kinetic energy and a balance on the potential and topography energy.

Proposition 1 *The discrete solution of the scheme* (6)–(7) *satisfies, for all* $i \in \mathscr{E}_{int}$ *and all* $n \in \{0, \ldots, N_t\}$*, the balance*

$$\frac{1}{2}\frac{\Delta x}{\Delta t}\left(h^{n+1}_{i+\frac{1}{2}}(u^{n+1}_{i+\frac{1}{2}})^2 - h^n_{i+\frac{1}{2}}(u^n_{i+\frac{1}{2}})^2\right) + \frac{1}{2}\left(q^n_{i+1}(\hat{u}^n_{i+1})^2 - q^n_i(\hat{u}^n_i)^2\right) + \frac{1}{2}g\left((h^{n+1}_{i+1})^2 - (h^{n+1}_i)^2\right)u^{n+1}_{i+\frac{1}{2}} + gh^{n+1}_{i+\frac{1}{2}}u^{n+1}_{i+\frac{1}{2}}(z_{i+1} - z_i) = -R^{n+1}_{i+\frac{1}{2}} \tag{11}$$

with

$$\begin{aligned} R^{n+1}_{i+\frac{1}{2}} := &\frac{1}{2}\frac{\Delta x}{\Delta t}h^{n+1}_{i+\frac{1}{2}}\left(u^{n+1}_{i+\frac{1}{2}} - u^n_{i+\frac{1}{2}}\right)^2 + \frac{1}{2}\left((q^n_{i+1})^-\left(u^n_{i+\frac{3}{2}} - u^n_{i+\frac{1}{2}}\right)^2\right. \\ &\left. - (q^n_i)^+\left(u^n_{i-\frac{1}{2}} - u^n_{i+\frac{1}{2}}\right)^2\right) - \left[(q^n_{i+1})^-\left(u^n_{i+\frac{3}{2}} - u^n_{i+\frac{1}{2}}\right)\right. \\ &\left. - (q^n_i)^+\left(u^n_{i-\frac{1}{2}} - u^n_{i+\frac{1}{2}}\right)\right]\left(u^{n+1}_{i+\frac{1}{2}} - u^n_{i+\frac{1}{2}}\right), \end{aligned} \tag{12}$$

and, for all $i \in \mathscr{M}$ *and all* $n \in \{0, \ldots, N_t\}$*, the balance*

$$\frac{\Delta x}{\Delta t}\left(\frac{1}{2}g(h^{n+1}_i)^2 - \frac{1}{2}g(h^n_i)^2\right) + \left(\frac{1}{2}g(\hat{h}^n_{i+\frac{1}{2}})^2u^n_{i+\frac{1}{2}} - \frac{1}{2}g(\hat{h}^n_{i-\frac{1}{2}})^2u^n_{i-\frac{1}{2}}\right) + \frac{1}{2}g(h^n_i)^2\left(u^n_{i+\frac{1}{2}} - u^n_{i-\frac{1}{2}}\right) = -R^{n+1}_i \tag{13}$$

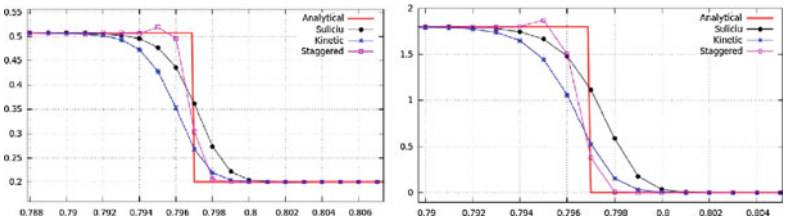


Fig. 1 Dam break on a wet bed. Water height (*left*) and velocity (*right*) around the shock at time $t = 0.1$. The number of cells is $Nx = 1000$

with

$$R_i^{n+1} := \frac{1}{2} g \frac{\Delta x}{\Delta t} \left(h_i^{n+1} - h_i^n\right)^2 + \frac{1}{2} \left(h_{i+1}^n - h_i^n\right)^2 (u_{i+\frac{1}{2}}^n)^- - \frac{1}{2} \left(h_{i-1}^n - h_i^n\right)^2 (u_{i-\frac{1}{2}}^n)^+ + g \left(h_i^{n+1} - h_i^n\right) \left(q_{i+\frac{1}{2}}^n - q_{i-\frac{1}{2}}^n\right). \quad (14)$$

Using the above balances we can deduce the following global consistency result.

Theorem 1 *Let $Q_T := (0, T) \times \Omega$. Let $(h^{(k)}, u^{(k)})_{k \in \mathbb{N}}$ be a sequence of discrete solutions. If the sequence satisfies some estimates (that we do not specify here) and converges in $L^p(Q_T) \times L^p(Q_T)$, with $p \in [1, +\infty)$, then its limit (h, u) satisfy*

$$\int_{Q_T} \left(\eta(h, hu)\partial_t \varphi + G(h, hu)\partial_x \varphi\right) \mathrm{d}x\mathrm{d}t \geq 0, \quad (15)$$

for all nonnegative test functions $\varphi \in C_c^\infty(Q_T)$.

Remark 5 A similar consistency study in the case of a flat topography can be found in [4].

4 Numerical Tests

4.1 Dam Break on a Wet Bed

The domain is $\Omega = [0, 1]$ and the topography is flat. The initial velocity is zero and the initial water height is $h_{\text{ini}}(x) = \mathbb{1}_{x<0.5} + 0.2\mathbb{1}_{x \geq 0.5}$. The solution of this problems consists of a shock and a rarefaction wave. The numerical solution obtained with the staggered scheme is in very good agreement with the exact solution (see Fig. 1). The staggered scheme yields a sharper shock than the kinetic and Suliciu schemes. However, a small overshoot in the water height and the velocity can be observed upstream from the shock for the staggered scheme. This overshoot does not occur in the kinetic and Suliciu schemes. The numerical convergence rate of

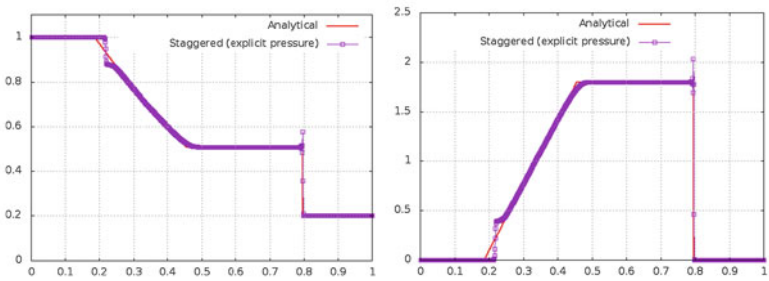


Fig. 2 Dam break on a wet bed. Water height (*left*) and velocity (*right*) at time $t = 0.1$. The number of cells is $Nx = 1000$

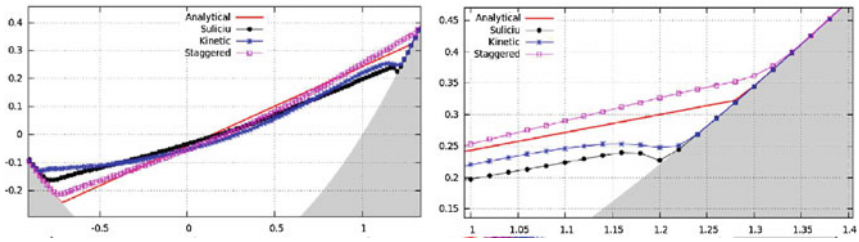


Fig. 3 Oscillations in a parabola. Water level $h + z$ at time $t = 2.7$. The *right* figure is a zoom of the *left* figure. The number of cells is $Nx = 200$

the staggered scheme in L^1-norm is slightly less than 0.8 for h and about 0.8 for u. These convergence rates are comparable to those obtained with the kinetic and Suliciu schemes.

The staggered scheme with an explicit pressure, discussed in Remark 1, has also been tested on the dam break problem. The non-entropic character of this variant is confirmed by the simulations; see the occurrence of a non-entropic shock at the left end of the rarefaction wave in Fig. 2.

4.2 Oscillations in a Parabola

The domain is $\Omega = [-2, 2]$. The topography is a parabola $z(x) = 0.5\left(x^2 - 1\right)$. The initial velocity is zero and the initial water height is $h_{\text{ini}}(x) = -z(x + 0.5)\mathbb{1}_{-1.5<x<0.5}$. The solution is periodic in time: the water oscillates in the parabola, the surface remaining planar. The numerical simulations show that the staggered scheme is able to treat accurately dry-wet transitions (see Fig. 3). For this problem, the staggered scheme is even more accurate than the collocated schemes. The numerical convergence rate of the staggered scheme in L^1-norm is very close to 1 for both h and hu and seems slightly better than the kinetic and Suliciu schemes.

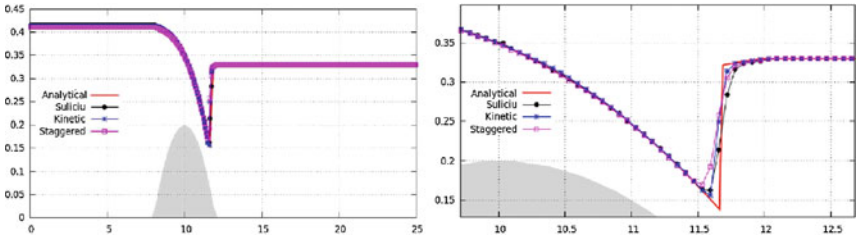


Fig. 4 Transcritical flow with shock. Water level $h + z$ at time $t = 200$. The *right* figure is a zoom of the *left* figure. The number of cells is $Nx = 200$

4.3 Transcritical Flow with Shock

This problem simulates a flow over a bump. The domain is $\Omega = [0, 25]$ and the topography is $z(x) = 0.2 - 0.05(x - 10)^2 \mathbb{1}_{8<x<12}$. The initial conditions are $h_{\text{ini}} = 0.33 - z$ and $u_{\text{ini}} = 0.18/0.33$. At the upstream end of the domain, the discharge is enforced ($hu(x = 0) = 0.18$), while, at the downstream end, the water height is prescribed ($u(x = 25) = 0.33$). After a time long enough ($t \geq 100$), the flow reaches a steady state. Upstream from the top of the bump, the flow is subcritical; downstream, it becomes supercritical, then, after a hydraulic jump, it is subcritical again. The staggered scheme reproduces quite accurately the final steady state and captures well the location of the shock. (see Fig. 4). The kinetic and Suliciu schemes seem to yield a slightly better shock profile than the staggered scheme. The numerical convergence rate of the staggered scheme in L^1-norm is close to 1 for both h and u, which is comparable to the kinetic and Suliciu schemes.

5 Conclusion

The staggered scheme is an alternative to the collocated schemes for solving the shallow water equations. The numerical tests demonstrate that the staggered scheme is as accurate and robust as the collocated schemes. The staggered scheme presents the same mathematical guarantees as the collocated schemes. In particular, the staggered scheme is consistent with the global entropy inequality. The componentwise computation of numerical fluxes is simpler than the approximate Riemann solvers and the well-balanced discretization of the topography source term is straightforward on uniform grids.

References

1. Bouchut, F.: Nonlinear Stability of Finite Volume Methods for Hyperbolic Conservation Laws and Well-Balanced Schemes for Sources. Birkhäuser, Basel (2004)
2. Delestre, O., Lucas, C., Ksinant, P.A., Darboux, F., Laguerre, C., Vo, T.N.T., James, F., Cordier, S.: Swashes: a compilation of shallow water analytic solutions for hydraulic and environmental studies. Int. J. Numer. Meth. Fluids **72**(3), 269–300 (2013)
3. Herbin, R., Kheriji, W., Latché, J.C., et al.: Consistent Semi-Implicit Staggered Schemes for Compressible Flows. Part I: The Barotropic Euler Equations. submitted (2013)
4. Herbin, R., Latché, J.C., Nguyen, T.T., et al.: Consistent Explicit Staggered Schemes for Compressible Flows Part I: The Barotropic Euler Equations. submitted (2013)
5. Stelling, G.S., Duinmeijer, S.P.A.: A staggered conservative scheme for every froude number in rapidly varied shallow water flows. Int. J. Numer. Meth. Fluids **43**(12), 1329–1354 (2003)
6. Toro, E.F.: Riemann Solvers and Numerical Methods for Fluid Dynamics. Springer, Berlin (2009)

A Uniformly Converging Scheme for Fractal Conservation Laws

Jérôme Droniou and Espen R. Jakobsen

Abstract The fractal conservation law $\partial_t u + \partial_x(f(u)) + (-\Delta)^{\alpha/2} u = 0$ changes characteristics as $\alpha \to 2$ from non-local and weakly diffusive to local and strongly diffusive. In this paper we present a corrected finite difference quadrature method for $(-\Delta)^{\alpha/2}$ with $\alpha \in [0, 2]$, combined with usual finite volume methods for the hyperbolic term, that automatically adjusts to this change and is uniformly convergent with respect to $\alpha \in [\eta, 2]$ for any $\eta > 0$. We provide numerical results which illustrate this asymptotic-preserving property as well as the non-uniformity of previous finite difference or finite volume type of methods.

1 Introduction

We consider the following fractional conservation law

$$\begin{aligned} &\partial_t u_\alpha + \partial_x(f(u_\alpha)) + \mathscr{L}_\alpha[u_\alpha] = 0\,,\ t > 0\,,\ x \in \mathbb{R}\,, \\ &u_\alpha(0, x) = u_{\text{ini}}(x)\,, \qquad\qquad x \in \mathbb{R}\,, \end{aligned} \tag{1}$$

where $\alpha \in [0, 2]$, $\mathscr{L}_\alpha = (-\Delta)^{\alpha/2}$,

$$u_{\text{ini}} \in L^\infty(\mathbb{R}) \cap BV(\mathbb{R}) \quad \text{and} \quad f : \mathbb{R} \to \mathbb{R} \text{ is locally Lipschitz-continuous.} \tag{2}$$

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

E. R. Jakobsen
Department of Mathematical Sciences, Norwegian University of Science and Technology, 7491 Trondheim, Norway
e-mail: erj@math.ntnu.no

J. Fuhrmann et al. (eds.), *Finite Volumes for Complex Applications VII - Methods and Theoretical Aspects*, Springer Proceedings in Mathematics & Statistics 77,
DOI: 10.1007/978-3-319-05684-5_22,

Such models appear for example in mathematical finance, gas detonation or semiconductor growth [1, 11, 23, 26]. The fractional Laplacian $\mathscr{L}_\alpha = (-\Delta)^{\alpha/2}$ can be defined e.g. as a Fourier multiplier, but for our purpose the following equivalent definition, valid for any $\varphi \in C_c^\infty(\mathbb{R})$ (set of smooth compactly supported functions), is more useful:

$$\begin{cases} \mathscr{L}_0[\varphi](x) = \varphi(x), & \alpha = 0, \\ \mathscr{L}_\alpha[\varphi](x) = -c_\alpha \displaystyle\int_{\mathbb{R}} \frac{\varphi(x+z)-\varphi(x)-\varphi'(x)z\mathbf{1}_{[-1,1]}(z)}{|z|^{1+\alpha}}\, dz, & \alpha \in (0,2), \\ \mathscr{L}_2[\varphi](x) = -\Delta\varphi(x), & \alpha = 2, \end{cases} \tag{3}$$

where $\mathbf{1}_{[-1,1]}$ is the characteristic function of $[-1, 1]$, $c_\alpha = (2\pi)^\alpha \frac{\alpha\Gamma(\frac{1+\alpha}{2})}{2\pi^{\frac{1}{2}+\alpha}\Gamma(1-\frac{\alpha}{2})}$ and Γ is the Euler function [15].

As $\alpha \to 2$, the operator $\mathscr{L}_\alpha$ changes nature and properties. For $\alpha \in (0, 2)$, $\mathscr{L}_\alpha$ is a *non-local* pseudo-differential operator of order < 2, and it has relatively weak diffusive properties since the decay at infinity of the fundamental solution of $\partial_t u + \mathscr{L}_\alpha[u] = 0$ is polynomial. At $\alpha = 2$, $\mathscr{L}_\alpha = -\Delta$ is a *local* operator with strong diffusive properties and a fundamental solution with super-exponential decay. When α vary over $[0, 2]$, the qualitative behaviour of the solution u_α of (1) also changes. In the case that $\alpha = 2$, it is well-known that u_α becomes instantly smooth for $t > 0$ even when the initial data is discontinuous. On the contrary, for $\alpha = 0$, the solution may develop shocks and uniqueness of the solution requires additional entropy conditions and the corresponding notion of entropy solution [22]. The study of the fractional case $\alpha \in (0, 2)$ dates back to [6], with some restrictions on α and f. The first complete study in the case $\alpha > 1$ for any locally Lipschitz f and bounded initial data u_{ini} can be found in [14]. Here it is proved that the solution becomes instantly smooth even if u_{ini} is only bounded (see also [15]). If $\alpha < 1$, then the solution can develop shocks [4] and the weak solution need not be unique [3]. The notion of entropy solution of [2] is therefore required to obtain a well-posed formulation.

There exists a vast literature on the numerical approximation of scalar conservation laws (i.e. (1) without $\mathscr{L}_\alpha$), see e.g. [17–19] and references therein. The study of numerical methods for fractal conservation laws is much more recent with a corresponding less extensive literature. Probabilistic methods have been studied in [21, 24], but must be applied to the equation satisfied by $\partial_x u_\alpha$ in order to avoid noisy results, and recovering from this a numerical approximation of u_α may be challenging in dimension greater than 1. Deterministic methods for (1) like finite difference, volume, and element methods (discontinuous Galerkin) are given in [8, 10, 13], while a high order spectral vanishing viscosity method is introduced in [9]. The latter method and its analysis is very different from the former three methods, with convergence and (non-optimal) error estimates that are independent of $\alpha \in (0, 2)$. As opposed to the spectral method, the other methods are monotone or have low order monotone variants.

Surprisingly, *for all the non-spectral monotone methods the convergence deteriorates as* $\alpha \to 2$, and the schemes themselves are not even defined in the limit $\alpha = 2$. The purpose of this paper is to present an *asymptotic-preserving monotone* scheme for (1) defined for any $\alpha \in [0, 2]$, i.e. a scheme that provides a monotone approximation of u_α which is uniform with respect to $\alpha \in [0, 2]$. In particular, our scheme naturally adapts to the change of behaviour of $\mathscr{L}_\alpha$ as $\alpha \to 2$ and $\alpha \to 0$ and its convergence properties do not deteriorate in these extreme cases. The idea behind our scheme is to add a correction term in the form of a suitably chosen vanishing local viscosity term. Similar ideas have been used for other equations before, see e.g. [12] for linear equations and [20] for fully nonlinear equations. A stochastic interpretation can be found in [5].

This paper is organised as follows. The numerical method is presented in Sect. 2, and its asymptotic-preserving characteristics are discussed. Due to lack of space and the technical nature of the proofs, we skip them and refer instead to [16]. In Sects. 3 and 4, we define precisely what asymptotic preserving means and the we give a couple numerical simulations to illustrate this property of the method.

2 The Scheme

The new scheme is based on monotone conservative finite difference approximations of the local terms combined with quadrature, truncation of $\frac{1}{|z|^{1+\alpha}}$ near the singularity, and a second order correction term (vanishing viscosity) for the non-local term. Except for the correction term, the scheme is similar to the schemes of [8, 13] and of [10] with P_0-elements. It is monotone, conservative, and converges in L^1_{loc} uniformly in $\alpha \in [\eta, 2]$ for all $\eta > 0$.

For given space and time steps $\delta x, \delta t > 0$, we introduce the grid $t_n := n\delta t$ and $x_i := i\delta x + \frac{\delta x}{2}$ for $n \in \mathbb{N}_0$ and $i \in \mathbb{Z}$. We identify sequences $(\varphi_i)_{i\in\mathbb{Z}}$ of numbers with piecewise constant functions $\varphi_{\delta x} : \mathbb{R} \to \mathbb{R}$ equal to φ_i on $[i\delta x, (i+1)\delta x)$ for all $i \in \mathbb{Z}$. Similarly, $(\varphi_i^n)_{n\geq 0,\, i\in\mathbb{Z}}$ is identified with $\varphi_{\delta x,\delta t} : [0, \infty) \times \mathbb{R} \to \mathbb{R}$ equal to φ_i^n on $[n\delta t, (n+1)\delta t) \times [i\delta x, (i+1)\delta x)$ for all $n \geq 0$ and $i \in \mathbb{Z}$. The discretisation of (1) can then we written as: find $u_{\alpha,\delta x,\delta t} = (u_i^n)_{n\geq 0,\, i\in\mathbb{Z}}$ such that

$$u_i^0 = \frac{1}{\delta x} \int_{[i\delta x,(i+1)\delta x)} u_0(x)\, dx \quad \text{for all } i \in \mathbb{Z}, \tag{4}$$

$$\frac{u_i^{n+1} - u_i^n}{\delta t} + \mathscr{F}_{\delta x}(u^n)_i + \mathscr{L}_{\alpha,\delta x}[u^{n+1}]_i = 0 \quad \text{for all } n \geq 0 \text{ and all } i \in \mathbb{Z}, \tag{5}$$

where $\mathscr{F}_{\delta x}$ is any monotone consistent and conservative discretization of $\partial_x(f(u))$ (see e.g. [17–19]), and $\mathscr{L}_{\alpha,\delta x}$ is a monotone discretisation of $\mathscr{L}_\alpha$ to be defined. Note that the scheme has explicit convection and implicit diffusion terms.

The first and simplest idea to obtain a monotone discretization of $\mathscr{L}_\alpha$ for $\alpha \in (0,2)$ is to discretize the integral in (3) using a simple (weighted) midpoint type quadrature rule, see e.g. [8, 10, 13]. For $\varphi \in C_c^\infty(\mathbb{R})$ and letting $\varphi_l = \varphi(x_l)$ if $l \in \mathbb{Z}$, this leads to

$$\mathscr{L}_\alpha[\varphi](x_i) \approx \tilde{\mathscr{L}}_{\alpha,\delta x}[\varphi]_i := -\sum_{j\in\mathbb{Z}\setminus\{0\}} \left(\varphi_{i+j} - \varphi_i\right) \int_{(j\delta x - \frac{\delta x}{2}, j\delta x + \frac{\delta x}{2})} \frac{c_\alpha}{|z|^{1+\alpha}}\, dz. \tag{6}$$

However, as $\alpha \to 2$ we have $c_\alpha \to 0$ and therefore $\tilde{\mathscr{L}}_{\alpha,\delta x} \to 0$ for fixed δx. In the limit $\alpha \to 2$ the scheme then converges to

$$\frac{u_i^{n+1} - u_i^n}{\delta t} + \mathscr{F}_{\delta x}(u^n)_i = 0 \quad \text{for all } n \geq 0 \text{ and all } i \in \mathbb{Z},$$

which is a discretisation of $\partial_t u + \partial_x(f(u)) = 0$ and not $\partial_t u + \partial_x(f(u)) - \Delta u = 0$. Hence the limits $\alpha \to 2$ and $\delta x \to 0$ do not commute and the scheme is not asymptotic-preserving.

Note that $\tilde{\mathscr{L}}_{\alpha,\delta x}$ vanishes in the limit because the measure $\frac{c_\alpha\, dz}{|z|^{1+\alpha}}$ concentrates around 0 as $\alpha \to 2$, while in the above midpoint rule the integral in (3) over $(-\frac{\delta x}{2}, \frac{\delta x}{2})$ will always be zero by symmetry. We therefore need to replace the midpoint rule on this interval by a more accurate rule based on the second order interpolation polynomial P_i of φ around the node x_i. We find that this polynomial satisfies $P_i(x_i+z) - P_i(x_i) - P_i'(x_i)z = \frac{1}{2\delta x^2}\left(z^2\varphi_{i-1} - 2z^2\varphi_i + z^2\varphi_{i+1}\right)$ and the new discretization therefore becomes

$$\begin{aligned}\hat{\mathscr{L}}_{\alpha,\delta x}[\varphi]_i &:= -c_\alpha \int_{-\frac{\delta x}{2}}^{\frac{\delta x}{2}} \frac{P(x_i+z) - P(x_i) - P'(x_i)z}{|z|^{1+\alpha}}\, dz + \tilde{\mathscr{L}}_{\alpha,\delta x}[\varphi]_i \\ &= \frac{\varphi_{i+1} - 2\varphi_i + \varphi_{i-1}}{\delta x^2} \int_{(-\frac{\delta x}{2}, \frac{\delta x}{2})} \frac{c_\alpha |z|^{1-\alpha}}{2}\, dz + \tilde{\mathscr{L}}_{\alpha,\delta x}[\varphi]_i.\end{aligned}$$

We can check that the new approximation has the following truncation error [16]:

$$\begin{aligned}&|\mathscr{L}_\alpha[\varphi](x_i) - \hat{\mathscr{L}}_{\alpha,\delta x}[\varphi]_i| \\ &\quad \leq C\left(\|\varphi^{(4)}\|_{L^\infty}\delta x^{4-\alpha} + \|\varphi''\|_{L^\infty} c_\alpha \left(\tfrac{1}{\alpha} + \tfrac{1}{|1-\alpha|}\right)\delta x^{\min(1,2-\alpha)} + \|\varphi'\|_{L^\infty}\delta x\right),\end{aligned}$$

which is $O(\delta x) + o_\alpha(1)$ as $\alpha \to 2$ and therefore does not deteriorate in this limit. Note that if $\alpha = 1$, then $\frac{1}{|1-\alpha|}\delta x^{\min(1,2-\alpha)}$ must be replaced with $\delta x|\ln(\delta x)|$.

In order to obtain an approximation which uses only a finite number of discrete values, we also truncate the sum in (6) as in [13] at some index $J_{\delta x} > 0$ (which may depend upon α) where $J_{\delta x}\delta x \to \infty$ as $\delta x \to 0$. The final approximate operator $\mathscr{L}_{\alpha,\delta x}$ is therefore

$$\mathscr{L}_{\alpha,\delta x}[\varphi]_i = -\sum_{0<|j|\le J_{\delta x}} W^j_{\alpha,\delta x}(\varphi_{i+j}-\varphi_i) - W^{J_{\delta x}+1}_{\alpha,\delta x}\left(\varphi_{i-J_{\delta x}-1}-\varphi_i\right) - W^{J_{\delta x}+1}_{\alpha,\delta x}\left(\varphi_{i+J_{\delta x}+1}-\varphi_i\right) - W^0_{\alpha,\delta x}\frac{\varphi_{i+1}-2\varphi_i+\varphi_{i-1}}{\delta x^2}, \tag{7}$$

with weights

$$\begin{aligned} W^0_{\alpha,\delta x} &= \int_{(-\frac{\delta x}{2},\frac{\delta x}{2})} \frac{c_\alpha |z|^{1-\alpha}}{2}\,dz\,, \\ W^j_{\alpha,\delta x} &= \int_{(j\delta x-\frac{\delta x}{2}, j\delta x+\frac{\delta x}{2})} \frac{c_\alpha}{|z|^{1+\alpha}}\,dz \qquad \text{for} \qquad 0<|j|\le J_{\delta x}, \\ W^{J_{\delta x}+1}_{\alpha,\delta x} &= \int_{z>J_{\delta x}\delta x+\frac{\delta x}{2}} \frac{c_\alpha}{|z|^{1+\alpha}}\,dz = \int_{z<-J_{\delta x}\delta x-\frac{\delta x}{2}} \frac{c_\alpha}{|z|^{1+\alpha}}\,dz. \end{aligned} \tag{8}$$

The last term in (7) contains the classical discretization of $\varphi''(x_i)$ and is the new correction term compared with the discretisations of [8, 10, 13]. Discretisation (7), (8) fits in the generic framework of [13] from which we can conclude:

Theorem 1 ([16]) *Under a standard CFL condition for the convection term,*

1. *There is a unique solution $u_{\alpha,\delta x,\delta t}$ of the scheme defined by (4), (5), (7) and (8), satisfying $\|u_{\alpha,\delta x,\delta t}\|_{L^\infty} \le \|u_{ini}\|_{L^\infty}$ and $|u_{\alpha,\delta x,\delta t}(t,\cdot)|_{BV} \le |u_{ini}|_{BV}$ for all $t>0$.*
2. *For fixed α, $u_{\alpha,\delta x,\delta t}$ converges in $L^1_{\mathrm{loc}}([0,\infty)\times\mathbb{R})$ as $(\delta x,\delta t)\to 0$ to the unique entropy solution u_α of (1).*

Remark 1 We set $\mathscr{L}_{2,\delta x}[\varphi]_i = -(\varphi_{i+1}-2\varphi_i+\varphi_{i-1})/\delta x^2$ and $\mathscr{L}_{0,\delta x}[\varphi]_i = \varphi_i$. This consists in fixing δx and sending $\alpha\to 2$ or $\alpha\to 0$ in (7). Taking the limits in the scheme (5), we obtain the classical implicit scheme for the (1) with $\alpha=2$ or $\alpha=0$.

3 The Asymptotic-Preserving Property

The scheme is asymptotic-preserving if its solution $u_{\alpha,\delta x,\delta t}$ satisfies the following uniform approximation result away from $\alpha=0$ (see [16] for the case $\alpha=0$):

$$\forall \eta>0\,,\quad \sup_{\alpha\in[\eta,2]} d_{L^1_{\mathrm{loc}}([0,\infty)\times\mathbb{R})}(u_{\alpha,\delta x,\delta t},u_\alpha)\to 0 \text{ as } (\delta x,\delta t)\to 0 \tag{9}$$

where $d_{L^1_{\mathrm{loc}}([0,\infty)\times\mathbb{R})}(u,v)=\sum_{n=1}^\infty 2^{-n}\min(1,||u-v||_{L^1([0,n)\times(-n,n))})$ is the usual distance defining the topology of $L^1_{\mathrm{loc}}([0,\infty)\times\mathbb{R})$. Here and elsewhere, the convergence $(\delta x,\delta t)\to 0$ is always taken under a standard CFL condition depending on the definition of the convective flux $\mathscr{F}$ in (5) (see e.g. [8, 10, 13]). This formulation

of the asymptotic-preserving property is very general and does not require an explicit error estimate independent on α. Such an estimate seems particularly challenging to obtain in the absence of regularity of the solution as $t \to 0$.

Theorem 2 ([16]) *Under a standard CFL for the convection part, the numerical scheme defined by* (4) (5)*,* (7) *and* (8) *is asymptotic-preserving.*

Next we want to illustrate this property numerically. As it is formulated now, this would require us to have access to the exact solution u_α, which is not the case. We overcome this difficulty by using instead the following equivalent reformulation of (9) (see [16]), which can be checked by computing approximate solutions only:

$$\forall \alpha_0 \in (0,2]\,, \quad \text{for any sequence } (\delta x_k, \delta t_k)_{k\in\mathbb{N}} \text{ converging to } 0\text{:}$$
$$\sup_{k\geq 1} d_{L^1_{\mathrm{loc}}([0,\infty)\times\mathbb{R})}(u_{\alpha,\delta x_k,\delta t_k}, u_{\alpha_0,\delta x_k,\delta t_k}) \to 0 \text{ as } \alpha \to \alpha_0. \tag{10}$$

Remark 2 The matrix of $\mathscr{L}_{\alpha,\delta x}$ defined by (7) is a semi-definite Toepliz matrix as in [8, 10, 13]. Implementation of the scheme thus takes advantage of super-fast multiplication and inversion algorithms for these matrices [7, 25]. Computing several approximate solutions, as required in (10), is therefore not very expensive.

4 Numerical Results

In all these tests, we take the Burgers flux $f(u) = \frac{u^2}{2}$ and $\mathscr{F}_{\delta x}$ given by a MUSCL method. The final time is $T = 1$ and the spatial computational domain is $[-1, 1]$. We use the same truncation parameters (in particular $J_{\delta x}$) as in [13, Sect. 4.1.2].

For each test, we choose the discretisation steps $(\delta x_k, \delta t_k) = (\frac{1}{2^k\times 50}, \frac{1}{2^k\times 100})$ for $k = 1, \ldots, 4$, which all satisfy the CFL for (5). We also select four values $(\alpha_m)_{m=1,\ldots,4} = (1.8, 1.9, 1.99, 1.999)$ which are near $\alpha_0 = 2$, the difficult case in assessing the uniformity of the convergence in (10) and the reason why we introduced the correction term in (7). We then indicate, for $m = 1, \ldots, 4$, the value of

$$E_m = \max_{k=1,\ldots,4} \sup_{t\in[0,1]} \|u_{\alpha_m,\delta x_k,\delta t_k}(\cdot,t) - u_{\alpha_0,\delta x_k,\delta t_k}(\cdot,t)\|_{L^1([-1,1])},$$

that is $\max_{k=1,\ldots,4}$ the $L^\infty(L^1)$ norm of $u_{\alpha_m,\delta x_k,\delta t_k} - u_{\alpha_0,\delta x_k,\delta t_k}$ on the computational domain. This is a stronger norm that the $L^1(L^1)$ norm used in (10). Hence, E_m approaching 0 as m increases is an even better indication that the scheme is asymptotic-preserving.

Test 1 (rarefaction): we select a Riemann initial condition, $u_{\mathrm{ini}} = -1$ if $x < 0$ and $u_{\mathrm{ini}} = 1$ if $x > 0$. In this case both convection and diffusion work to smooth out

Table 1 Comparison between the uncorrected scheme of [13] and our corrected scheme, $u_{\text{ini}} = -1$ on $(-\infty, 0)$, $u_{\text{ini}} = 1$ on $(0, \infty)$

	E_1	E_2	E_3	E_4
Uncorrected scheme	1.8E-1	3E-1	8.8E-1	9.1E-1
Corrected scheme	5.1E-2	2.2E-2	1.7E-4	1.7E-5

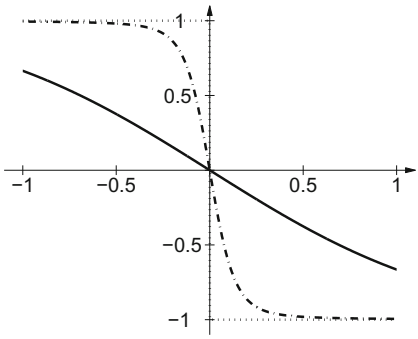


Fig. 1 Approximate solutions provided at $T = 1$ by the corrected (*continuous*) and uncorrected (*dashed*) schemes for (1) with $\alpha = 1.99$. The *dotted line* is both the initial condition and the solution to $\partial_t u + \partial_x(f(u)) = 0$

Table 2 Comparison between the uncorrected scheme of [13] and our corrected scheme, $u_{\text{ini}} = 1$ on $(-\infty, 0)$, $u_{\text{ini}} = -1$ on $(0, \infty)$

	E_1	E_2	E_3	E_4
Uncorrected scheme	2.1E-1	3.9E-1	1.3	1.3
Corrected scheme	5.3E-2	2.3E-2	3.2E-4	4.2E-5

the initial data. Table 1 shows the values of $(E_m)_{m=1,\ldots,4}$ for both the uncorrected scheme from [13] based on (6) and our corrected scheme based on (7).

Test 2 (smooth shock): the initial condition is $u_{\text{ini}}(x) = 1$ if $x < 0$ and $u_{\text{ini}}(x) = -1$ if $x > 0$. Here the hyperbolic and non-local terms in (1) compete to maintain or diffuse the initial shock. Since α_m is near 2 however, any solution is instantly smooth, but has much larger gradients near $x = 0$ than the solution in Test 1 (Table 2).

Both tests confirm that the scheme defined by (4), (5), (7) and (8) is asymptotic-preserving. They also confirm that, without the order 2 correction in (7), the scheme deteriorates as $\alpha \to 2$ and does not provide a correct numerical solution at any reasonable resolution. This is also illustrated in Fig. 1, where we plot the solutions of both schemes for $\alpha = 1.99$ for the initial condition of Test 2 and $(\delta x, \delta t) = (\frac{1}{2^4 \times 50}, \frac{1}{2^4 \times 100})$. Even at this very high resolution, the uncorrected scheme provides an incorrect approximate solution which, as expected, is closer to the solution of $\partial_t u + \partial_x(f(u)) = 0$ than to the solution of (1).

5 Conclusion

We have presented a monotone numerical method for fractional conservation laws which is asymptotic-preserving with respect to the fractional power of the Laplacian. The scheme automatically adjusts to the change of nature of the equation as the power of the Laplacian goes to 1 (i.e. $\alpha \to 2$ in (1)) and therefore provides accurate approximate solutions for any power of the fractional Laplacian. We have given numerical results to illustrate the asymptotic-preserving property of our method, as well as the necessity of modifying previously studied monotone methods to obtain this property.

The complete theoretical study of such monotone asymptotic-preserving schemes will be presented in the forthcoming paper [16]. Here a general class of fractional degenerate parabolic equations are considered that include (1) as a special case.

References

1. Alfaro, M., Droniou, J.: General fractal conservation laws arising from a model of detonations in gases. Appl. Math. Res. Express **2012**, 127–151 (2012)
2. Alibaud, N.: Entropy formulation for fractal conservation laws. J. Evol. Equ. **7**(1), 145–175 (2007)
3. Alibaud, N., Andreianov, B.: Non-uniqueness of weak solutions for the fractal Burgers equation. Ann. Inst. H. Poincaré Anal. Non Linéaire **27**(4), 997–1016 (2010)
4. Alibaud, N., Droniou, J., Vovelle, J.: Occurrence and non-appearance of shocks in fractal burgers equations. J. Hyperbolic Differ. Equ. **4**(3), 479–499 (2007)
5. Asmussen, S., Rosiński, J.: Approximations of small jumps of Lévy processes with a view towards simulation. J. Appl. Probab. **38**(2), 482–493 (2001)
6. Biler, P., Karch, G., Woyczynski, W.: Fractal Burgers equations. J. Diff. Equ. **148**, 9–46 (1998)
7. Chan, R., Ng, M.: Conjugate gradient methods for toeplitz systems. SIAM Rev. **38**(3), 427–482 (1996)
8. Cifani, S., Jakobsen, E.R.: On numerical methods and error estimates for degenerate fractional convection-diffusion equations. To appear in Numer. Math. doi:10.1007/s00211-013-0590-0
9. Cifani, S., Jakobsen, E.R.: On the spectral vanishing viscosity method for periodic fractional conservation laws. Math. Comp. **82**(283), 1489–1514 (2013)
10. Cifani, S., Jakobsen, E.R., Karlsen, K.H.: The discontinuous galerkin method for fractal conservation laws. IMA J. Numer. Anal. **31**(3), 1090–1122 (2011)
11. Clavin, P.: Instabilities and Nonlinear Patterns of Overdriven Detonations in Gases. Kluwer (2002)
12. Cont, R., Tankov, P.: Financial modelling with jump processes. Chapman & Hall/CRC Financial Mathematics Series. Chapman and Hall/CRC, Boca Raton (2004)
13. Droniou, J.: A numerical method for fractal conservation laws. Math. Comp. **79**(269), 95–124 (2010)
14. Droniou, J., Gallouët, T., Vovelle, J.: Global solution and smoothing effect for a non-local regularization of an hyperbolic equation. J. Evol. Equ. **3**(3), 499–521 (2003)
15. Droniou, J., Imbert, C.: Fractal first order partial differential equations. Arch. Ration. Mech. Anal. **182**(2), 299–331 (2006)
16. Droniou, J., Jakobsen, E.R.: An asymptotic-preserving scheme for fractal conservation laws and fractional degenerate parabolic equations (In preparation)

17. Eymard, R., Gallouët, T., Herbin, R.: Finite volume methods. In: Ciarlet, P.G., Lions, J.L. (eds.) Techniques of Scientific Computing, Part III, Handbook of Numerical Analysis, VII, pp. 713–1020. North-Holland, Amsterdam (2000)
18. Godlewski, E., Raviart, P.A.: Numerical approximation of hyperbolic systems of conservation laws. Applied Mathematical Sciences, vol. 118. Springer, New-York (1996)
19. Holden, H., Risebro, N.H.: Front Tracking for Hyperbolic Conservation Laws. Springer, New York (2002)
20. Jakobsen, E.R., Karlsen, K.H., La Chioma, C.: Error estimates for approximate solutions to Bellman equations associated with controlled jump-diffusions. Numer. Math. **110**(2), 221–255 (2008)
21. Jourdain, B., Méléard, S., Woyczynski, W.: Probabilistic approximation and inviscid limits for one-dimensional fractional conservation laws. Bernoulli **11**(4), 689–714 (2005)
22. Kruzhkov, S.N.: First order quasilinear equations with several independent variables. Math. Sb. (N.S.) **81**(123), 228–255 (1970)
23. Soner, H.: Optimal control with state-space constraint ii. SIAM J. Control Optim. **24**(6), 1110–1122 (1986)
24. Stanescu, D., Kim, D., Woyczynski, W.: Numerical study of interacting particles approximation for integro-differential equations. J. Comput. Phys. **206**, 706–726 (2005)
25. Van Loan, C.: Computational Frameworks for the Fast Fourier Transform. Frontiers in Applied Mathematics, vol. 10. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA (1992)
26. Woyczynski, W.: Lévy Processes in the Physical Sciences. Birkhäuser, Boston (2001)

Uniform-in-Time Convergence of Numerical Schemes for Richards' and Stefan's Models

Jérôme Droniou, Robert Eymard and Cindy Guichard

Abstract We prove that all Gradient Schemes—which include Finite Element, Mixed Finite Element, Finite Volume methods—converge uniformly in time when applied to a family of nonlinear parabolic equations which contains Richards and Stefan's models. We also provide numerical results to confirm our theoretical analysis.

1 Introduction

Let us consider the following generic nonlinear parabolic model

$$\begin{aligned} \partial_t \beta(\overline{u}) - \Delta \zeta(\overline{u}) &= f \text{ in } \Omega \times (0,T), \\ \beta(\overline{u})(x,0) &= \beta(u_{\text{ini}})(x) \text{ in } \Omega, \\ \zeta(\overline{u}) &= 0 \text{ on } \partial\Omega \times (0,T), \end{aligned} \tag{1}$$

where β, ζ are non-decreasing. This model includes both Richards' model (with $\zeta(s) = s$), which describes the flow of water in an underground medium, and Stefan's model (with $\beta(s) = s$), which arises in the study of the heat diffusion in a melting medium. The numerical approximation of both Richards' and Stefan's models has

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

R. Eymard
Laboratoire d'Analyse et de Mathématiques Appliquées, CNRS, UPEM, UPEC, 5 boulevard Descartes, Champs-sur-Marne, 77454 Marne-la-Vallée Cedex 2, France
e-mail: Robert.Eymard@u-pem.fr

C. Guichard
Laboratoire Jacques-Louis Lions, Sorbonne Universités, UPMC Univ Paris 06, UMR 7598, F-75005 Paris, France
e-mail: guichard@ljll.math.upmc.fr

J. Fuhrmann et al. (eds.), *Finite Volumes for Complex Applications VII - Methods and Theoretical Aspects,* Springer Proceedings in Mathematics & Statistics 77,
DOI: 10.1007/978-3-319-05684-5_23,

been extensively studied in the literature (see the fundamental work on the Stefan problem [13, 14] for a review of some numerical approximations, and see [12] for the Richards problem), but the convergence analysis of the considered schemes received a much reduced coverage and consists mostly in establishing space-time averaged (e.g. in $L^2(\Omega \times (0,T))$ results (in the case of finite volume schemes, see for example [6, 9]). Yet, the quantity of interest is often not $\overline{u}$ on $\Omega \times (0,T)$ but $\overline{u}$ at a given time, for example $t = T$. Existing numerical analysis results therefore do not ensure that this quantity of interest is really properly approximated by numerical methods.

The usual way to obtain pointwise-in-time approximation results for numerical schemes is to prove estimates in $L^\infty(0,T;L^2(\Omega))$ on $u - \overline{u}$, where u is the approximated solution. Establishing such error estimates is however only feasible when uniqueness of the solution $\overline{u}$ to (1) can be proved (which is the case for Richards' and Stefan's problem, but not for more complex non-linear parabolic problems) and requires moreover some regularity assumptions on $\overline{u}$. These assumptions clearly fail for (1) for which, because of the possible plateaux of β and ζ, the solution can develop jumps in its gradient.

The purpose of this article is to prove that, using Discrete Functional Analysis techniques (i.e. the translation to numerical analysis of nonlinear analysis techniques), one can establish an $L^\infty(0,T;L^2(\Omega))$ convergence result for numerical approximations of (1) without having to assume non-physical regularity assumptions on the data. Note that, although Richards' and Stefan's models are formally equivalent when β and ζ are strictly increasing (consider $\beta = \zeta^{-1}$ to pass from one model to the other), they change nature when these functions are allowed to have plateaux. Richards' model can degenerate to an ODE (if $\zeta = 0$ on the range of u_{ini}) and Stefan's model can become a non-transient elliptic equation (if $\beta = 0$). The technique we develop in this paper is however generic enough to work directly on (1), as well as on a vast number of numerical methods.

That being said, we nevertheless require a particular numerical framework to work in, in order to write precise equations and estimates. The framework we choose is that of Gradient Schemes, which has the double benefit of covering a vast number of numerical methods—Finite Element schemes, Mimetic Finite Difference schemes, Finite Volume schemes, etc.—and of having already been studied for many models—elliptic, parabolic, linear or non-linear, possibly degenerate, etc. We refer the reader to [3–5, 8, 10] for more details.

The paper is organised as follows. In the next section, we present the assumption and the notion of weak solution for (1). Section 3 presents the Gradient Schemes for (1). In Sect. 4, we state our uniform convergence result and give a short proof of it, based on the space-time averaged convergence results available in the literature. Finally, Sect. 5 provides some numerical results to illustrate our uniform-in-time convergence theorem.

Note that more complete proofs, as well as an entirely unified convergence analysis (not relying on previous convergence results) of Gradient Schemes for a more general and more non-linear model than (1), can be found in [2].

2 Assumptions and Weak Solution for (1)

The notion of solution of (1) is that of a weak one in the following sense

$$\begin{cases} \overline{u} \in L^2(\Omega \times (0,T)) \,,\ \zeta(\overline{u}) \in L^2(0,T; H_0^1(\Omega)) \,, \partial_t \beta(\overline{u}) \in L^2(0,T; H^{-1}(\Omega)), \\ \beta(\overline{u}) \in C([0,T]; L^2(\Omega)\text{-w}), \\ \beta(\overline{u})(\cdot,0) = \beta(u_{\rm ini}) \text{ in } L^2(\Omega), \\ \displaystyle\int_0^T \langle \partial_t \beta(\overline{u})(\cdot,t), \overline{v}(\cdot,t)\rangle_{H^{-1},H_0^1} \mathrm{d}t + \int_0^T \int_\Omega \nabla \zeta(\overline{u})(x,t) \cdot \nabla \overline{v}(x,t) \mathrm{d}x \mathrm{d}t \\ \quad = \displaystyle\int_0^T \int_\Omega f(x,t) \overline{v}(x,t) \mathrm{d}x \mathrm{d}t \,, \qquad \forall \overline{v} \in L^2(0;T; H_0^1(\Omega)) \end{cases} \tag{2}$$

where $C([0,T], L^2(\Omega)\text{-w})$ is the set of functions $[0,T] \to L^2(\Omega)$ which are continuous for the weak topology of $L^2(\Omega)$. We assume throughout this paper that

$$\begin{aligned} &\beta, \zeta : \mathbb{R} \mapsto \mathbb{R} \text{ are non-decreasing and Lipschitz-continuous,} \\ &\beta(0) = \zeta(0) = 0 \text{ and } \exists A, B > 0 \text{ such that, for all } s \in \mathbb{R}, |\zeta(s)| \geq A|s| - B, \end{aligned} \tag{3a}$$

$$\beta = \mathrm{Id} \text{ or } \zeta = \mathrm{Id} \quad (\textit{we let } \gamma = \zeta \textit{ if } \beta = \mathrm{Id} \textit{ and } \gamma = \beta \textit{ if } \zeta = \mathrm{Id}) \tag{3b}$$

$$\begin{aligned} &\Omega \text{ is an open bounded subset of } \mathbb{R}^d, d \in \mathbb{N}^\star, \\ &u_{\rm ini} \in L^2(\Omega), \quad f \in L^2(\Omega \times (0,T)). \end{aligned} \tag{3c}$$

Under these assumptions, the weak continuity of $\beta(\overline{u}) : [0,T] \mapsto L^2(\Omega)$ is actually a consequence of the other regularity properties on $\overline{u}$, $\zeta(\overline{u})$, $\beta(\overline{u})$ and of the equation, see [2].

3 Gradient Scheme

The presentation of Gradient Schemes given here is minimal, we refer the reader to [3, 4, 7] for more details. A gradient scheme can be viewed as a general formulation of several discretisations of (1) which are based on approximations of the weak formulation (2). These approximations are based on some discrete spaces and mappings, the set of which we call a gradient discretisation.

Definition 1 We say that $\mathscr{D} = (X_{\mathscr{D},0}, \Pi_{\mathscr{D}}, \nabla_{\mathscr{D}}, \mathscr{I}_{\mathscr{D}}, (t^{(n)})_{n=0,\ldots,N})$ is a gradient discretisation for (1) if

1. $X_{\mathscr{D},0}$ is a finite dimensional real vector space (set of unknowns),
2. the linear mapping $\Pi_{\mathscr{D}} : X_{\mathscr{D},0} \to L^\infty(\Omega)$ is a piecewise constant reconstruction operator, that is there exists a set I of degrees of freedom such that $X_{\mathscr{D},0} = \mathbb{R}^I$ and there exists a family $(\Omega_i)_{i\in I}$ of disjoint subsets of Ω such that $\overline{\Omega} = \bigcup_{i\in I} \overline{\Omega_i}$ and, for all $u = (u_i)_{i\in I} \in X_{\mathscr{D},0}$ and all $i \in I, \Pi_{\mathscr{D}} u = u_i$ on Ω_i,
3. the linear mapping $\nabla_{\mathscr{D}} : X_{\mathscr{D},0} \to L^2(\Omega)^d$ gives a reconstructed discrete gradient.

4. $\mathscr{I}_{\mathscr{D}} : L^2(\Omega) \to X_{\mathscr{D},0}$ is a linear interpolation operator,
5. $t^{(0)} = 0 < t^{(1)} < t^{(2)} < \ldots < t^{(N)} = T$.

For any function $\chi : \mathbb{R} \mapsto \mathbb{R}$ and any $u \in X_{\mathscr{D},0}$, we denote by $\chi(u) \in X_{\mathscr{D},0}$ the element defined by $(\chi(u))_i = \chi(u_i)$ for any $i \in I$. Since $\Pi_{\mathscr{D}}$ is a piecewise constant reconstruction, we then have $\Pi_{\mathscr{D}}\chi(u) = \chi(\Pi_{\mathscr{D}}u)$. It is also customary to use the notations $\Pi_{\mathscr{D}}$ and $\nabla_{\mathscr{D}}$ for space-time dependent functions. We will also need a notation for the jump-in-time of piecewise constant functions in time. Hence, if $(v^{(n)})_{n=0,\ldots,N} \subset X_{\mathscr{D},0}$, we set

$$\begin{aligned}&\text{for a.e. } x \in \Omega,\ \Pi_{\mathscr{D}}v(x,0) = \Pi_{\mathscr{D}}v^{(0)}(x) \text{ and } \forall n = 0,\ldots,N-1,\ \forall t \in (t^{(n)}, t^{(n+1)}] :\\ &\quad \Pi_{\mathscr{D}}v(x,t) = \Pi_{\mathscr{D}}v^{(n+1)}(x),\ \nabla_{\mathscr{D}}v(x,t) = \nabla_{\mathscr{D}}v^{(n+1)}(x)\\ &\quad \text{and } \delta_{\mathscr{D}}v(t) = \delta_{\mathscr{D}}^{(n+\frac{1}{2})}v := \frac{v^{(n+1)} - v^{(n)}}{t^{(n+1)} - t^{(n)}}.\end{aligned}$$

With these notations, the gradient scheme corresponding to a given gradient discretisation $\mathscr{D}$ is obtained by replacing the continuous functions and gradients in (2) with their discrete counterpart and using an implicit-in-time discretisation. It is therefore written: find $(u^{(n)})_{n=0,\ldots,N} \subset X_{\mathscr{D},0}$ such that

$$\left\{\begin{aligned}&u^{(0)} = \mathscr{I}_{\mathscr{D}}u_{\rm ini} \text{ and, for all } v = (v^{(n)})_{n=0,\ldots,N} \subset X_{\mathscr{D},0},\\ &\int_0^T \int_\Omega \Big[\Pi_{\mathscr{D}}\delta_{\mathscr{D}}\beta(u)(x,t)\Pi_{\mathscr{D}}v(x,t) + \nabla_{\mathscr{D}}\zeta(u)(x,t)\cdot\nabla_{\mathscr{D}}v(x,t)\Big]{\rm d}x{\rm d}t\\ &\quad = \int_0^T \int_\Omega f(x,t)\Pi_{\mathscr{D}}v(x,t){\rm d}x{\rm d}t.\end{aligned}\right. \tag{4}$$

As mentioned in the introduction, gradient schemes cover a wide number of well-known numerical methods [3]. Their convergence analysis is moreover based on a few (four, to be precise) properties that a gradient discretisation must satisfy: *coercivity*, *consistency*, *limit-conformity* and *compactness*. As we will not directly make much use of these properties but only of the following initial convergence result, we just refer the reader to [3, 4] for their precise definition.

Theorem 1 ([5, 8]) *Under Assumption (3a)–(3c), there exists a unique solution to the gradient scheme (4). Moreover, if $(\mathscr{D}_m)_{m\in\mathbb{N}}$ is a coercive, consistent, limit-conforming and compact sequence of gradient discretisations, if $(u_m)_{m\in\mathbb{N}}$ are the solutions to the corresponding gradient schemes and if $\overline{u}$ is the solution to (2) then, as $m \to \infty$, $\Pi_{\mathscr{D}_m}u_m \to \overline{u}$ weakly in $L^2(\Omega \times (0,T))$, $\Pi_{\mathscr{D}_m}\gamma(u_m) \to \gamma(\overline{u})$ in $L^2(\Omega \times (0,T))$ and $\nabla_{\mathscr{D}_m}\zeta(u_m) \to \nabla\zeta(\overline{u})$ in $L^2(\Omega \times (0,T))^d$.*

4 Uniform Convergence Result

Our main result is the following. As mentioned in the introduction, we only sketch its proof and refer the reader to [2] for the details.

Theorem 2 ([2]) Under the assumptions and notations of Theorem 1, $\Pi_{\mathscr{D}_m}\gamma(u_m) \to \gamma(\overline{u})$ strongly in $L^\infty(0,T;L^2(\Omega))$.

Proof The keys to this proof are the following integration-by-parts properties satisfied by the continuous and discrete solutions. Defining $\beta_r(s) =$ closest z to 0 such that $\beta(z) = s$ (pseudo-inverse of β) and $B(z) = \int_0^z \zeta(\beta_r(s))\mathrm{d}s$, we have, for any $T_0 \in (0,T]$,

$$\int_\Omega B(\beta(\overline{u})(x,T_0))\mathrm{d}x + \int_0^{T_0}\int_\Omega \nabla\zeta(\overline{u})(x,t)\cdot\nabla\zeta(\overline{u})(x,t)\mathrm{d}x\mathrm{d}t = \int_\Omega B(\beta(u_{\mathrm{ini}}(x)))\mathrm{d}x + \int_0^{T_0}\int_\Omega f(x,t)\zeta(\overline{u})(x,t)\mathrm{d}x\mathrm{d}t \tag{5}$$

and

$$\int_\Omega B(\Pi_{\mathscr{D}_m}\beta(u_m)(x,T_0))\mathrm{d}x + \int_0^{T_0}\int_\Omega \nabla_{\mathscr{D}_m}\zeta(u_m)(x,t)\cdot\nabla_{\mathscr{D}_m}\zeta(u_m)(x,t)\mathrm{d}x\mathrm{d}t \le \int_\Omega B(\Pi_{\mathscr{D}_m}\beta(\mathscr{I}_{\mathscr{D}_m}u_{\mathrm{ini}})(x))\mathrm{d}x + \int_0^{t^{(k)}}\int_\Omega f(x,t)\Pi_{\mathscr{D}_m}\zeta(u_m)(x,t)\mathrm{d}x\mathrm{d}t, \tag{6}$$

where $k \in \{1,\dots,N\}$ is such that $t^{(k-1)} < T_0 \le t^{(k)}$. These formula are obtained by plugging respectively $\bar{v} = \zeta(\overline{u})$ and $v = u_m$ in (2) and (4). Properly justifying (5) is however not straightforward because of the lack of regularity of $\overline{u}$.

Let $T_0 \in [0,T]$ and $(T_m)_{m\in\mathbb{N}}$ which converges to T_0. We apply (6) to $T_0 = T_m$ and let $m \to \infty$. The consistency of $(\mathscr{D}_m)_{m\in\mathbb{N}}$ ensures that $\mathscr{I}_{\mathscr{D}_m}u_{\mathrm{ini}} \to u_{\mathrm{ini}}$ in $L^2(\Omega)$. Hence, using the strong convergence in $L^2(\Omega\times(0,T))^d$ of $\nabla_{\mathscr{D}_m}\zeta(u_m)$ to $\nabla\zeta(\overline{u})$ and (5), we find

$$\limsup_{m\to\infty}\int_\Omega B(\Pi_{\mathscr{D}_m}\beta(u_m)(x,T_m))\mathrm{d}x \le \int_\Omega B(\beta(\overline{u})(x,T_0))\mathrm{d}x. \tag{7}$$

Using the scheme (4), we easily obtain, for any $\varphi \in L^2(\Omega)$, estimates on the variations of $t \mapsto \langle\Pi_{\mathscr{D}_m}\beta(u_m)(t),\varphi\rangle_{L^2(\Omega)}$ which show that $(\Pi_{\mathscr{D}_m}\beta(u_m))_{m\in\mathscr{D}_m}$ is relatively compact in $L^\infty(0,T;L^2(\Omega)\text{-w})$ and therefore converges uniformly in time for the weak topology of $L^2(\Omega)$. We deduce that $\Pi_{\mathscr{D}_m}\beta(u_m)(T_m) \to \beta(\bar{u})(T_0)$ weakly in $L^2(\Omega)$ and the convexity of B therefore ensures that

$$\int_\Omega B(\beta(\overline{u})(x,T_0))\mathrm{d}x \le \liminf_{m\to\infty}\int_\Omega B(\Pi_{\mathscr{D}_m}\beta(u_m)(x,T_m))\mathrm{d}x. \tag{8}$$

Combining (7) and (8) we find that

$$\lim_{m\to\infty}\int_\Omega B(\Pi_{\mathscr{D}_m}\beta(u_m)(x,T_m))\mathrm{d}x=\int_\Omega B(\beta(\overline{u})(x,T_0))\mathrm{d}x. \tag{9}$$

We also notice that, by weak convergence in $L^2(\Omega)$ of $\Pi_{\mathscr{D}_m}\beta(u_m)(T_m)$ to $\beta(\overline{u})(T_0)$ and the convexity of B,

$$\int_\Omega B(\beta(\overline{u})(x,T_0))\mathrm{d}x\le\liminf_{m\to\infty}\int_\Omega B\left(\frac{\Pi_{\mathscr{D}_m}\beta(u_m)(x,T_m)+\beta(\overline{u})(x,T_0)}{2}\right)\mathrm{d}x. \tag{10}$$

The definition of B ensures that, for all $s,s'\in\mathbb{R}$,

$$(\gamma(s)-\gamma(s'))^2\le C_1\left[B(\beta(s))+B(\beta(s'))-2B\left(\frac{\beta(s)+\beta(s')}{2}\right)\right]$$

where C_1 only depends on the Lipschitz constants of β and ζ. We deduce that

$$\begin{aligned}&\|\gamma(\Pi_{\mathscr{D}_m}u_m)(\cdot,T_m)-\gamma(\overline{u})(\cdot,T_0)\|^2_{L^2(\Omega)}\\&\le C_1\int_\Omega\left[B(\beta(\Pi_{\mathscr{D}_m}u_m)(x,T_m))+B(\beta(\overline{u})(x,T_0))\right]\mathrm{d}x\\&\quad-2C_1\int_\Omega B\left(\frac{\beta(\Pi_{\mathscr{D}_m}u_m)(x,T_m)+\beta(\overline{u})(x,T_0)}{2}\right)\mathrm{d}x.\end{aligned}$$

Taking the lim sup as $m\to\infty$ of this relation and using (9) and (10), we find that $\Pi_{\mathscr{D}_m}\gamma(u_m(\cdot,T_m))\to\gamma(\overline{u})(\cdot,T_0)$ in $L^2(\Omega)$ as $m\to\infty$. Since this is true for any sequence $T_m\to T_0$, and since we can prove that $\gamma(\overline{u})$ is continuous $[0,T]\mapsto L^2(\Omega)$, this proves that $\Pi_{\mathscr{D}_m}\gamma(u_m)\to\gamma(\overline{u})$ uniformly on $[0,T]$ for the topology of $L^2(\Omega)$.

5 Numerical Tests

In order to illustrate the uniform-in-time convergence properties, we first present the gradient scheme which has been selected for running the test cases. The gradient scheme is built on a conforming simplicial mesh of the polyhedral domain Ω (see [1] for the precise definitions of such a mesh). The degrees of freedom of any $u\in X_{\mathscr{D},0}$ are the values u_s for all interior vertices s of the mesh. Then $\Pi_{\mathscr{D}}u$ is taken piecewise constant in the regions K_s (see Fig. 1), whereas $\nabla_{\mathscr{D}}u$ is the gradient of the P^1 finite element function obtained from the values u_s.

For both following tests, the meshes used for the discretisation of the domain $\Omega=(0,1)^2$ come from from the FVCA5 2D benchmark on anisotropic diffusion problem [11]. These triangle meshes show no symmetry which could artificially increase the convergence rate, and all angles of triangles are acute. This family of meshes is built using the same pattern reproduced at different scales: the first (coarsest) mesh and the third mesh are shown in Fig. 2. We consider the two cases

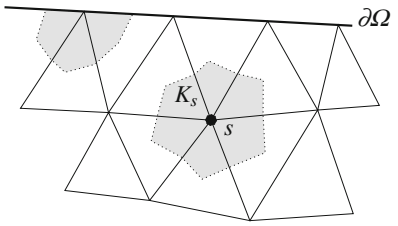


Fig. 1 Definition of K_s

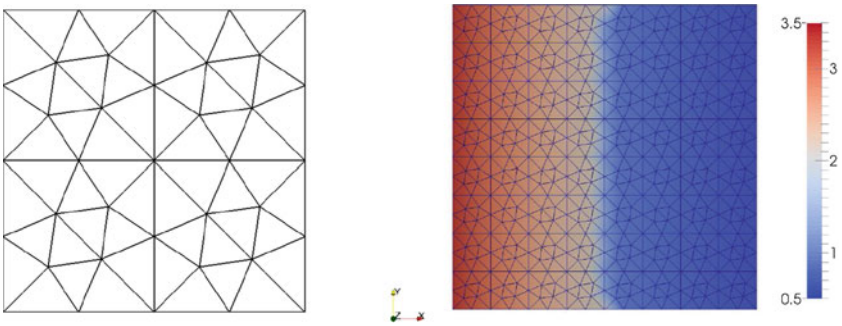


Fig. 2 First mesh and approximate solution u at time 0.5 for the Stefan problem on the third mesh

Table 1 Data and numerical results for Stefan and Richards problems

test case	Stefan	Richards
$\beta(u)$	u	$\begin{cases} 0 & \text{if } u \le 0 \\ u & \text{if } 0 \le u \end{cases}$
$\zeta(u)$	$\begin{cases} u & \text{if } u \le 1 \\ 1 & \text{if } 1 \le u \le 2 \\ u-1 & \text{if } 2 \le u \end{cases}$	u
analytical sol. $u(x_1,x_2,t)$	$\begin{cases} 2\exp(t-x_1)\ (>2) & \text{if } x_1 < t \\ \exp(t-x_1)\ (<1) & \text{if } t < x_1 \end{cases}$	$\begin{cases} \exp(t-x_1)-1\ (\ge 0) & \text{if } x_1 \le t \\ t-x_1\ (\le 0) & \text{if } t \le x_1 \end{cases}$

h	δt	error on $\zeta(u)$ in $L^\infty(0,T;L^2(\Omega))$	num. order	error on $\beta(u)$ in $L^\infty(0,T;L^2(\Omega))$	num. order
0.250	0.01024	0.362E-01	-	0.223E-02	-
0.125	0.00256	0.191E-01	0.920	0.891E-03	1.327
0.063	0.00064	0.895E-02	1.096	0.297E-03	1.584
0.031	0.00016	0.392E-02	1.192	0.101E-03	1.562
0.016	0.00004	0.175E-02	1.166	0.334E-04	1.590

of a Stefan problem and of a Richards problems, for which there exists an analytical solution with $f = 0$. These analytical solutions show the regularity properties of "natural" solutions on the time period [0, 1] during which a free boundary moves from $x_1 = 0$ to $x_1 = 1$. In the Stefan problem, this free boundary is the surface between two thermodynamical states of a material. In the Richards problem, it is the limit between a fully saturated zone and a partially saturated zone. These test cases

are built on 1D solutions, using fully 2D meshes, hence providing realistic conditions (an example of a numerical solution is shown in the right part of Fig. 2). For both of them, the corresponding data and numerical results are given in Table 1. The convergence orders are computed from the values of h, and the constant time steps have been taken proportional to h^2. Note that in both cases, the proposed analytical solution is a strong solution for $x_1 < t$ and $x_1 > t$ and the Rankine-Hugoniot condition holds at the free boundary $x_1 = t$. It is therefore a weak solution to (2), extended to the case of non-homogeneous Dirichlet boundary conditions. These results confirm our uniform-in-time convergence result (Theorem 2). We also observe that in the Stefan case, where u is discontinuous and $\zeta(u)$ is only of class H^1, the convergence order remains close to 1 whereas in the Richards case, where u is of class C^1 in space, the convergence orders are greater.

References

1. Ciarlet, P.: The finite element method. In: Ciarlet, P.G., Lions, J.L. (eds.) Part I, Handbook of Numerical Analysis. III. North-Holland, Amsterdam (1991)
2. Droniou, J., Eymard, R.: Uniform-in-time convergence results of numerical methods for non-linear parabolic equations. http://hal.archives-ouvertes.fr/hal-00949682 (2014)
3. Droniou, J., Eymard, R., Gallouët, T., Guichard, C., Herbin, R.: Gradient schemes for elliptic and parabolic problems (2014). (In preparation)
4. Droniou, J., Eymard, R., Gallouët, T., Herbin, R.: Gradient schemes: a generic framework for the discretisation of linear, nonlinear and nonlocal elliptic and parabolic equations. Math. Models Methods Appl. Sci. (M3AS) **23**(13), 2395–2432 (2013)
5. Eymard, R., Féron, P., Gallouet, T., Herbin, R., Guichard, C.: Gradient schemes for the Stefan problem. Int. J. Finite Vol. **10s** (2013). http://hal.archives-ouvertes.fr/hal-00751555
6. Eymard, R., Gallouët, T., Hilhorst, D., Naït Slimane, Y.: Finite volumes and nonlinear diffusion equations. RAIRO Modél. Math. Anal. Numér. **32**(6), 747–761 (1998)
7. Eymard, R., Guichard, C., Herbin, R.: Small-stencil 3d schemes for diffusive flows in porous media. M2AN **46**, 265–290 (2012)
8. Eymard, R., Guichard, C., Herbin, R., Masson, R.: Gradient schemes for two-phase flow in heterogeneous porous media and Richards equation. ZAMM p. accepted for publication (2013)
9. Eymard, R., Gutnic, M., Hilhorst, D.: The finite volume method for Richards equation. Comput. Geosci. **3**(3–4), 259–294 (2000) (1999). doi:10.1023/A:1011547513583. http://dx.doi.org/10.1023/A:1011547513583
10. Eymard, R., Herbin, R.: Gradient scheme approximations for diffusion problems. In: Finite Volumes for Complex Applications VI Problems & Perspectives, pp. 439–447 (2011)
11. Herbin, R., Hubert, F.: Benchmark on discretization schemes for anisotropic diffusion problems on general grids. In: Finite volumes for complex applications V, pp. 659–692. ISTE, London (2008)
12. Maitre, E.: Numerical analysis of nonlinear elliptic-parabolic equations. M2AN Math. Model. Numer. Anal. **36**(1), 143–153 (2002). doi:10.1051/m2an:2002006. http://dx.doi.org/10.1051/m2an:2002006
13. Nochetto, R.H., Verdi, C.: Approximation of degenerate parabolic problems using numerical integration. SIAM J. Numer. Anal. **25**(4), 784–814 (1988)
14. Pop, I.S.: Numerical schemes for degenerate parabolic problems. In: Progress in Industrial Mathematics at ECMI 2004, Math. Ind., vol. 8, pp. 513–517. Springer, Berlin (2006)

Comparison of Two Couplings of the Finite Volume Method and the Boundary Element Method

Christoph Erath

Abstract In many fluid dynamics problems the boundary conditions may be unknown, or the domain may be unbounded. Also mass conservation and stability with respect to dominating convection is substantial. Therefore, we test two coupling methods to address these issues on the prototype of a flow and transport problem. More precisely, we couple the vertex-centered and the cell-centered finite volume method with the boundary element method, FVM-BEM and CFVM-BEM, respectively. Also robust refinement indicators are considered which allow us to steer an adaptive mesh-refinement algorithm to treat efficiently problems with singularities or boundary/internal layers—shown on two examples.

1 Model Problem and Notation

Due to the conservation of mass property and a stable approximation for convection dominated problems finite volume methods are well established in fluid dynamics. Boundary element methods can be used if the fundamental solution of the problem is known. Since they reduce the approximation problem from a domain to its boundary, they can be employed for problems on unbounded domains (with radiation conditions) without truncating the domain. In a sense they also feature local conservation. Two coupling methods of both schemes are considered here to benefit and merge their properties. In an interior domain $\Omega \subset \mathbb{R}^d$ $(d = 2, 3)$, which is a bounded and simply connected domain with polygonal/polyhedral Lipschitz boundary Γ, we consider the prototype of a flow and transport problem and discretize it with a FVM. Whereas in the corresponding unbounded exterior domain $\Omega_e = \mathbb{R}^d \backslash \overline{\Omega}$ we approximate a diffusive process with the BEM. Special care has to be taken on the so called

C. Erath (✉)
Department of Mathematics, University of Vienna, Oskar-Morgenstern-Platz 1, Vienna 1090, Austria
e-mail: christoph.erath@univie.ac.at

J. Fuhrmann et al. (eds.), *Finite Volumes for Complex Applications VII - Methods and Theoretical Aspects,* Springer Proceedings in Mathematics & Statistics 77,
DOI: 10.1007/978-3-319-05684-5_24,

coupling boundary $\Gamma = \partial\Omega = \partial\Omega_e$, which is divided in an inflow and outflow part, namely $\Gamma^{in} := \{x \in \Gamma \,|\, \mathbf{b}(x) \cdot \mathbf{n}(x) < 0\}$ and $\Gamma^{out} := \{x \in \Gamma \,|\, \mathbf{b}(x) \cdot \mathbf{n}(x) \geq 0\}$, respectively. Here $\mathbf{n}$ is the normal vector on Γ pointing outward with respect to Ω. The mathematical formulation of our problem reads: Find u and u_e such that

$$\begin{aligned}
\operatorname{div}(-\alpha\nabla u + \mathbf{b}u) + \gamma u &= f && \text{in } \Omega, && \text{(1a)}\\
\Delta u_e &= 0 && \text{in } \Omega_e, && \text{(1b)}\\
u_e(x) &= a_\infty + b_\infty \log|x| + o(1) && \text{for} |x| \to \infty, d = 2, && \text{(1c)}\\
u_e(x) &= \mathcal{O}(|x|^{-1}) && \text{for} |x| \to \infty, d = 3, && \text{(1d)}\\
u &= u_e + u_0 && \text{on } \Gamma, && \text{(1e)}\\
(\alpha\nabla u - \mathbf{b}u)\cdot\mathbf{n} &= \frac{\partial u_e}{\partial \mathbf{n}} + t_0 && \text{on } \Gamma^{in}, && \text{(1f)}\\
(\alpha\nabla u)\cdot\mathbf{n} &= \frac{\partial u_e}{\partial \mathbf{n}} + t_0 && \text{on } \Gamma^{out} \quad . && \text{(1g)}
\end{aligned}$$

In the two dimensional case we can fix either $a_\infty \in \mathbb{R}$ or $b_\infty \in \mathbb{R}$; see [1]. The given data satisfy $f \in L^2(\Omega)$, $u_0 \in H^{1/2}(\Gamma)$ and $t_0 \in L^2(\Gamma)$, where $L^m(\cdot)$ and $H^m(\cdot)$, $m > 0$, denote the standard Lebesgue and Sobolev spaces equipped with the usual norms $\|\cdot\|_{L^2(\cdot)}$ and $\|\cdot\|_{H^m(\cdot)}$, respectively. The diffusion coefficient α is positive, and there holds $(\operatorname{div}\mathbf{b})/2 + \gamma \geq C_1 \geq 0$ and $\|\operatorname{div}\mathbf{b} + \gamma\|_{L^\infty(\Omega)} \leq C_2 C_1$ with the constants $C_1, C_2 \geq 0$ for the convection vector function $\mathbf{b} \in W^{1,\infty}(\Omega)^d$ (vector of Lipschitz continuous functions) and the reaction function $\gamma \in L^\infty(\Omega)$.

It is shown in [1] that in a weak sense there exists a unique solution $u \in H^1(\Omega)$ and $u_e \in H^1_{\ell oc}(\Omega_e)$ (set of all local H^1-functions) of the model problem (1). Although not explicitly stated in [1] the result is also valid (verbatim) for the 3-D case. The proof is based on the fact that we can transform the unbounded exterior problem (1b)–(1d) into an integral equation—the exterior Calderón system—with the Cauchy data $\xi := u_e|_\Gamma \in H^{1/2}(\Gamma)$ and $\phi := \partial u_e/\partial\mathbf{n}|_\Gamma \in H^{-1/2}(\Gamma)$. The weak form of this system and the interior weak form are coupled through the conditions (1e)–(1g). For more details we refer to [1] and only remark that the Calderón system is based on some bounded and linear integral operators $\mathcal{V} \in L\big(H^{s-1/2}(\Gamma); H^{s+1/2}(\Gamma)\big)$ (single layer op.), $\mathcal{K} \in L\big(H^{s+1/2}(\Gamma); H^{s+1/2}(\Gamma)\big)$ (double layer op.), $\mathcal{K}^* \in L\big(H^{s-1/2}(\Gamma); H^{s-1/2}(\Gamma)\big)$ (adjoint double layer op.) and $\mathcal{W} \in L\big(H^{s+1/2}(\Gamma); H^{s-1/2}(\Gamma)\big)$ (hypersingular integral op.) for $s \in [-1/2, 1/2]$. These operators are based on the fundamental solution $-\frac{1}{2\pi}\log|x|$ for the 2-D case and $\frac{1}{4\pi}\frac{1}{|x|}$ for the 3-D case of the exterior problem; for more details see e.g. [1].

Triangulation: To simplify notation and the language we only note the construction for the 2-D case. Throughout, $\mathcal{T}$ denotes a triangulation, the *primal mesh*, of Ω, where $\mathcal{N}$ and $\mathcal{E}$ are the corresponding set of nodes and edges, respectively. The notation in this work is consistent in the sense that $\mathcal{N}_I$ and $\mathcal{N}_\Gamma$ denote the set of nodes in the interior and on the boundary, respectively, $\mathcal{E}_\Gamma^{in} \subset \mathcal{E}_\Gamma$ denotes all coupling edges on Γ^{in}, $\mathcal{E}_T$ all edges of T, and so on. For brevity, the elements $T \in \mathcal{T}$ are

non-degenerated triangles. The Euclidean diameter of $T \in \mathscr{T}$ is $h_T := \sup_{x,y\in T} |x - y|$ and h_E denotes the length of an edge $E \in \mathscr{E}$. The triangulation is regular, i.e., the ratio h_T of any element $T \in \mathscr{T}$ to the diameter of its largest inscribed ball is bounded by a constant independent of h_T. Additionally, we assume that the triangulation $\mathscr{T}$ is aligned with the discontinuities (if any) of any given data, and $\mathbf{n}$ denotes the unit normal vector to the boundary pointing outward the domain.

Dual mesh: If we connect the center of gravity of an element $T \in \mathscr{T}$ with the midpoints of the edges $E \in \mathscr{E}_T$ we get the *dual mesh* $\mathscr{T}^*$ with its boxes $V \in \mathscr{T}^*$. A box associated with a vertex $a_i \in \mathscr{N}$ (from the primal mesh, $i = 1 \dots \#\mathscr{N}$, which lies in the box) is denoted by $V_i \in \mathscr{T}^*$. Note that this vertex is unique.

We denote by $\mathscr{C}(\cdot)$ all continuous functions. The L^2 scalar product is $(\cdot,\cdot)_\omega$, $\omega \subset \Omega$. The duality between $H^m(\Gamma)$ and $H^{-m}(\Gamma)$ is given by the extended L^2-scalar product $\langle\cdot,\cdot\rangle_\Gamma$. Moreover, we define the piecewise affine and globally continuous function space on $\mathscr{T}$ by $\mathscr{S}^1(\mathscr{T}) := \left\{v \in \mathscr{C}(\Omega) \,\middle|\, v|_T \text{ affine for all } T \in \mathscr{T}\right\}$ and the piecewise constant space on $\mathscr{T}$ by $\mathscr{P}^0(\mathscr{T}) := \left\{v \in L^2(\Omega) \,\middle|\, v|_T \text{ const. for all } T \in \mathscr{T}\right\}$. The spaces $\mathscr{S}^1(\mathscr{E}_\Gamma)$, $\mathscr{P}^0(\mathscr{E}_\Gamma)$, and $\mathscr{P}^0(\mathscr{T}^*)$ are equivalently defined as above and $\mathscr{S}^1_*(\mathscr{E}_\Gamma)$ is $\mathscr{S}^1(\mathscr{E}_\Gamma)$ with integral mean zero over $\mathscr{E}_\Gamma$.

2 FVM (Vertex-Centered) and BEM Coupling

A detailed description and motivation of this type of coupling can be found in [1]. The discrete system reads for $a_\infty = 0$: Find a discrete solution $u_h \in \mathscr{S}^1(\mathscr{T})$, $\xi_h \in \mathscr{S}^1_*(\mathscr{E}_\Gamma)$ and $\phi_h \in \mathscr{P}^0(\mathscr{E}_\Gamma)$ of our model problem such that

$$\mathscr{A}_V(u_h, v^*) - \left(\phi_h, v^*\right)_\Gamma = F(v^*), \tag{2a}$$

$$-\langle u_h, \psi_h\rangle_\Gamma - \langle \mathscr{V}\phi_h, \psi_h\rangle_\Gamma + \langle (1/2 + \mathscr{K})\xi_h, \psi_h\rangle_\Gamma = -\langle u_0, \psi_h\rangle_\Gamma\,, \tag{2b}$$

$$\left\langle (1/2 + \mathscr{K}^*)\phi_h, \theta_h\right\rangle_\Gamma + \langle \mathscr{W}\xi_h, \theta_h\rangle_\Gamma = 0 \tag{2c}$$

for all $v^* \in \mathscr{P}^0(\mathscr{T}^*)$ ($v^* := \sum_{x_i\in\mathscr{N}} v_i^*\chi_i^*$, $v_i^* \in \mathbb{R}$), $\theta_h \in \mathscr{S}^1_*(\mathscr{E}_\Gamma)$, $\psi_h \in \mathscr{P}^0(\mathscr{E}_\Gamma)$. The bilinear form $\mathscr{A}_V$ and the right-hand side $F(v^*)$ are defined as

$$\mathscr{A}_V(u_h, v^*) := \sum_{a_i\in\mathscr{N}} v_i^* \left(\int_{\partial V_i\setminus\Gamma} (-\alpha\nabla u_h + \mathbf{b}u_h)\cdot\mathbf{n}\, ds + \int_{V_i} \gamma u_h\, dx + \int_{\partial V_i\cap\Gamma^{out}} \mathbf{b}\cdot\mathbf{n}\, u_h\, ds \right),$$

$$F(v^*) := \sum_{a_i\in\mathscr{N}} v_i^* \left(\int_{V_i} f\, dx + \int_{\partial V_i\cap\Gamma} t_0\, ds \right).$$

Note that the discretization in the interior domain follows along the dual mesh $\mathscr{T}^*$, u_h approximates u, $\xi_h \approx \xi$, and $\phi_h \approx \phi$ and the two are coupled through ϕ_h in (2a) and u_h in the Calderón system (2b)–(2c). See [1, Remark 3.1] why ξ_h has to be the integral mean. If we want to apply a full upwind scheme for the finite volume scheme, we replace $\mathbf{b}u_h$ in $\mathscr{A}_V$ by its full upwind value $\mathbf{b}u_{h,ij}$. Note that there exists a $\tau_{ij} = V_i \cap V_j \neq \emptyset$ for $V_i, V_j \in \mathscr{T}^*$, i.e., τ_{ij} consists two straight lines and is a

part of ∂V_i and ∂V_j. With $\beta_{ij} := \big(\int_{\tau_{ij}} \mathbf{b}\cdot\mathbf{n}_i\, ds\big)/|\tau_{ij}|$ the upwind value is defined by $u_{h,ij} := u_h(a_i)$ if $\beta_{ij} \geq 0$, and $u_{h,ij} := u_h(a_j)$ otherwise. For a sufficient small mesh size the discrete solution of system (2) and its upwind version exists, is unique, and is of first order; see [1]. This result is also valid for three dimensions.

3 Cell-Centered FVM and BEM Coupling

The CFVM-BEM coupling reads for $a_\infty = 0$: Find $u_h \in \mathscr{P}^0(\mathscr{T})$, $u_{h,\Gamma} \in \mathscr{S}^1(\mathscr{E}_\Gamma)$, $\xi_h \in \mathscr{S}^1_*(\mathscr{E}_\Gamma)$ and $\phi_h \in \mathscr{P}^0(\mathscr{E}_\Gamma)$ such that

$$\sum_{E\in\mathscr{E}_T\setminus\mathscr{E}_\Gamma} F^D_{T,E}(u_h) + \sum_{E\in\mathscr{E}_T\setminus\mathscr{E}^{in}_\Gamma} F^C_{T,E}(u_h) + F^R_T(u_h) - \int_{\partial T\cap\Gamma} \phi_h\, ds = \int_T f\, dx + \int_{\partial T\cap\Gamma} t_0\, ds, \tag{3a}$$

$$-u_a + \overline{u}_a + \overline{\varsigma}_{a,h} = -\overline{\varsigma}_{a,t_0}, \tag{3b}$$

$$-\big\langle u_{h,\Gamma}, \psi_h\big\rangle_\Gamma - \langle \mathscr{V}\phi_h, \psi_h\rangle_\Gamma + \langle (1/2+\mathscr{K})\xi_h, \psi_h\rangle_\Gamma = -\langle u_0, \psi_h\rangle_\Gamma\,, \tag{3c}$$

$$\big\langle (1/2+\mathscr{K}^*)\phi_h, \theta_h\big\rangle_\Gamma + \langle \mathscr{W}\xi_h, \theta_h\rangle_\Gamma = 0 \tag{3d}$$

for all $T \in \mathscr{T}$, $a \in \mathscr{N}_\Gamma$, $\theta_h \in \mathscr{S}^1_*(\mathscr{E}_\Gamma)$ and $\psi_h \in \mathscr{P}^0(\mathscr{E}_\Gamma)$. A detailed description of this coupling method can be found in [2]. Note that the discretization in the interior domain follows along the primal mesh $\mathscr{T}$, u_h approximates u, $\xi_h \approx \xi$, and $\phi_h \approx \phi$. To allow local mesh-refinement we approximate the diffusion flux $F^D_{T,E}(u_h)$ by the diamond path method as in [4]. For the convection flux $F^C_{T,E}(u_h)$ we can choose the full upwind scheme as described in Sect. 2 and $F^R_T(u_h)$ is simply the integral of γ over T. The approximation of u_a is done by an interpolation value $\overline{u}_a$ of certain values u_T of $T \in \mathscr{T}$, see also [4], and a mean value $\overline{\varsigma}_a = \overline{\varsigma}_{a,h} + \overline{\varsigma}_{a,t_0}$. The latter is the approximated conormal of u_e on Γ, which is given by the solution ϕ_h of the boundary element method for the exterior problem and the jump term t_0. The piecewise affine discrete solution reads $u_{h,\Gamma} := \sum_{a\in\mathscr{N}_\Gamma} u_a\eta_a(x)$ with the standard nodal linear basis function η_a on $\mathscr{E}_\Gamma$. Note that the unknown u_a on Γ is also needed for the diamond path. CFVM and BEM are coupled through ϕ_h in (3a) and $u_{h,\Gamma}$ in the Calderón system (3c)–(3d). We want to point out that there is neither an existence proof nor an a priori result available for this type of coupling. Thus, we assume that this systems is well-defined and gives a unique solution.

4 A Posteriori Error Estimator

For convection or reaction dominated problems robust a posteriori estimators are essential. Therefore, we define $\beta_T := \min_{x\in T}\big\{(\operatorname{div}\mathbf{b}(x))/2 + \gamma(x)\big\}$ for all $T \in \mathscr{T}$ and $\beta_E := \min\big\{\beta_{T_1}, \beta_{T_2}\big\}$ for $E \in \mathscr{E}_I$ with $E \subset T_1 \cap T_2$ or $\beta_E := \beta_T$ for $E \in \mathscr{E}_\Gamma$ with

$E \in \mathscr{E}_T$. Furthermore, we introduce the quantities $\mu_T := \min\left\{\beta_T^{-1/2}, h_T\alpha^{-1/2}\right\}$ and $\mu_E := \min\left\{\beta_E^{-1/2}, h_E\alpha^{-1/2}\right\}$. We provide an a posteriori estimator of residual type for both coupling schemes, which is based on the primal mesh $\mathscr{T}$. For the CFVM-BEM the estimator post processes the original piecewise finite volume approximation in the interior domain to a conforming finite element space which leads to the so called Morley interpolant $\mathscr{I}u_h$; see [2]. Let us write $u_{FVM} = u_h$ for FVM-BEM and $u_{FVM} = \mathscr{I}u_h$ for CFVM-BEM. Then the residual reads $R := f - \operatorname{div}(-\alpha\nabla u_{FVM} + \mathbf{b}u_{FVM}) - \gamma\, u_{FVM}$ and an edge-residual $J : L^2(\mathscr{E}) \to \mathbb{R}$ is given by

$$J|_E := \begin{cases} (-\alpha\nabla u_{FVM}|_{T'} + \alpha\nabla u_{FVM}|_T)\cdot\mathbf{n} & \text{for all } E \in \mathscr{E}_I, \\ (-\alpha\nabla u_{FVM} + \mathbf{b}u_{FVM})\cdot\mathbf{n} + \phi_h + t_0 & \text{for all } E \in \mathscr{E}_\Gamma^{in}, \\ -(\alpha\nabla u_{FVM})\cdot\mathbf{n} + \phi_h + t_0 & \text{for all } E \in \mathscr{E}_\Gamma^{out}, \end{cases}$$

with $E = T \cap T'$, $T, T' \in \mathscr{T}$ for $E \in \mathscr{E}_I$ and $E \in \mathscr{E}_T$ otherwise, and $\mathbf{n}$ pointing outward of T. First we define the refinement indicator for each element $T \in \mathscr{T}$ by

$$\begin{aligned} \eta_T^2 := \mu_T^2\|R\|_{L^2(T)}^2 &+ \frac{1}{2}\sum_{E\in\mathscr{E}_I\cap\mathscr{E}_T} \alpha^{-1/2}\mu_E\|J\|_{L^2(E)}^2 + \sum_{E\in\mathscr{E}_\Gamma\cap\mathscr{E}_T} \alpha^{-1/2}\mu_E\|J\|_{L^2(E)}^2 \\ &+ \sum_{E\in\mathscr{E}_\Gamma\cap\mathscr{E}_T} h_E\|\partial u_{h,\Gamma}/\partial s - \partial/\partial s\,(u_0 - \mathscr{V}\phi_h + (1/2 + \mathscr{K})\xi_h)\,\|_{L^2(E)}^2 \\ &+ \sum_{E\in\mathscr{E}_\Gamma\cap\mathscr{E}_T} h_E\|\mathscr{W}\xi_h + (1/2 + \mathscr{K}^*)\phi_h\|_{L^2(E)}^2, \end{aligned} \tag{4}$$

where $\partial/\partial s$ denotes the arc length derivative. Note that $u_{h,\Gamma}$ will be replaced by u_h for FVM-BEM. For the upwind FVM-BEM, we additionally define

$$\eta_{T,up}^2 := \alpha_T^{-1/2}\mu_T \sum_{\tau_{ij}^T\in\mathscr{D}^T} \|\mathbf{b}\cdot\mathbf{n}_i(u_h - u_{h,ij}^T)\|_{L^2(\tau_{ij}^T)}^2 \tag{5}$$

for $T \in \mathscr{T}$, where $\mathscr{D}^T := \left\{\tau_{ij}^T \,\middle|\, \tau_{ij}^T = V_i \cap V_j \cap T \neq \emptyset \text{ for } V_i, V_j \in \mathscr{T}^* \text{with} V_i \neq V_j\right\}$ and $u_{h,ij}^T$ is the upwind value (see Sect. 2). The upper bound is proven in [2, 3] (for FVM-BEM even for the 3-D case) and reads

$$\begin{aligned} C_{\text{rel}}^{-2}\big(|\!|\!|u - u_{FVM}|\!|\!|_\Omega + \|\xi - \xi_h\|_{H^{1/2}(\Gamma)} + \|\phi - \phi_h\|_{H^{-1/2}(\Gamma)}\big)^2 \\ \leq \eta^2 := \sum_{T\in\mathscr{T}} \big(\eta_T^2 (+\eta_{T,up}^2)\big). \end{aligned} \tag{6}$$

The constant C_{rel} depends only on the shape of the elements $\mathscr{T}$ but not on the size, the number of elements or the model data, and

$$|||v|||_\Omega^2 := \|\alpha^{1/2}\nabla v\|_{L^2(\Omega)}^2 + \left\|\left((\operatorname{div}\mathbf{b})/2+\gamma\right)^{1/2} v\right\|_{L^2(\Omega)}^2 \quad \text{for all } v \in H^1(\Omega)$$

defines the energy (semi)norm. In [2] robustness is also shown against a piecewise constant α. Furthermore, one can also find a proof for a local lower bound of (6) in [2, 3], where the constant additionally depends on the local Péclet number.

5 Numerical Experiments

With the refinement indicators of (4) (plus (5) for FVM-BEM upwinding), we run a standard refinement algorithm with the following criterion: construct a minimal subset $\mathscr{M}^{(k)}$ of $\mathscr{T}^{(k)}$ at step k such that

$$\theta \sum_{T\in\mathscr{T}^{(k)}} \left(\eta_T^2\,(+\eta_{T,up}^2)\right) \le \sum_{T\in\mathscr{M}^{(k)}} \left(\eta_T^2\,(+\eta_{T,up}^2)\right)$$

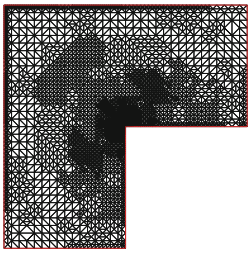
and mark all elements in $\mathscr{M}^{(k)}$ for refinement. We use $\theta = 1/2$ for adaptive mesh-refinement. The shape regularity constant is bounded in all our examples which can be guaranteed by a red-green-blue refinement strategy.

5.1 The Classical L-Shaped Laplace Problem

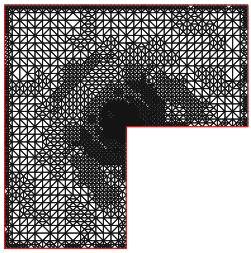
The Laplace problem with $\Omega = (-1/4, 1/4)^2 \backslash \big([0, 1/4] \times [-1/4, 0]\big)$ can be seen as a benchmark problem to test discrete systems, especially with adaptive mesh refinement techniques. The given exact solutions read: $u(x_1, x_2) = r^{2/3}\sin(2\varphi/3)$ with $(x_1, x_2) = r(\cos\varphi, \sin\varphi)$ ($r \in \mathbb{R}_0^+, \varphi \in [0, 2\pi[$) for (1a) and for (1b) and (1c) $u_e(x_1, x_2) = \log\sqrt{(x_1 + 0.125)^2 + (x_2 - 0.125)^2}$ with $a_\infty = 0$ and $b_\infty = 1$. The right-hand side is $f = 0$ if we choose $\alpha = 1$, $\mathbf{b} = (0, 0)^T$ and $\gamma = 0$ for our model problem. The jumps u_0 and t_0 are calculated appropriately. We stress that u has a generic singularity at the reentrant corner $(0, 0)$. It is well known that a first order scheme leads to a suboptimal $\mathscr{O}(N^{-1/3})$ order of convergence with respect to the number of elements $N := \#\mathscr{T}$, or $\mathscr{O}(h^{2/3})$ if h denotes the uniform mesh-size. An adaptive refinement algorithm may give us back the optimal order of $\mathscr{O}(N^{-1/2})$. Table 1 shows the energy norm errors starting with a uniform mesh $\#\mathscr{T}^{(0)} = 12$. Note that the (not computable) BEM norms are estimated up to a constant because of $\|\xi - \xi_h\|_{H^{1/2}(\Gamma)}^2 \sim |||\xi - \xi_h|||_{\mathscr{W}}^2 := \langle \mathscr{W}(\xi - \xi_h), \xi - \xi_h\rangle_\Gamma$ and $\|\phi - \phi_h\|_{H^{-1/2}(\Gamma)}^2 \sim |||\phi - \phi_h|||_{\mathscr{V}}^2 := \langle \mathscr{V}(\phi - \phi_h), \phi - \phi_h\rangle_\Gamma$. We stress that both coupling schemes recover the optimal convergence rate $\mathscr{O}(N^{-1/2})$ in the sum of the energy norms. However, the CFVM-BEM coupling has a stronger pre-refinement phase in $|||u - u_{FVM}|||_\Omega$, whereas all other norms are similar with respect to N.

Table 1 Energy norms for different refinement levels k for both coupling systems for example 5.1

k	Scheme	N	$\|u - u_{FVM}\|_\Omega$	$\|\xi - \xi_h\|_\mathscr{W}$	$\|\phi - \phi_h\|_\mathscr{V}$	$\|u - u_{FVM}\|_{L^2(\Omega)}$
8	FVM-BEM	1106	$1.70e-02$	$4.85e-03$	$4.72e-03$	$9.90e-05$
	CFVM-BEM	256	$2.96e-02$	$2.83e-02$	$2.41e-02$	$9.67e-04$
12	FVM-BEM	11592	$5.02e-03$	$6.12e-04$	$6.19e-04$	$1.02e-05$
	CFVM-BEM	1148	$8.25e-03$	$4.63e-03$	$4.20e-03$	$1.22e-04$
16	FVM-BEM	121544	$1.54e-03$	$1.00e-04$	$9.59e-05$	$1.34e-06$
	CFVM-BEM	9983	$2.39e-03$	$1.03e-03$	$8.13e-04$	$1.28e-05$
20	CFVM-BEM	94008	$7.44e-04$	$1.74e-04$	$1.69e-04$	$1.61e-06$



$\#\mathscr{T}^{(11)} = 6808.$



$\#\mathscr{T}^{(15)} = 5633.$

Fig. 1 Adaptively generated mesh for FVM-BEM (*left*) and CFVM-BEM (*right*) for example 5.1

Figure 1 shows adaptively refined meshes where we have chosen a mesh with almost the same number of elements. Both look similar. As expected the refinement happens around the singularity and a little bit on the coupling boundary.

5.2 A More Practical Example

Let us choose the same Ω as above and $\alpha = 0.1$, $\mathbf{b} = (15, 10)^T$ and $\gamma = 10^{-2}$. The volume force f is in the lower square, i.e., $f = 5$ for $-0.2 \le x_1 \le -0.1$, $-0.2 \le x_2 \le -0.05$ and $f = 0$ elsewhere. This example describes the stationary concentration of a chemical dissolved and distributed in a fluid, where we have a convection dominated problem in Ω and a diffusion distribution in Ω_e. This is a prototype of a transport problem but here without boundary conditions (which are "replaced" by the exterior problem). We prescribe the jumps $u_0 = 0$ and $t_0 = 0$ and fix the radiation condition $b_\infty = 0$ and get additionally the constraint $\int_\Gamma \phi_h \, ds = 2\pi b_\infty$. Note that we have an additional term $\langle a_\infty, \psi_h \rangle_\Gamma$ on the left-hand side of (2b) and (3c) with the unknown a_∞ and an additional equation as the counterpart. In Fig. 2 we see that the refinement for both schemes happens from f along the convection $\mathbf{b}$ and the layers at the boundary. However, the CFVM-BEM refinement is *more* local. The contour lines are generated at the same level and show the flow also into the unbounded domain and look very

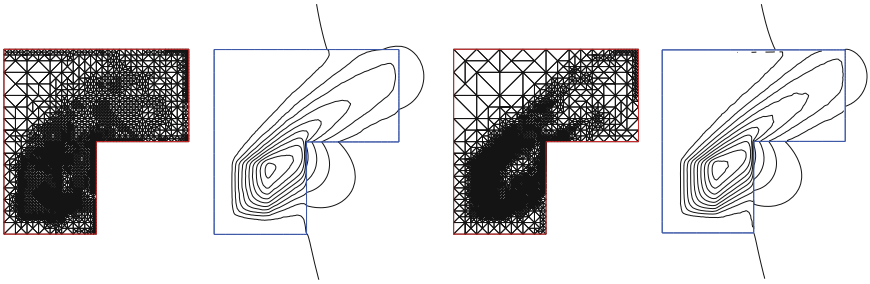


Fig. 2 Adaptively generated mesh and contour lines for FVM-BEM (*left*) and CFVM-BEM (*right*) for example 5.2

similar. The values in Ω_e can be calculated by the *representation formula* from the Cauchy data ξ_h and ϕ_h; see [1, 2].

6 Conclusions

We have illustrated on practical experiments the effectiveness of both conservative adaptive coupling methods. Contrary to FEM-BEM couplings FVM-BEM and CFVM-BEM do not have a *global* Galerkin orthogonality which leads to some difficulties in their analysis. CFVM-BEM uses the primal mesh (local conservation of the fluxes) for the (non conforming) interior piecewise constant numerical solution, which could be an advantages for using meshes with hanging nodes. On the other hand CFVM-BEM has an additional block compared to FVM-BEM and one should do more tests to show the robustness of this additional interpolation. With an interior piecewise affine and globally continuous solution FEM-BEM is closer to the spirit of FEM-BEM but with the robustness of a finite volume scheme in the interior domain and mass conservation (but local fluxes on the dual mesh). The a posteriori estimation for CFVM-BEM is more complicated because it relies on a post processed Morley-type interpolant. Both a posteriori estimates are of residual type and robust and semi-robust in the upper and lower bound, respectively. More rigorous testing has to be done to recommend one over the other for a particular problem.

References

1. Erath, C.: Coupling of the finite volume element method and the boundary element method: an a priori convergence result. SIAM J. Numer. Anal. **50**(2), 574–594 (2012)
2. Erath, C.: A new conservative numerical scheme for flow problems on unstructured grids and unbounded domains. J. Comput. Phys. **245**, 476–492 (2013)

3. Erath, C.: A posteriori error estimates and adaptive mesh refinement for the coupling of the finite volume method and the boundary element method. SIAM J. Numer. Anal. **51**(3), 1777–1804 (2013)
4. Erath, C., Praetorius, D.: A posteriori error estimate and adaptive mesh refinement for the cell-centered finite volume method for elliptic boundary value problems. SIAM J. Numer. Anal. **47**(1), 109–135 (2008)

Gradient Schemes for Stokes Problem

Robert Eymard and Pierre Feron

Abstract We provide a framework which encompasses a large family of conforming and nonconforming numerical schemes, for the approximation of the steady state incompressible Stokes equations with homogeneous Dirichlet's boundary conditions. Three examples (Taylor-Hood, extended MAC and Crouzeix-Raviart schemes) are shown to enter into this framework. The convergence of the scheme is proved by compactness arguments, thanks to estimates on the discrete solution that allow to prove the weak convergence to the unique continuous solution of the problem. Then strong convergence results are obtained thanks to the limit problem. An error estimate result is provided, applying on solutions with low regularity.

1 Incompressible Steady Stokes Problem

We consider the incompressible steady Stokes problem:

$$\left\{\begin{array}{rl} \eta\overline{u} - \Delta\overline{u} + \nabla\overline{p} = f & \text{in } \Omega \\ \text{div}\overline{u} = 0 & \text{in } \Omega \\ \overline{u} = 0 & \text{on } \partial\Omega \end{array}\right. \tag{1}$$

where $\overline{u}$ represents the velocity field and $\overline{p}$ the pressure, under the following hypotheses (called Hypotheses H in the following): Ω an open bounded Lipschitz domain of $\mathbb{R}^d$ with $d = 2$ or 3, $f \in L^2(\Omega)^d$ and $\eta \in [0, +\infty)$.

R. Eymard (✉) · P. Feron
LAMA (UMR 8050) UPEM, UPEC, CNRS, 77454, Marne-la-Valle, France
e-mail: Robert.Eymard@u-pem.fr

P. Feron
e-mail: Pierre.Feron@u-pem.fr

J. Fuhrmann et al. (eds.), *Finite Volumes for Complex Applications VII - Methods and Theoretical Aspects*, Springer Proceedings in Mathematics & Statistics 77,
DOI: 10.1007/978-3-319-05684-5_25,

Definition 1 Under Hypotheses (H), denoting, for $(\xi^{(i)}, \chi^{(i)})_{i=1,\dots,d}$ with $\xi^{(i)}, \chi^{(i)} \in \mathbb{R}^d$, by $\xi : \chi = \sum_{i=1}^d \xi^{(i)} \cdot \chi^{(i)}$, $(\overline{u}, \overline{p})$ is called a weak solution to (1) if

$$\begin{cases} \overline{u} \in H_0^1(\Omega)^d, \ \overline{p} \in L_0^2(\Omega) \text{ where } L_0^2(\Omega) = \{q \in L^2(\Omega), \displaystyle\int_\Omega q \mathrm{d}x = 0\}, \\ \eta \displaystyle\int_\Omega \overline{u} \cdot \overline{v} \mathrm{d}x + \int_\Omega \nabla \overline{u} : \nabla \overline{v} \mathrm{d}x - \int_\Omega \overline{p} \mathrm{div} \overline{v} \mathrm{d}x = \int_\Omega f \cdot \overline{v} \mathrm{d}x, \forall \overline{v} \in H_0^1(\Omega)^d, \\ \mathrm{div}(\overline{u}) = 0 \text{ a.e. in } \Omega. \end{cases} \tag{2}$$

Theorem 1 (Existence and uniqueness [10]) *Under Hypotheses* (H)*, there exists one and only one weak solution* $(\overline{u}, \overline{p})$ *to Problem (1) in the sense of Definition 1.*

The aim of this paper is to provide a theoretical framework which includes several useful schemes for the Stokes (and the Navier-Stokes) problems. This framework is given in Sect. 2, as an extension of the notion of gradient schemes provided for scalar elliptic problems [3–7]. Among the schemes which are included in this framework, we briefly present in Sect. 3 the Taylor-Hood scheme, an extended version [1] of the Marker-And-Cell (MAC) scheme [8, 9, 11] and the Crouzeix-Raviart scheme [2] (these three schemes are useful in many industrial applications). In Sect. 4, we provide the convergence result for this general framework, followed by an error estimate result, also providing a proof for the convergence of the general scheme, but needing slightly more regularity than the convergence result.

2 Gradient Scheme

Definition 2 *(Gradient Discretisation for the Stokes problem)* A gradient discretisation D for the Stokes problem with homogeneous Dirichlet's boundary conditions, is defined by $D = (X_{D,0}, \Pi_D, \nabla_D, Y_D, \chi_D, \mathrm{div}_D)$, with:

1. $X_{D,0}$ is a vector space on $\mathbb{R}$ with finite dimension.
2. Y_D is a vector space on $\mathbb{R}$ with finite dimension.
3. The linear mapping $\Pi_D : X_{D,0} \to L^2(\Omega)^d$ is the reconstruction of the approximate velocity field.
4. The linear mapping $\chi_D : Y_D \to L^2(\Omega)$ is the reconstruction of the approximate pressure, and must be chosen such that $\|\chi_D \cdot\|_{L^2(\Omega)}$ is a norm on Y_D. We then denote $Y_{D,0} = \{q \in Y_D, \int_\Omega \chi_D q \mathrm{d}x = 0\}$.
5. The linear mapping $\nabla_D : X_{D,0} \to L^2(\Omega)^{d \times d}$ is the discrete gradient operator. It must be chosen such that $\|\cdot\|_D := \|\nabla_D \cdot\|_{L^2(\Omega)^{d \times d}}$ is a norm on $X_{D,0}$.
6. The linear mapping $\mathrm{div}_D : X_{D,0} \to L^2(\Omega)$ is the discrete divergence operator.

Remark 1 (Boundary conditions) The definition of $\|\cdot\|_D$ depends on the considered boundary conditions. Here for simplicity we only consider homogeneous Dirichlet's boundary conditions (leading to the notation $X_{D,0}$) but other can easily be addressed.

Definition 3 *(Coercivity)* Let D be a discretisation in the sense of definition 2. Let C_D and β_D be defined by

$$C_D = \max_{v\in X_{D,0},\|v\|_D=1} \|\Pi_D v\|_{L^2(\Omega)} + \max_{v\in X_{D,0},\|v\|_D=1} \|\mathrm{div}_D v\|_{L^2(\Omega)},$$
$$\beta_D = \min\{\max_{v\in X_{D,0},\|v\|_D=1} \int_\Omega \chi_D q \,\mathrm{div}_D v \,\mathrm{d}x,\ q \in Y_{D,0} \text{ such that } \|\chi_D q\|_{L^2(\Omega)} = 1\}.$$

A sequence $(D_m)_{m\in\mathbb{N}}$ of gradient discretisation is said to be **coercive** if there exist $C_P \in \mathbb{R}_+$ such that $C_{D_m} \le C_P$ (discrete Poincaré inequality and control of discrete divergence) and if there exists $\beta \in (0, +\infty)$ such that $\beta_{D_m} \ge \beta$ (discrete LBB condition), for all $m \in \mathbb{N}$.

Definition 4 *(Consistency)* Let D be a gradient discretisation in the sense of definition 2, and let $I_D : H_0^1(\Omega)^d \mapsto X_{D,0}$, $S_D : H_0^1(\Omega)^d \to [0, +\infty)$, $\widetilde{I}_D : L_0^2(\Omega) \mapsto Y_{D,0}$ and $\widetilde{S}_D : L_0^2(\Omega) \to [0, +\infty)$ be defined by

$$\forall \varphi \in H_0^1(\Omega)^d,\ \forall v \in X_{D,0},$$
$$E_D(v,\varphi) = \|\Pi_D v - \varphi\|_{L^2(\Omega)^d} + \|\nabla_D v - \nabla\varphi\|_{L^2(\Omega)^{d\times d}} + \|\mathrm{div}_D v - \mathrm{div}\varphi\|_{L^2(\Omega)},$$
$$\forall \varphi \in H_0^1(\Omega)^d,\ I_D(\varphi) = \operatorname*{argmin}_{v\in X_{D,0}} E_D(v,\varphi),\ S_D(\varphi) = E_D(I_D(\varphi),\varphi),$$

and

$$\forall \psi \in L_0^2(\Omega),\quad \widetilde{I}_D(\psi) = \operatorname*{argmin}_{z\in Y_{D,0}} \|\chi_D z - \psi\|_{L^2(\Omega)},\ \widetilde{S}_D(\psi) = \|\chi_D \widetilde{I}_D(\psi) - \psi\|_{L^2(\Omega)}.$$

A sequence $(D_m)_{m\in\mathbb{N}}$ of gradient discretisation is said to be **consistent** if, for all $\varphi \in H_0^1(\Omega)^d$, $S_{D_m}(\varphi)$ tends to 0 when $m \to \infty$ and for all $\psi \in L_0^2(\Omega)$, $\widetilde{S}_{D_m}(\psi)$ tends to 0 as $m \to \infty$.

Definition 5 *(Limit-conformity)* Let D be a gradient discretisation in the sense of definition 2, and let $W_D : H_{\mathrm{div}}(\Omega)^d \to [0, +\infty)$ and $\widetilde{W}_D : H^1(\Omega) \to [0, +\infty)$ be respectively defined by

$$\forall \varphi \in H_{\mathrm{div}}(\Omega)^d,\ W_D(\varphi) = \max_{v\in X_{D,0},\|v\|_D=1} \left(\int_\Omega (\nabla_D v : \varphi + \Pi_D v \cdot \mathrm{div}\varphi)\,\mathrm{d}x \right),$$
$$\forall \psi \in H^1(\Omega),\ \widetilde{W}_D(\psi) = \max_{v\in X_{D,0},\|v\|_D=1} \left(\int_\Omega (\Pi_D v \cdot \nabla\psi + \psi\,\mathrm{div}_D v)\mathrm{d}x \right).$$

A sequence $(D_m)_{m\in\mathbb{N}}$ of gradient discretisation is said to be **limit-conforming** if, for all $\varphi \in H_{\mathrm{div}}(\Omega)^d$, $W_{D_m}(\varphi)$ tends to 0 when $m \to \infty$ and if, for all $\psi \in H^1(\Omega)$, $\widetilde{W}_{D_m}(\psi)$ tends to 0 as $m \to \infty$.

Under Hypotheses (H), let D be a gradient discretisation of Ω in the sense of definition 2. The gradient scheme for the approximation of Problem (1) is given by

$$\begin{cases} (u,\, p) \in X_{D,0} \times Y_{D,0},\ \forall v \in X_{D,0}, \\ \eta \displaystyle\int_{\Omega} \Pi_D u \cdot \Pi_D v \mathrm{d}x + \int_{\Omega} \nabla_D u : \nabla_D v \mathrm{d}x - \int_{\Omega} \chi_D p \, \mathrm{div}_D v \mathrm{d}x = \int_{\Omega} f \cdot \Pi_D v \mathrm{d}x, \\ \displaystyle\int_{\Omega} \mathrm{div}_D u \chi_D q \, \mathrm{d}x = 0, \quad \forall q \in Y_{D,0}. \end{cases} \tag{3}$$

3 Examples of Gradient Schemes for the Stokes Problem

Conforming Taylor-Hood scheme

For this example, $X_{D,0}$ (resp. Y_D) is the vector space of the degrees of freedom for the velocity (resp. the pressure) in the Taylor-Hood element, Π_D and χ_D are obtained through the finite element basis functions, and we define the conforming operators $\nabla_D = \nabla \circ \Pi_D$ and $\mathrm{div}_D = \mathrm{div} \circ \Pi_D$ (this implies that W_D and $\widetilde{W}_D$ are identically null).

The extended MAC scheme for non conforming meshes

This example is detailed in [1]. We consider 2D or 3D meshes of Ω, which are such that all internal faces have their normal vector parallel to one of the basis vector $e^{(k)}$ of the space $\mathbb{R}^d$, for some $k = 1, \ldots, d$ (see an example at left part of Fig. 1). Note that on the other hand, the external faces need not be aligned with the axes: they are only assumed to be planar. Hence curved boundaries may be meshed with such grids, by using local refinement close to the boundaries, such as in Fig. 1. These meshes are used for the approximation of the pressure and of the divergence operator. Then the gradient scheme is defined as follows.

1. $X_{D,0}$ is the vector space on $\mathbb{R}$ of all families of normal velocities to all internal edges of the mesh (see the left part of Fig. 1).
2. Y_D is the vector space on $\mathbb{R}$ of all families of values in the cells of the mesh.
3. The linear mapping $\Pi_D : X_{D,0} \to L^2(\Omega)^d$ is the reconstruction of the approximate velocity field, defined as the piecewise constant value of each component of the velocity in the corresponding Voronoï cells (the right part of Fig. 1 presents the grid for the horizontal velocity).
4. The linear mapping $\chi_D : Y_D \to L^2(\Omega)$ is the piecewise constant reconstruction of the approximate pressure in the pressure mesh.
5. The linear mapping $\nabla_D : X_{D,0} \to L^2(\Omega)^{d \times d}$ is the discrete gradient operator, obtained as the gradient of the P^1 reconstruction of each component of the velocity in the corresponding triangular grid (the medium part of Fig. 1 shows this

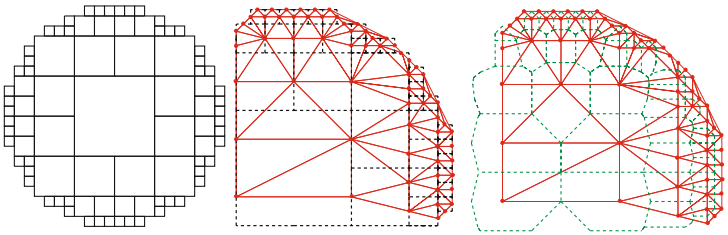


Fig. 1 *Left* pressure grid. *Middle* Zoom on the top right triangular velocity grid, used for the gradient reconstruction of the horizontal velocity (pressure grid recalled by discontinuous lines). *Right* Voronoï cells used for the velocity reconstruction

triangular grid for the horizontal velocity, joining the barycenters of the vertical edges of the pressure grid).

6. The linear mapping $\mathrm{div}_D : X_{D,0} \to L^2(\Omega)$ is the discrete divergence operator, simply computed as the piecewise constant value in the cells of the pressure grid obtained through the balance of the normal velocities integrated over the faces of the mesh.

The Crouzeix-Raviart Scheme

We consider 2D or 3D simplicial meshes of Ω (triangles in 2D, tetrahedra in 3D). Then the Crouzeix-Raviart scheme [2] can be defined as a gradient scheme by the following way:

1. $X_{D,0}$ is the vector space on $\mathbb{R}$ of all families of vectors of $\mathbb{R}^d$ at the center of all internal faces of the mesh.
2. Y_D is the vector space on $\mathbb{R}$ of all families of values in the simplices.
3. The linear mapping $\Pi_D : X_{D,0} \to L^2(\Omega)^d$ is the nonconforming piecewise affine reconstruction of each component of the velocity.
4. The linear mapping $\chi_D : Y_D \to L^2(\Omega)$ is the piecewise constant reconstruction in the simplices.
5. The linear mapping $\nabla_D : X_{D,0} \to L^2(\Omega)^{d\times d}$ is the so-called "broken gradient" of the velocity, defined as the piecewise constant field of the gradient of the affine components of the velocity in all the simplices.
6. The linear mapping $\mathrm{div}_D : X_{D,0} \to L^2(\Omega)$ is the discrete divergence operator, simply computed as the piecewise constant value in the cells of the pressure grid obtained through the balance of the normal velocities integrated over the faces of the mesh.

4 Convergence Results

Lemma 1 (Estimates) *Under Hypotheses* (H)*, Let* D *a gradient discretisation of* Ω *in the sense of definition 2 such that* $\beta_D > 0$ *(see Definition 3). Let* (u, p) *be a solution of (3). Then, there exists* $C_1 \geq 0$*, only depending on* Ω*, et* d*,* η*, and any* $C \geq C_D + \frac{1}{\beta_D}$ *such that:*

$$\|u\|_D \leq C_1 \|f\|_{L^2(\Omega)^d} \text{ and } \|\chi_D p\|_{L^2(\Omega)} \leq C_1 \|f\|_{L^2(\Omega)^d}. \tag{4}$$

As an immediate consequence, there exists one and only one (u, p)*, solution to (3).*

Proof One first set $v = u$ in (3). This immediately provides the left part of (4), thanks to the Cauchy-Schwarz inequality and to Definition 3. Then one selects some $v \in X_{D,0}$ such that $\|v\|_D = 1$ and $\beta_D \|\chi_D p\|_{L^2(\Omega)} \leq \int_\Omega \chi_D p \, \mathrm{div}_D v \, \mathrm{d}x$. Choosing this v in (3) leads to the right part of (4). The existence and uniqueness of (u, p) results from the fact that it is the solution of a square linear system, with kernel reduced to $(0, 0)$.

Theorem 2 (Convergence of the scheme) *Under Hypotheses* (H)*, Let* $(\overline{u}, \overline{p})$ *the unique weak solution of the incompressible steady Stokes problem (1) in the sense of definition 1 and let* $(D^{(m)})_{m\in\mathbb{N}}$ *be a sequence of gradient discretisation on* Ω *in the sense of definition 2, which is consistent, limit-conforming and coercive in the sense of the above definitions. Let* (u_m, p_m) *be the unique solution of the scheme (3) for* $D = D_m$*. Then, as* $m \to \infty$*,*

- $\Pi_{D^{(m)}} u_m$ *converges to* $\overline{u}$ *in* $L^2(\Omega)^d$*,*
- $\nabla_{D^{(m)}} u_m$ *converges to* $\nabla \overline{u}$ *in* $L^2(\Omega)^{d\times d}$*,*
- $\chi_{D^{(m)}} p_m$ *converges to* $\overline{p}$ *in* $L^2(\Omega)$*.*

Proof In the following proof, we use simplified notations for the integrals for shortness reasons, and we replace all indices $D^{(m)}$ by m, hence denoting by $D^{(m)} = (X_m, \Pi_m, \nabla_m, Y_m, \chi_m, \mathrm{div}_m)$, and the values provided by Definition 3 are denoted by C_m and $\beta_m \geq \beta > 0$. We first observe that, thanks to Lemma 1 and to the limit-conformity property, up to a subsequence, there exists $\overline{u} \in H^1_0(\Omega)^d$ and $\overline{p} \in L^2_0(\Omega)$ such that weak convergence properties hold for the discrete reconstructions of the approximate velocity, its approximate gradient, its approximate divergence, and the approximate pressure. Then, for any test function $\overline{v} \in H^1_0(\Omega)^d$, we set in (3) $v = I_m \overline{v}$ and $q = \widetilde{I}_m(\mathrm{div}\overline{u})$ (we have $\mathrm{div}\overline{u} \in L^2_0(\Omega)$ since $\int_\Omega \mathrm{div}\overline{u} = \int_{\partial\Omega} \overline{u} \cdot n = 0$, see Definition 5). Then we pass to the limit $m \to \infty$ in the scheme. We get, by weak/strong convergence, that $\int_\Omega (\mathrm{div}\overline{u})^2 = 0$ and that (2) holds. This proves that $(\overline{u}, \overline{p})$ is the unique weak solution of the incompressible steady Stokes problem (1). The uniqueness of the limit shows that the whole sequence converges.

Passing to the limit in the scheme with $v = u_m$ shows the convergence of the norm of the discrete velocity gradient $\nabla_m u_m$ to its continuous counterpart $\nabla\overline{u}$. This shows the strong convergence of the gradient. The coercivity property and interpolation of the limit shows that the reconstruction of the velocity $\Pi_m u_m$ is strongly convergent.

Let us now turn to the convergence of the approximate pressure in $L^2(\Omega)$. We select $v_m \in X_m$ such that $\|v_m\|_{D^{(m)}} = 1$ and

$$\beta\|\chi_m(\widetilde{I}_m\overline{p} - p_m)\|_{L^2(\Omega)} \leq \int_\Omega \chi_m(\widetilde{I}_m\overline{p} - p_m)\mathrm{div}_m v_m.$$

Letting $v = v_m$ in the scheme, we get

$$\eta\int_\Omega \Pi_m u_m \cdot \Pi_m v_m + \int_\Omega \nabla_m u_m : \nabla_m v_m - \int_\Omega \chi_m p_m \,\mathrm{div}_m v_m = \int_\Omega f \cdot \Pi_m v_m.$$

Combining the two above relations and using the triangle inequality, we deduce

$$\begin{aligned}\beta\|\overline{p} - \chi_m p_m\|_{L^2(\Omega)} \leq \beta\|\chi_m\widetilde{I}_m\overline{p} - \overline{p}\|_{L^2(\Omega)} + \int_\Omega f\cdot\Pi_m v_m + \int_\Omega \chi_m\widetilde{I}_m\overline{p}\,\mathrm{div}_m v_m \\ - \eta\int_\Omega \Pi_m u_m\cdot\Pi_m v_m - \int_\Omega \nabla_m u_m : \nabla_m v_m.\end{aligned}$$

Up to the extraction of a subsequence, we may assume that there exists $\overline{v} \in H^1_0(\Omega)^d$ such that the following weak convergences hold in L^2: $\Pi_m v_m$ to $\overline{v}$, $\nabla_m v_m$ to $\nabla\overline{v}$ and $\mathrm{div}_m v_m$ to $\mathrm{div}\overline{v}$. Using the (already proved) strong convergence properties for the velocity, we may now pass to the limit $m \to \infty$, since all integrals involve weak/strong convergence properties. We get

$$\beta\limsup_{m\to\infty}\|\overline{p} - \chi_m p_m\|_{L^2(\Omega)} \leq \int_\Omega f\cdot\overline{v} + \int_\Omega \overline{p}\,\mathrm{div}\overline{v} - \eta\int_\Omega \overline{u}\cdot\overline{v} - \int_\Omega \nabla\overline{u} : \nabla\overline{v}.$$

It now suffices to use the fact that we already proved that $(\overline{u}, \overline{p})$ is a weak solution to the Stokes equation. We then get that the right hand side of the previous inequality vanishes, which shows the convergence in L^2 for this subsequence. Using a standard uniqueness argument, we deduce that the whole sequence converges.

The following error estimate needs more regularity hypotheses than that which have been done for the above convergence theorem.

Theorem 3 *Under Hypotheses (H), Let $(\overline{u}, \overline{p})$ be the unique solution of the incompressible steady Stokes problem (1) in the sense of definition 2 such that $\overline{p} \in H^1(\Omega)$ (which implies that $\nabla\overline{u} \in H_{\mathrm{div}}(\Omega)^d$). Let D be a gradient discretisation on Ω in the sense of definition 1 such that $\beta_D > 0$ (see Definition 3). Let $(u, p) \in V_D$ be the unique solution of the scheme (3). Then there exists C_e, which increasingly depends on only η, C_D and $\frac{1}{\beta_D}$, such that there holds*

$$\|\overline{u} - \Pi_D u\|_D + \|\overline{p} - \chi_D p\|_{L^2(\Omega)} \leq C_e\left(W_D(\nabla\overline{u}) + \widetilde{W}_D(\overline{p}) + S_D(\overline{u}) + \widetilde{S}_D(\overline{p})\right).$$

Proof Under the hypotheses of the theorem, since $-\Delta\overline{u} = -\mathrm{div}(\nabla\overline{u}) = f - \nabla\overline{p} - \eta\overline{u}$ a.e., we get that $\nabla\overline{u} \in H_{\mathrm{div}}(\Omega)^d$. Using this expression in $W_D(\nabla\overline{u})$ and using (3), we can write, for any $v \in X_{D,0}$,

$$\int_\Omega \big(\eta(\overline{u} - \Pi_D u_D) \cdot \Pi_D v + (\nabla \overline{u} - \nabla_D u_D) : \nabla_D v + (\chi_D p_D \operatorname{div}_D v + \nabla \overline{p} \cdot \Pi_D v)\big) \mathrm{d}x \leq W_D(\nabla \overline{u}) \|v\|_D.$$

Then, introducing the expression of $\widetilde{W}_D(\overline{p})$, we get

$$\int_\Omega \big(\eta(\overline{u} - \Pi_D u_D) \cdot \Pi_D v + (\nabla \overline{u} - \nabla_D u_D) : \nabla_D v + (\chi_D p_D - \overline{p}) \operatorname{div}_D v\big) \mathrm{d}x \leq (W_D(\nabla \overline{u}) + \widetilde{W}_D(\overline{p})) \|v\|_D.$$

We now use the values $I_D \overline{u}$ and $\widetilde{I}_D \overline{p}$ introduced in Definition 3, and we denote by $\varepsilon_D(\overline{u}, \overline{p}) = W_D(\nabla \overline{u}) + \widetilde{W}_D(\overline{p}) + S_D(\overline{u}) + \widetilde{S}_D(\overline{p})$. We can then write

$$\int_\Omega \big(\eta(\Pi_D I_D \overline{u} - \Pi_D u_D) \cdot \Pi_D v + (\nabla_D I_D \overline{u} - \nabla_D u_D) : \nabla_D v\big) \mathrm{d}x + \int_\Omega (\chi_D p_D - \chi_D \widetilde{I}_D \overline{p}) \operatorname{div}_D v \leq (1 + \eta) \varepsilon_D(\overline{u}, \overline{p}) \|v\|_D. \tag{5}$$

Thanks to Definition 3, let us now take $v \in X_{D,0}$ such that $\|v\|_D = 1$ and

$$\int_\Omega \chi_D (p_D - \widetilde{I}_D \overline{p}) \operatorname{div}_D v \, \mathrm{d}x \geq \beta_D \|\chi_D (p_D - \widetilde{I}_D \overline{p})\|_{L^2(\Omega)}.$$

We then get, from (5), and using Definition 3,

$$\|\chi_D (p_D - \widetilde{I}_D \overline{p})\|_{L^2(\Omega)} \leq \frac{1 + \eta}{\beta_D} \varepsilon_D(\overline{u}, \overline{p}) + \frac{1 + \eta C_D}{\beta_D} \|I_D \overline{u} - u_D\|_D. \tag{6}$$

Now setting $v = I_D \overline{u} - u_D$ in (5) and using $\int_\Omega \operatorname{div}_D u_D \, \chi_D q = 0$ for all $q \in Y_D$, we can write

$$\|I_D \overline{u} - u_D\|_D^2 + \int_\Omega \chi_D (p_D - \widetilde{I}_D \overline{p}) \operatorname{div}_D I_D \overline{u} \, \mathrm{d}x \leq (1 + \eta) \varepsilon_D(\overline{u}, \overline{p}) \|I_D \overline{u} - u_D\|_D,$$

which implies

$$\|I_D \overline{u} - u_D\|_D^2 \leq (1 + \eta) \varepsilon_D(\overline{u}, \overline{p}) \|I_D \overline{u} - u_D\|_D + S_D(\overline{u}) \|\chi_D (p_D - \widetilde{I}_D \overline{p})\|_{L^2(\Omega)}.$$

Thanks to (6) and to the inequality $ab \leq \frac{1}{2} a^2 + \frac{1}{2} b^2$, the above inequality yields the existence of C_1, increasing function of $1/\beta_D$, C_D and η, such that $\|I_D \overline{u} - u_D\|_D \leq C_1 \varepsilon_D(\overline{u}, \overline{p})$. The conclusion follows, thanks to the definitions of $I_D \overline{u}$ and $\widetilde{I}_D \overline{p}$, to Definition 3, to the triangle inequality and to the use of (6).

References

1. Chénier, E., Eymard, R., Gallouet, T., Herbin, R.: An extension of the MAC scheme to locally refined meshes : convergence analysis for the full tensor time-dependent Navier-Stokes equations. http://hal.archives-ouvertes.fr/hal-00751556
2. Crouzeix, M., Raviart, P.A.: Conforming and nonconforming finite element methods for solving the stationary Stokes equations. I. Rev. Française Automat. Informat. Recherche Opérationnelle Sér. Rouge 7(R-3), 33–75 (1973).
3. Droniou, J., Eymard, R., Gallouët, T., Guichard, C., Herbin, R.: Gradient Schemes for Elliptic and Parabolic Problems (2014) (In preparation)
4. Droniou, J., Eymard, R., Gallouët, T., Herbin, R.: Gradient schemes: a generic framework for the discretisation of linear, nonlinear and nonlocal elliptic and parabolic equations. Math. Models Methods Appl. Sci. (M3AS) **23**(13), **23**95–2432 (2013)
5. Eymard, R., Féron, P., Gallouet, T., Herbin, R., Guichard, C.: Gradient schemes for the Stefan problem. IJFV **10s** (2013). http://hal.archives-ouvertes.fr/hal-00751555
6. Eymard, R., Guichard, C., Herbin, R., Masson, R.: Gradient schemes for two-phase flow in heterogeneous porous media and Richards equation. ZAMM p. accepted for publication (2013).
7. Eymard, R., Herbin, R.: Gradient scheme approximations for diffusion problems. Finite Volumes for Complex Applications VI Problems & Perspectives pp. 439–447 (2011).
8. Harlow, F., Welch, J.: Numerical calculation of time-dependent viscous incompressible flow of fluid with a free surface. Physics of Fluids **8**, 2182–2189 (1965)
9. Patankar, S.: Numerical heat transfer and fluid flow. Series in Computational Methods in Mechanics and Thermal Sciences, vol. XIII. Washington - New York - London: Hemisphere Publishing Corporation; New York. McGraw-Hill Book Company (1980).
10. Temam, R.: Navier-Stokes equations, Studies in Mathematics and its Applications, vol. 2, third edn. North-Holland Publishing Co., Amsterdam, : Theory and numerical analysis. With an appendix by F, Thomasset (1984)
11. Wesseling, P.: Principles of Computational Fluid Dynamics. Springer (2001).

Uniform Estimate of the Relative Free Energy by the Dissipation Rate for Finite Volume Discretized Reaction-Diffusion Systems

André Fiebach and Annegret Glitzky

Abstract We prove a uniform Poincaré-like estimate of the relative free energy by the dissipation rate for implicit Euler, finite volume discretized reaction-diffusion systems. This result is proven indirectly and ensures the exponential decay of the relative free energy with a unified decay rate for admissible finite volume meshes.

MSC2010: 65M08, 46E39, 35B40, 35K57, 35R05

1 Introduction

In a heterostructured domain $\Omega \subset \mathbb{R}^N$, we consider m diffusing species X_i with initial densities U_i which undergo a finite number of reversible chemical reactions. Besides the densities u_i of the species X_i we introduce their (dimensionless) chemical potentials v_i and chemical activities a_i. According to Boltzmann statistics we have $u_i = \overline{u}_i \mathrm{e}^{v_i} = \overline{u}_i a_i$, $i = 1, \dots, m$, where the reference densities $\overline{u}_i$ express the heterogeneity of the system. For the fluxes j_i of the species X_i we make the ansatz $j_i = -d_i u_i \nabla v_i = -d_i \overline{u}_i \mathrm{e}^{v_i} \nabla v_i = -d_i \overline{u}_i \nabla a_i$, $i = 1, \dots, m$, with diffusion coefficients d_i. Let $\mathscr{R} \subset \mathbb{Z}_+^m \times \mathbb{Z}_+^m$ be a finite subset. Each pair $(\alpha, \beta) \in \mathscr{R}$ represents the vectors of stoichiometric coefficients of a reversible reaction

$$\alpha_1 \mathsf{X}_1 + \dots + \alpha_m \mathsf{X}_m \rightleftharpoons \beta_1 \mathsf{X}_1 + \dots + \beta_m \mathsf{X}_m .$$

A. Fiebach (✉)
Physikalisch-Technische Bundesanstalt, Abbestraße 2-12, 10587 Berlin, Germany
e-mail: andre.fiebach@ptb.de

A. Glitzky
Weierstrass Institute, Mohrenstraße 39, 10117 Berlin, Germany
e-mail: glitzky@wias-berlin.de

J. Fuhrmann et al. (eds.), *Finite Volumes for Complex Applications VII - Methods and Theoretical Aspects*, Springer Proceedings in Mathematics & Statistics 77,
DOI: 10.1007/978-3-319-05684-5_26,

According to the mass action law, the net rate of this pair of reactions is of the form $k_{\alpha\beta}(a^\alpha - a^\beta)$, where $k_{\alpha\beta}$ is a reaction rate coefficient and $a^\alpha := \prod_{i=1}^m a_i^{\alpha_i}$. The net production rate of species X_i resulting from all reactions taking place is

$$R_i := \sum_{(\alpha,\beta)\in\mathscr{R}} k_{\alpha\beta}(a^\alpha - a^\beta)(\beta_i - \alpha_i).$$

The problem under consideration consists of the m continuity equations

$$\left.\begin{aligned} \tfrac{\partial u_i}{\partial t} + \nabla\cdot j_i = R_i \text{ in } \mathbb{R}_+\times\Omega, \quad \nu\cdot j_i = 0 \text{ on } \mathbb{R}_+\times\partial\Omega,\\ u_i(0) = U_i \text{ in } \Omega, \quad i = 1,\dots,m. \end{aligned}\right\} \tag{P}$$

The set $\mathscr{S} := \operatorname{span}\{\alpha - \beta : (\alpha,\beta)\in\mathscr{R}\} \subset \mathbb{R}^m$ represents the stoichiometric subspace defined by the reaction system. Our essential assumptions on the data are

(A1) Ω is an open, bounded, polyhedral domain in $\mathbb{R}^N$, $N = 2, 3$;
$\overline{u}_i,\ U_i \in L^\infty_+(\Omega),\ \overline{u}_i,\ U_i \ge \delta > 0,\ i = 1,\dots,m,\ \mathscr{R}\subset\mathbb{Z}^m_+\times\mathbb{Z}^m_+$ finite subset,
$k_{\alpha\beta},\ d_i : \Omega\times\mathbb{R}^m_+ \to \mathbb{R}_+$ Carathéodory functions satisfying
$d_i(x,a) \ge \delta,\ c \ge k_{\alpha\beta}(x,a) \ge b_{\alpha\beta}(x)$ f.a.a. $x\in\Omega$, and all $a\in\mathbb{R}^m_+$,
where $\|b_{\alpha\beta}\|_{L^1} > 0$ for all $(\alpha,\beta)\in\mathscr{R}$.
If $N = 3$ then $\max_{(\alpha,\beta)\in\mathscr{R}} \max\left\{\sum_{i=1}^m \alpha_i, \sum_{i=1}^m \beta_i\right\} \le 3$,
$\mathscr{A}\cap\partial\mathbb{R}^m_+ = \emptyset$, where
$\mathscr{A} := \left\{a\in\mathbb{R}^m_+ : a^\alpha = a^\beta \text{ for all } (\alpha,\beta)\in\mathscr{R},\ \int_\Omega(\bar{u}a - U)\,\mathrm{d}x \in \mathscr{S}\right\}$.

These assumptions allow us to handle a general class of reaction-diffusion systems, including heterogeneous materials, reactions occurring in subdomains and diffusion and reaction rate coefficients depending on the state variables, see [3, Remark 1].

The aim of the paper is to show for finite volume discretized versions of Problem (P) a Poincaré-like estimate of the discrete relative free energy by the discrete dissipation rate uniformly for all meshes with (A2), see Theorem 1. The essential new result is that our proof works without any restriction on the mesh size which is needed in [4, Theorem 3.2]. Using discrete functional inequalities from [1] instead of results in [5] the estimate is generalized from Voronoi meshes to admissible finite volume meshes. More general reaction rate and diffusion coefficients are treated, too. Finally, for Euler backward in time and finite volume in space discretization schemes, the discretized free energy along the discrete solutions decays exponentially to its equilibrium value with a uniform decay rate for all discretizations fulfilling (A2), see Theorem 2. This gives the discrete counterpart to the behavior in the continuous case characterized by [6, Theorem 4.3] in a more general setting.

2 Discretization Scheme and Main Result

An admissible mesh of Ω (see [2]) denoted by $\mathscr{M} = (\mathscr{P}, \mathscr{T}, \mathscr{E})$ is formed by a family of grid points $\mathscr{P}$ in $\bar{\Omega}$, a family $\mathscr{T}$ of control volumes and a family $\mathscr{E}$ of parts of hyperplanes in $\mathbb{R}^N$ (which represent the faces of the boxes). Let M be the number of grid points $x_K \in \mathscr{P}$, $M = \#\mathscr{P}$. $|K|$ denotes the measure of the box $K \in \mathscr{T}$. For $K, L \in \mathscr{T}$ with $K \neq L$ either the $(N-1)$ dimensional Lebesgue measure of $\bar{K} \cap \bar{L}$ is zero or $\bar{K} \cap \bar{L} = \bar{\sigma}$ for some $\sigma \in \mathscr{E}$. The symbol $\sigma = K|L$ denotes the surface between K and L. The set of interior surfaces is called $\mathscr{E}_{int} \subset \mathscr{E}$. Moreover, for $\sigma \in \mathscr{E}$ we denote by m_σ the $(N-1)$ dimensional Lebesgue measure of the face σ. For $\sigma = K|L \in \mathscr{E}_{int}$ let d_σ be the Euclidean distance of x_K and x_L and σ is assumed to be orthogonal to the line connecting x_K and x_L. $\mathscr{E}_K$ is the subset of $\mathscr{E}$ such that $\partial K = \bar{K} \setminus K = \cup_{\sigma \in \mathscr{E}_K} \bar{\sigma}$. Concerning the discretization we suppose

(A2) Let $\mathscr{M}$ be an admissible finite volume mesh with
$\operatorname{dist}(x_K, \sigma) \geq \theta d_\sigma \ \forall K \in \mathscr{T} \ \forall \sigma \in \mathscr{E}_K \cap \mathscr{E}_{int}$ $(\theta > 0)$.
Let $\mathscr{Z} = \{t_0, t_1, \ldots, t_n, \ldots\}$ be a partition of $\mathbb{R}_+$ with $t_0 = 0$, $t_n \in \mathbb{R}_+$,
$t_{n-1} < t_n$, $n \in \mathbb{N}$, $t_n \to +\infty$ as $n \to \infty$, $\sup_{n \in \mathbb{N}}(t_n - t_{n-1}) \leq \tau < \infty$.

$X(\mathscr{M})$ represents the set of functions from Ω to $\mathbb{R}$ which are constant on each box of the mesh. For $w_h \in X(\mathscr{M})$ the value at the box $K \in \mathscr{T}$ is called w_K. For $w_h \in X(\mathscr{M})$ the discrete H^1 seminorm $|w_h|_{1,\mathscr{M}}$ and H^1 norm $\|w_h\|_{1,\mathscr{M}}$ are defined by

$$|w_h|^2_{1,\mathscr{M}} = \sum_{\sigma = K|L \in \mathscr{E}_{int}} \frac{m_\sigma}{d_\sigma} |w_K - w_L|^2, \quad \|w_h\|^2_{1,\mathscr{M}} = |w_h|^2_{1,\mathscr{M}} + \|w_h\|^2_{L^2}. \quad (1)$$

For $K \in \mathscr{T}$ we denote by $u_{iK}(t_n)$ the constant density on K at t_n. Associated to the grid points we have chemical potentials $v_{iK}(t_n)$ and chemical activities $a_{iK}(t_n)$, $i = 1, \ldots, m$. Moreover we work with the vectors $\mathbf{u}$, $\mathbf{v}$, $\mathbf{a} \in \mathbb{R}^{Mm}$ and the vectors on a box $\mathbf{u}_K$, $\mathbf{v}_K$, $\mathbf{a}_K \in \mathbb{R}^m$. We introduce the mean values on the control volumes $K \in \mathscr{T}$,

$$\overline{u}_{iK} = \tfrac{1}{|K|} \int_K \overline{u}_i(x)\,\mathrm{d}x, \quad k_{\alpha\beta K}(\cdot) = \tfrac{1}{|K|} \int_K k_{\alpha\beta}(x, \cdot)\,\mathrm{d}x$$

and the corresponding piecewise constant functions $\bar{u}_{ih}$ and $k_{\alpha\beta h}$. The discrete version of Problem (P) is

$$\left.\begin{aligned} \frac{u_{iK}(t_n) - u_{iK}(t_{n-1})}{t_n - t_{n-1}} |K| - \sum_{\sigma = K|L \in \mathscr{E}_K} Y_i^\sigma(t_n)\big(a_{iL}(t_n) - a_{iK}(t_n)\big) \frac{m_\sigma}{d_\sigma} = R_i^K(t_n), \\ u_{iK}(t_n) = \overline{u}_{iK}\, \mathrm{e}^{v_{iK}(t_n)} = \overline{u}_{iK}\, a_{iK}(t_n), \quad i = 1, \ldots, m, \ n \geq 1, \\ u_{iK}(0) = U_{iK} := \tfrac{1}{|K|} \textstyle\int_\Omega U_i\,\mathrm{d}x, \quad i = 1, \ldots, m, \quad K \in \mathscr{T}, \end{aligned}\right\} (P_{\mathscr{M}})$$

where $Y_i^\sigma = Y_i^\sigma(\mathbf{a})$ is a mean of $d_i(x, a)\overline{u}_i(x)$ on the face σ and R_i^K are given by

$$R_i^K = R_i^K(\mathbf{a}_K) = \sum_{(\alpha,\beta)\in\mathscr{R}} (\beta_i - \alpha_i) k_{\alpha\beta K}(\mathbf{a}_K)\Big(\mathbf{a}_K^\alpha - \mathbf{a}_K^\beta\Big)|K|.$$

We introduce the operator $\widehat{E} : \mathbb{R}^{Mm} \to \mathbb{R}^{Mm}$, $\widehat{E}\mathbf{v} = \big((\overline{u}_{iK}\mathrm{e}^{v_{iK}})_{K\in\mathscr{T}}\big)_{i=1,\dots,m}$ and

$$\widehat{\mathscr{U}} = \Big\{\mathbf{u} \in \mathbb{R}^{Mm} : \Big(\sum_{K\in\mathscr{T}} u_{1K}|K|, \dots, \sum_{K\in\mathscr{T}} u_{mK}|K|\Big) \in \mathscr{S}\Big\}.$$

The discrete dissipation rate $\widehat{D} : \mathbb{R}^{Mm} \to \mathbb{R}$ corresponding to Problem ($\mathrm{P}_{\mathscr{M}}$) and the discrete free energy $\widehat{F} : \mathbb{R}^{Mm} \to \overline{\mathbb{R}}$ take the form

$$\begin{aligned}\widehat{D}(\mathbf{v}) = &\sum_{i=1}^m \sum_{\sigma=K|L\in\mathscr{E}_{int}} Y_i^\sigma\,(\mathrm{e}^{v_{iK}} - \mathrm{e}^{v_{iL}})(v_{iK} - v_{iL})\frac{m_\sigma}{d_\sigma}\\ &+ \sum_{(\alpha,\beta)\in\mathscr{R}} \sum_{K\in\mathscr{T}} k_{\alpha\beta K}\Big(\mathrm{e}^{\alpha\cdot\mathbf{v}_K} - \mathrm{e}^{\beta\cdot\mathbf{v}_K}\Big)(\alpha-\beta)\cdot\mathbf{v}_K|K|,\\ \widehat{F}(\mathbf{u}) = &\sum_{i=1}^m \sum_{K\in\mathscr{T}} \Big(u_{iK}\ln\frac{u_{iK}}{\overline{u}_{iK}} - u_{iK} + \overline{u}_{iK}\Big)|K|.\end{aligned}$$

Assuming (A1), Problem (P) has exactly one weak stationary solution (u^*, v^*) fulfilling $\int_\Omega (u^* - U)\,\mathrm{d}x \in \mathscr{S}$, see [6, Theorem 3.2]. It is the thermodynamic equilibrium and the corresponding constant vector of chemical activities a^* lies in $\mathscr{A}$. Also the discrete Problem ($\mathrm{P}_{\mathscr{M}}$) has a unique stationary solution $(\mathbf{u}^*, \mathbf{v}^*)$ with $\mathbf{u}^* - \mathbf{U} \in \widehat{\mathscr{U}}$ which again represents the thermodynamic equilibrium of the discrete problem ($\mathrm{P}_{\mathscr{M}}$), see [4, Theorem 3.1]. Let u_h^*, v_h^*, $a_h^* \in X(\mathscr{M})$ be the piecewise constant functions corresponding to $\mathbf{u}^*$, $\mathbf{v}^*$, $\mathbf{a}^*$. According to [4, Corollary 3.1] we have $u_{ih}^* = u_i^*\,\overline{u}_{ih}/\overline{u}_i$, $i = 1, \dots, m$, $v_h^* = v^*$, and $a_h^* = a^*$. Both results from [4] hold true for admissible meshes, too.

We now prove a Poincaré type inequality (similar to [6, Theorem 4.2] for the continuous case) which gives for the discretized situation a uniform estimate of the relative free energy $\widehat{F}(\mathbf{u}) - \widehat{F}(\mathbf{u}^*)$ by the dissipation rate $\widehat{D}$ being independent on the underlying mesh $\mathscr{M}$. [4, Theorem 3.2] contains a proof for Voronoi meshes with mesh sizes less than some constant depending on the data of the problem. Here we establish a uniform estimate for all admissible finite volume meshes fulfilling (A2).

Theorem 1 *We assume* (A1) *and* (A2). *Let* $(\mathbf{u}^*, \mathbf{v}^*)$ *be the thermodynamic equilibrium of* ($\mathrm{P}_{\mathscr{M}}$). *Then for every* $\rho > 0$ *there is a constant* $c_\rho > 0$ *such that*

$$\widehat{F}(\widehat{E}\mathbf{v}) - \widehat{F}(\mathbf{u}^*) \le c_\rho \widehat{D}(\mathbf{v}) \tag{2}$$

for all $\mathbf{v} \in \widehat{\mathscr{N}_\rho} := \Big\{\mathbf{v} \in \mathbb{R}^{Mm} : \widehat{F}(\widehat{E}\mathbf{v}) - \widehat{F}(\mathbf{u}^*) \le \rho,\ \mathbf{u} = \widehat{E}\mathbf{v} \in \mathbf{U} + \widehat{\mathscr{U}}\Big\}$, *uniformly for all admissible finite volume meshes.*

Proof In this proof we denote by c (possibly different) positive constants depending only on the data but not depending on the mesh. Let $\rho > 0$ be arbitrarily given.

1. Let $\mathbf{u} = \widehat{E}\mathbf{v} \in \mathbf{U} + \widehat{\mathscr{U}}$. By [4, Lemma 3.1] there exist constants $c_1,\ c_2 > 0$ not depending on the mesh $\mathscr{M}$ such that

$$c_1 \sum_{i=1}^{m} \|\sqrt{u_{ih}} - \sqrt{u_{ih}^*}\|_{L^2}^2 \le \widehat{F}(\mathbf{u}) - \widehat{F}(\mathbf{u}^*) \le c_2 \sum_{i=1}^{m} \|u_{ih} - u_{ih}^*\|_{L^2}^2. \tag{3}$$

Using (A1) and the inequality $(x-y)\ln\frac{x}{y} \ge |\sqrt{x} - \sqrt{y}|^2$ for $x,\ y > 0$, we estimate

$$\begin{aligned}\widehat{D}(\mathbf{v}) &\ge c \sum_{i=1}^{m} \sum_{\sigma \in \mathscr{E}_{int}} |\sqrt{\mathrm{e}^{v_{iK}}} - \sqrt{\mathrm{e}^{v_{iL}}}|^2 \frac{m_\sigma}{d_\sigma} \\ &\quad + c \sum_{(\alpha,\beta)\in\mathscr{R}} \int_\Omega b_{\alpha\beta h} \Big(\mathrm{e}^{v_h\cdot\alpha/2} - \mathrm{e}^{v_h\cdot\beta/2}\Big)^2 \mathrm{d}x =: D_1(\mathbf{v}), \quad \mathbf{v} \in \mathbb{R}^{Mm}.\end{aligned}$$

Therefore it suffices to prove the inequality

$$\widehat{F}(\mathbf{u}) - \widehat{F}(\mathbf{u}^*) \le C D_1(\mathbf{v}) \quad \forall \mathbf{v} \in \widehat{\mathscr{N}_\rho} \tag{4}$$

with some constant $C > 0$ not depending on the mesh $\mathscr{M}$.

2. If (4) would be false, then there would be a sequence of admissible meshes $\mathscr{M}_n$ and corresponding $\mathbf{v}_n \in \widehat{\mathscr{N}_\rho},\ \mathbf{u}_n = \widehat{E}\mathbf{v}_n \in \mathbf{U}_n + \widehat{\mathscr{U}},\ n \in \mathbb{N}$, such that

$$\widehat{F}(\mathbf{u}_n) - \widehat{F}(\mathbf{u}_n^*) = C_n D_1(\mathbf{v}_n) > 0, \tag{5}$$

and $\lim_{n\to\infty} C_n = +\infty$. Clearly, for each $\mathscr{M}_n$ we have to use the corresponding quantities $M, \widehat{E}, \widehat{F}, D_1$,... and sets $\mathscr{E}_{int}, \widehat{\mathscr{U}}, \widehat{\mathscr{N}_\rho}$. But we don't write them with an index $\mathscr{M}_n$. Let $a_{niK} = \mathrm{e}^{v_{niK}}$, $K \in \mathscr{T}_n$. By $u_{nih},\ v_{nih},\ a_{nih}, \ldots \in X(\mathscr{M}_n)$, $i = 1,\ldots,m$, we denote the corresponding piecewise constant functions. From (3) we obtain

$$\|\sqrt{a_{nih}} - \sqrt{a_{nih}^*}\|_{L^2}^2 \le c\|\sqrt{u_{nih}} - \sqrt{u_{nih}^*}\|_{L^2}^2 \le \frac{c}{c_1}\big(\widehat{F}(\mathbf{u}_n) - \widehat{F}(\mathbf{u}_n^*)\big) \le c(\rho). \tag{6}$$

Thus by assumption and because of $a_{nih}^* = a_i^*$ we find a suitable $\tilde{c}(\rho) < \infty$ with

$$\|\sqrt{a_{nih}}\|_{L^2} \le \tilde{c}(\rho), \quad i = 1,\ldots,m, \text{ for all } n. \tag{7}$$

3. The definition of D_1 and (4) gives $\sum_{i=1}^m |\sqrt{a_{nih}}|_{1,\mathscr{M}_n}^2 \le c D_1(\mathbf{v}_n) \to 0$. Applying the discrete Poincaré inequality for functions with general boundary values (see [1, Theorem 5]) we find for $\sqrt{a_{nih}} \in X(\mathscr{M}_n)$, $i = 1,\ldots,m$,

$$\sqrt{a_{nih}} - m_\Omega(\sqrt{a_{nih}}) \to 0 \quad \text{in} \quad L^2(\Omega), \quad \text{where} \quad m_\Omega(\sqrt{a_{nih}}) := \frac{1}{|\Omega|}\int_\Omega \sqrt{a_{nih}}\,\mathrm{d}x.$$

The discrete Sobolev-Poincaré inequality (see [1, Theorem 3]) gives for $q \in [1,\infty)$ if $N = 2$ and for $q \in [1,6]$ if $N = 3$ the estimate $\|\sqrt{a_{nih}} - m_\Omega(\sqrt{a_{nih}})\|_{L^q} \le c_q \|\sqrt{a_{nih}} - m_\Omega(\sqrt{a_{nih}})\|_{1,\mathscr{M}_n} \le \tilde{c}_q(|\sqrt{a_{nih}}|_{1,\mathscr{M}_n} + \|\sqrt{a_{nih}} - m_\Omega(\sqrt{a_{nih}})\|_{L^2}) \to 0$.

Since $m_\Omega(\sqrt{a_{nih}})\,|\Omega| = \|\sqrt{a_{nih}}\|_{L^1} \le c\|\sqrt{a_{nih}}\|_{L^2} \le c(\rho)$ by (7) for all $\mathscr{M}_n$ we find (for a subsequence, and we restrict our further investigations to this subsequence) $m_\Omega(\sqrt{a_{nih}}) \to \sqrt{\widehat{a}_i}$ in $\mathbb{R}$. Using that $|\sqrt{a_{nih}} - \sqrt{\widehat{a}_i}| \le |\sqrt{a_{nih}} - m_\Omega(\sqrt{a_{nih}})| + |m_\Omega(\sqrt{a_{nih}}) - \sqrt{\widehat{a}_i}|$ we conclude

$$\sqrt{a_{nih}} \to \sqrt{\widehat{a}_i} \quad \text{in} \quad L^q(\Omega), \quad i = 1,\dots,m, \tag{8}$$

for $q \in [1,\infty)$ if $N = 2$ and for $q \in [1,6]$ if $N = 3$. From

$$a_{nih} - \widehat{a}_i = (\sqrt{a_{nih}} - \sqrt{\widehat{a}_i})(\sqrt{a_{nih}} + \sqrt{\widehat{a}_i}) = (\sqrt{a_{nih}} - \sqrt{\widehat{a}_i})^2 + 2\sqrt{\widehat{a}_i}(\sqrt{a_{nih}} - \sqrt{\widehat{a}_i})$$

we find that

$$\|a_{nih} - \widehat{a}_i\|_{L^2} \le \|\sqrt{a_{nih}} - \sqrt{\widehat{a}_i}\|_{L^4}^2 + 2\sqrt{\widehat{a}_i}\|\sqrt{a_{nih}} - \sqrt{\widehat{a}_i}\|_{L^2} \to 0. \tag{9}$$

4. Let $r_{\alpha\beta}(a_h) := (a_h^{\alpha/2} - a_h^{\beta/2})^2$. Using $\|b_{\alpha\beta}\|_{L^1} = \|b_{\alpha\beta\,h}\|_{L^1}$, taking into account the restriction of the reaction order if $N = 3$ and (8) we have for all $(\alpha,\beta) \in \mathscr{R}$

$$\begin{aligned} 0 &\le \|b_{\alpha\beta} r_{\alpha\beta}(\widehat{a})\|_{L^1} = \|b_{\alpha\beta\,h} r_{\alpha\beta}(\widehat{a})\|_{L^1} \\ &\le \|b_{\alpha\beta\,h} r_{\alpha\beta}(a_{nh}) - b_{\alpha\beta\,h} r_{\alpha\beta}(\widehat{a})\|_{L^1} + \|b_{\alpha\beta\,h} r_{\alpha\beta}(a_{nh})\|_{L^1} \\ &\le \|b_{\alpha\beta\,h}\|_{L^\infty} \|r_{\alpha\beta}(a_{nh}) - r_{\alpha\beta}(\widehat{a})\|_{L^1} + cD_1(\mathbf{v}_n) \to 0. \end{aligned}$$

Therefore, with $\|b_{\alpha\beta}\|_{L^1} > 0$ we find necessarily that

$$\widehat{a}^\alpha = \widehat{a}^\beta \quad \forall (\alpha,\beta) \in \mathscr{R}. \tag{10}$$

5. We introduce $\widehat{u} := (\widehat{u}_1,\dots,\widehat{u}_m)$, $\widehat{u}_i := \overline{u}_i\,\widehat{a}_i$, and show $\int_\Omega (\widehat{u} - U)\,\mathrm{d}x \in \mathscr{S}$. Let $\gamma \in \mathscr{S}^\perp$ (orthogonal complement of $\mathscr{S}$ in $\mathbb{R}^m$) be arbitrarily given. Then

$$\left|\gamma \cdot \int_\Omega (\overline{u}\widehat{a} - U)\,\mathrm{d}x\right| \le \left|\gamma \cdot \int_\Omega (\widehat{a} - a_{nh})\overline{u}_{nh}\,\mathrm{d}x\right| + \left|\gamma \cdot \int_\Omega (a_{nh}\overline{u}_{nh} - U_{nh})\,\mathrm{d}x\right|.$$

By (9) the first integral on the right hand side tends to zero, the second is zero since $\mathbf{u}_n - \mathbf{U}_n \in \widehat{\mathscr{U}}$. Thus, together with (10) we find $\widehat{a} \in \mathscr{A}$, and according to (A1) we obtain that $\widehat{a} = a^*$. By the definition of $\widehat{u}$ this yields $\widehat{u} = u^*$.

6. Because of (3) and (9) we have

$$\lambda_n^2 := \widehat{F}(\mathbf{u}_n) - \widehat{F}(\mathbf{u}_n^*) \le c_2 \sum_{i=1}^m \|\overline{u}_i\|_{L^\infty} \|a_{nih} - a_{nih}^*\|_{L^2}^2 \to 0 \tag{11}$$

as $n \to \infty$. Additionally (according to (5)) we find

$$\frac{1}{C_n} = \frac{1}{\lambda_n^2} D_1(\mathbf{v}_n) \to 0 \text{ as } n \to \infty. \tag{12}$$

7. For all n we introduce

$$b_{nih} := \frac{1}{\lambda_n}\Big(\sqrt{\frac{a_{nih}}{\widehat{a}_i}} - 1\Big) \in X(\mathscr{M}_n), \quad i = 1, \dots, m.$$

Because of $(b_{niK} - b_{niL})^2 \le \frac{1}{\lambda_n^2 \widehat{a}_i}(\sqrt{a_{niK}} - \sqrt{a_{niL}})^2$ for all $\sigma = K|L \in \mathscr{E}_{int}$ it results $\sum_{i=1}^m |b_{nih}|^2_{1,\mathscr{M}_n} \le c D_1(\mathbf{v}_n)/\lambda_n^2 \to 0$. As demonstrated in Step 3 (for $\sqrt{a_{nih}}$), the discrete Poincaré and Sobolev-Poincaré inequality ensure for b_{nih} the convergence $\|b_{nih} - m_\Omega(b_{nih})\|_{L^q} \to 0$, $i = 1, \dots, m$, for $q \in [1, \infty)$ if $N = 2$ and for $q \in [1, 6]$ if $N = 3$. Using $\widehat{a}_i = a_i^* = a_{nih}^*$, (6) and (11) we obtain

$$\begin{aligned} |m_\Omega(b_{nih})|\,|\Omega| &\le \frac{1}{\lambda_n \sqrt{\widehat{a}_i}} \int_\Omega |\sqrt{a_{nih}} - \sqrt{\widehat{a}_i}|\,\mathrm{d}x \le \frac{1}{\lambda_n \sqrt{a_i^*}} \|\sqrt{a_{nih}} - \sqrt{a_{nih}^*}\|_{L^1} \\ &\le \frac{c}{\lambda_n} \|\sqrt{a_{nih}} - \sqrt{a_{nih}^*}\|_{L^2} \le \frac{c}{\lambda_n} (\widehat{F}(\mathbf{u}_n) - \widehat{F}(\mathbf{u}_n^*))^{1/2} \le \frac{c}{\lambda_n} \lambda_n = c \end{aligned}$$

for all $\mathscr{M}_n$. Thus we find (for a subsequence) $m_\Omega(b_{nih}) \to \widehat{b}_i$ in $\mathbb{R}$. By $|b_{nih} - \widehat{b}_i| \le |b_{nih} - m_\Omega(b_{nih})| + |m_\Omega(b_{nih}) - \widehat{b}_i|$ we conclude for $i = 1, \dots, m$ that

$$b_{nih} \to \widehat{b}_i \text{ in } L^q(\Omega) \text{ for } q \in [1, \infty) \text{ if } N = 2 \text{ and for } q \in [1, 6] \text{ if } N = 3. \tag{13}$$

8. We define $\widehat{y} = (\widehat{y}_1, \dots, \widehat{y}_m)$, $\widehat{y}_i := 2\widehat{b}_i u_i^* = 2\widehat{b}_i \widehat{a}_i \overline{u}_i$ and show $\int_\Omega \widehat{y}\,\mathrm{d}x \in \mathscr{S}$. Let $\gamma \in \mathscr{S}^\perp$. Since $2b_{nih}\widehat{a}_i \overline{u}_{nih} = (u_{nih} - u_{nih}^*)/\lambda_n + b_{nih}(\sqrt{\widehat{a}_i} - \sqrt{a_{nih}})\sqrt{\widehat{a}_i}\,\overline{u}_{nih}$ it results

$$\begin{aligned} \Big|\gamma \cdot \int_\Omega \widehat{y}\,\mathrm{d}x\Big| &= 2\Big|\sum_{i=1}^m \int_\Omega \widehat{b}_i \widehat{a}_i \overline{u}_{nih} \gamma_i\,\mathrm{d}x\Big| = 2\Big|\sum_{i=1}^m \int_\Omega \big(b_{nih}\widehat{a}_i \overline{u}_{nih}\gamma_i + (\widehat{b}_i - b_{nih})\widehat{a}_i \overline{u}_{nih}\gamma_i\big)\mathrm{d}x\Big| \\ &\le \Big|\gamma \cdot \int_\Omega \tfrac{u_{nh} - u_{nh}^*}{\lambda_n}\,\mathrm{d}x\Big| + c\|b_{nh}\|_{L^2}\|\sqrt{\widehat{a}_h} - \sqrt{a_{nh}}\|_{L^2} + c\|b_{nh} - \widehat{b}\|_{L^2}\|\widehat{a}\|_{L^2}, \end{aligned}$$

where the first term on the last line is zero since $\mathbf{u}_n$, $\mathbf{u}_n^* \in \widehat{\mathscr{U}} + \mathbf{U}_n$ and the last two terms tend to zero as $n \to \infty$ by (8) and (13), respectively. This leads to $\int_\Omega \widehat{y}\,\mathrm{d}x \in \mathscr{S}$.

9. By the definition of $r_{\alpha\beta}(a_{nh})$ and b_{nih} we obtain for all $(\alpha, \beta) \in \mathscr{R}$,

$$\begin{aligned} \widehat{a}^{-\alpha} r_{\alpha\beta}(a_{nh}) &= \Big(\prod_{i=1}^m (\lambda_n b_{nih} + 1)^{\alpha_i} - \prod_{i=1}^m (\lambda_n b_{nih} + 1)^{\beta_i}\Big)^2 \\ &= \Big(\lambda_n \sum_{i=1}^m b_{nih}(\alpha_i - \beta_i)\Big)^2 + Q_n, \end{aligned} \tag{14}$$

where $|Q_n| \le c\lambda_n^3(|b_{nh}|+1)^{p_0}$ with $0 \le p_0 \le 2\max_{(\alpha,\beta)\in\mathscr{R}} \max\{\sum_{i=1}^m \alpha_i, \sum_{i=1}^m \beta_i\}$. (A1) ensures $p_0 \le 6$ if $N=3$. Since $\lambda_n \to 0$ as $n\to\infty$ (see (11)), we find $\frac{1}{\lambda_n^2}\|Q_n\|_{L^1} \le c\lambda_n \int_\Omega (|b_{nh}|+1)^{p_0}\,\mathrm{d}x \to 0$ as $n\to\infty$. This together with (12) and (14) gives

$$\lim_{n\to\infty} \int_\Omega b_{\alpha\beta h}\Big(\sum_{i=1}^m b_{nih}(\alpha_i-\beta_i)\Big)^2\,\mathrm{d}x = 0 \quad \forall (\alpha,\beta)\in\mathscr{R}.$$

Therefore, from (A1) we conclude $\widehat{b} = (\widehat{b}_1,\dots,\widehat{b}_m) \in \mathscr{S}^\perp$. This together with the definition of $\widehat{y}$ and $\int_\Omega \widehat{y}\,\mathrm{d}x \in \mathscr{S}$ (see Step 8) leads to $\widehat{b}\cdot\int_\Omega \widehat{y}\,\mathrm{d}x = \sum_{i=1}^m \int_\Omega 2u_i^*\widehat{b}_i^2\,\mathrm{d}x = 0$ which ensures $\widehat{b}=0$ and $\widehat{y}=0$.

10. Using the definition of λ_n (see (11)), (3), $b_{nih}\to 0$ in $L^4(\Omega)$ and (8) we find

$$\begin{aligned}
1 &= \tfrac{1}{\lambda_n^2}\Big(\widehat{F}(\mathbf{u}_n)-\widehat{F}(\mathbf{u}_n^*)\Big) \le c\sum_{i=1}^m \|\overline{u}_{nih}\|_{L^\infty}\|\tfrac{a_{nih}-\widehat{a}_i}{\lambda_n}\|_{L^2}^2 \\
&\le c\sum_{i=1}^m \int_\Omega \tfrac{(\sqrt{a_{nih}}-\sqrt{\widehat{a}_i})^2}{\lambda_n^2}\Big(\sqrt{a_{nih}}+\sqrt{\widehat{a}_i}\Big)^2\,\mathrm{d}x \le c\sum_{i=1}^m b_{nih}^2\widehat{a}_i\Big(\widehat{a}_i+|\sqrt{a_{nih}}-\sqrt{\widehat{a}_i}|^2\Big)\,\mathrm{d}x \\
&\le c\sum_{i=1}^m \|b_{nih}\|_{L^4}^2\Big(1+\|\sqrt{a_{nih}}-\sqrt{\widehat{a}_i}\|_{L^4}^2\Big) \to 0.
\end{aligned}$$

This contradiction shows that the assumption made at the beginning of Step 2 of the proof was wrong, i.e., (4) holds, and the proof is complete. □

3 Conclusions

Since $\widehat{F}(\mathbf{U})-\widehat{F}(\mathbf{u}^*) \le c(U,u^*,\overline{u}) =: \rho$ uniformly for all discretizations we have $\mathbf{v}(t_n) \in \widehat{\mathscr{N}_\rho}$ for $n\ge 1$ for solutions $(\mathbf{u},\mathbf{v})$ to $(\mathrm{P}_{\mathscr{M}})$. Following the proof of [4, Theorem 3.3], but now using the improved result of our Theorem 1, we can choose in step 3 of that proof one $\lambda > 0$ such that $\lambda \mathrm{e}^{\lambda\tau}c_\rho < 1$ uniform for all $\mathscr{M}$, see (A2), too. Especially we do not have any upper restriction on the mesh size, can use admissible finite volume meshes, and obtain

Theorem 2 *We assume* (A1) *and* (A2). *Then there exists a universal* $\lambda > 0$ *such that for all solutions* $(\mathbf{u},\mathbf{v})$ *to* $(P_{\mathscr{M}})$ *the estimate*

$$\widehat{F}(\mathbf{u}(t_n))-\widehat{F}(\mathbf{u}^*) \le \mathrm{e}^{-\lambda t_n}\big(\widehat{F}(\mathbf{U})-\widehat{F}(\mathbf{u}^*)\big) \quad \forall n\ge 1$$

holds uniformly for all discretizations, especially the scheme $(\mathrm{P}_{\mathscr{M}})$ *is dissipative.*

Theorem 2 (as discrete version of [6, Theorem 4.3]) enables us to provide uniform positive lower bounds for the particle densities for the solutions of $(\mathrm{P}_{\mathscr{M}})$ if the order of all reactions is less or equal to two and $N = 2$, see [3, Lemma 4, Theorem 4].

Acknowledgments The work was partially supported by the European Commission within the 7th Framework Programme MD^3 "Material Development for Double Exposure and Double Patterning" and by the DFG Research Center MATHEON *Mathematics for Key Technologies* within project D22 "Modeling of Electronic Properties of Interfaces in Solar Cells".

References

1. Bessemoulin-Chatard, M., Chainais-Hillairet, C., Filbet, F.: On discrete functional inequalities for some finite volume schemes. preprint arXiv:1202.4860v2 (January 15, 2014)
2. Eymard, R., Gallouët, T., Herbin, R.: The finite volume method. In: Ciarlet, P., Lions, J.L. (eds.) Handbook of Numerical Analysis VII, pp. 723–1020. Elsevier (2000)
3. Fiebach, A., Glitzky, A., Linke, A.: Uniform global bounds for solutions of an implicit Voronoi finite volume method for reaction-diffusion problems. Published online in Numer. Math. (2014). doi:10.1007/s00211-014-0604-6
4. Glitzky, A.: Uniform exponential decay of the free energy for Voronoi finite volume discretized reaction-diffusion systems. Math. Nachr. **284**, 2159–2174 (2011)
5. Glitzky, A., Griepentrog, J.A.: Discrete Sobolev-Poincaré inequalities for Voronoi finite volume approximations. SIAM J. Numer. Anal. **48**, 372–391 (2010)
6. Glitzky, A., Gröger, K., Hünlich, R.: Free energy and dissipation rate for reaction-diffusion processes of electrically charged species. Appl. Anal. **60**, 201–217 (1996)

Modified Finite Volume Nodal Scheme for Euler Equations with Gravity and Friction

Emmanuel Franck

Abstract In this work we present a new finite volume scheme valid on unstructured meshes for the Euler equation with gravity and friction indeed the classical Godunov type schemes are not adapted to treat the hyperbolic systems with source terms. The new method is based on a finite volume nodal scheme modified to capture correctly the behavior induced by the source terms.

1 Introduction

In many physical applications appear hyperbolic systems with source terms which model the balance between the convective effects, acoustic effects and the external forces. A classical example of this type of problem is the Euler equations with friction and gravity given by

$$\begin{cases} \partial_t \rho + \dfrac{1}{\varepsilon}\operatorname{div}(\rho \mathbf{u}) = 0 \\ \partial_t \rho \mathbf{u} + \dfrac{1}{\varepsilon}\operatorname{div}(\rho \mathbf{u} \otimes \mathbf{u}) + \dfrac{1}{\varepsilon}\nabla p = \dfrac{1}{\varepsilon}\rho \mathbf{g} - \dfrac{\sigma}{\varepsilon^2}\rho \mathbf{u} \\ \partial_t \rho e + \dfrac{1}{\varepsilon}\operatorname{div}(\rho \mathbf{u} e) + \dfrac{1}{\varepsilon}\operatorname{div}(p\mathbf{u}) = \dfrac{1}{\varepsilon}\rho(\mathbf{g}, \mathbf{u}) - \dfrac{\sigma}{\varepsilon^2}\rho ||\mathbf{u}||^2 \end{cases} \tag{1}$$

with $\mathbf{g}$ a vector of gravity and ε a small parameter which comes from to a rescaling of the time and σ. The limit ε tend to zero correspond to the limit in long time for very large σ. This model is used for the astrophysics applications (for example atmospheric phenomena) and is an interesting model to begin the study of more complicated multi-fluid and multi-phases flows [5, 6]. At the numerical level, it is known that the classical Godunov and splitting schemes are not efficient to capture

E. Franck (✉)
IPP, 2 Boltzmannstrass, Garching, Germany
e-mail: emmanuel.franck@ipp.mpg.de

J. Fuhrmann et al. (eds.), *Finite Volumes for Complex Applications VII - Methods and Theoretical Aspects*, Springer Proceedings in Mathematics & Statistics 77,
DOI: 10.1007/978-3-319-05684-5_27,

the behavior induced by the balance between source terms and hyperbolic part. Since some years, specific numerical methods have been designed, in particular the asymptotic preserving schemes (which capture the asymptotic limit independently of the relaxation parameter ε) and well-balanced schemes which discretize the steady states with a high accuracy. Our aim is to extend this type of method on unstructured meshes to the Euler equations. Firstly we recall some properties at the analytical level.

Proposition 1 *The system (1) satisfies the following properties:*

- *The density and the energy are non negative*
- *The entropy inequality* $\partial_t(\rho S) + \operatorname{div}(\rho \mathbf{u} S) \geq 0$ *is satisfied for weak solutions*
- *When* ε *tends to zero the system tends to*

$$\begin{cases} \partial_t \rho + \operatorname{div}(\rho \mathbf{u}) = 0 \\ \partial_t \rho e + \operatorname{div}(\rho \mathbf{u} e) + p \operatorname{div} \mathbf{u} = 0 \\ \mathbf{u} = \dfrac{1}{\sigma}(\mathbf{g} - \dfrac{1}{\rho}\nabla p) \end{cases} \tag{2}$$

- *The solutions of* $\mathbf{u} = \mathbf{0}$ *and* $\nabla p = \rho \mathbf{g}$ *are steady states (hydrostatic equilibrium) of the system (1)*

Proof The first property is a classical property of the Euler equations. The second and fourth one are discussed in [5].
We give a proof of the asymptotic limit. To obtain this result, we use a Hilbert expansion. Each variable is decomposed on the following form $\rho = \rho^0 + \varepsilon\rho^1 + \varepsilon^2\rho^2 + o(\varepsilon^2)$. Next we plug these definitions in the model.
The terms homogeneous to $\frac{1}{\varepsilon^2}$ are $-\rho^0\mathbf{u}^0 = \mathbf{0}$ and $-\rho^0||\mathbf{u}^0||^2 = 0$. Since ρ is strictly positive we obtain that $\mathbf{u}^0 = \mathbf{0}$.
The term homogeneous to $\frac{1}{\varepsilon}$ is $\nabla p^0 = \rho^0\mathbf{g} - \sigma\rho^0\mathbf{u}^1$.
To finish we give the terms homogeneous to $\frac{1}{\varepsilon^0}$, using $\mathbf{u}^0 = \mathbf{0}$ we have:

$$\begin{aligned} &\partial_t \rho^0 + \operatorname{div}(\rho^0 \mathbf{u}^1) = 0, \\ &\partial_t \rho^0 e^0 + \operatorname{div}(\rho^0 \mathbf{u}^1 e^0) + \operatorname{div}(p^0 \mathbf{u}^1) = \rho^0(\mathbf{g}, \mathbf{u}^1) - \sigma\rho^0||\mathbf{u}_1||^2. \end{aligned} \tag{3}$$

Using the equation $\nabla p^0 = \rho^0\mathbf{g} - \sigma\rho^0\mathbf{u}^1$ we obtain

$$\begin{aligned} &\partial_t \rho^0 + \operatorname{div}(\rho^0 \mathbf{u}^1) = 0, \\ &\partial_t \rho^0 e^0 + \operatorname{div}(\rho^0 \mathbf{u}^1 e^0) + \operatorname{div}(p^0 \mathbf{u}^1) - (\mathbf{u}^1, \nabla p^0) = 0. \end{aligned} \tag{4}$$

with $\mathbf{u}^1 = \frac{1}{\sigma}(\mathbf{g} - \frac{1}{\rho}\nabla p_0)$. To conclude we use $\operatorname{div}(p\mathbf{u}) = (\mathbf{u}, \nabla p) + p \operatorname{div} \mathbf{u}$. □

Now we propose to design a scheme which captures and preserves these properties at the discrete level. To capture the diffusion limit system (2), we use asymptotic preserving (AP) methods.

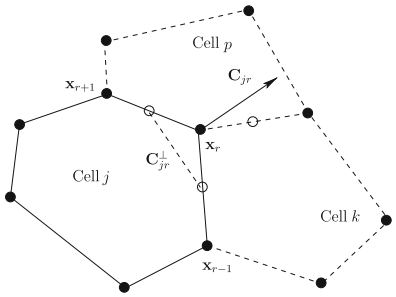


Fig. 1 Notations for nodal scheme. The corner quantity $\mathbf{C}_{jr}$ is equal to the orthogonal vector to the half of the vector that starts at $\mathbf{x}_{r-1}$ and finishes at $\mathbf{x}_{r+1}$. The center of the cell is an arbitrary point inside the cell

For a relaxation model as the Euler equations with the friction terms which depends of ε, the classical schemes like Godunov-type schemes admit a consistency error homogeneous to $O(\frac{\Delta x}{\varepsilon})$ and a CFL condition constrained by ε. However, for the AP schemes the consistency error and the CFL condition are independent of ε [1, 2, 7]. Whereas the well-balanced methods are schemes which discretize exactly or with a high accuracy the steady states [7, 8]. The idea to obtain good discretization is to plug the source terms in the fluxes to capture correctly the effects of these terms.

2 Derivation of the Scheme and Asymptotic Properties

Some asymptotic preserving and well-balanced schemes for Euler equations have been designed in 1D [5, 6]. However the situation is more complicated in 2D. Indeed in [2] we show that the classical extension of the AP scheme for linear hyperbolic systems with diffusion limit does not converge on unstructured meshes. In fact the limit diffusion scheme called 5-points scheme is not consistent on unstructured meshes. To treat this problem we have proposed new scheme based on a nodal formulation (these schemes localize the fluxes at the corner) for different models [2, 3]. Now extend these methods to solve the Euler equations. We use a modified Lagrange+remap scheme (nodal scheme for the Lagrangian part defined in [4] and a nodal advection scheme for the remap part).

Let us consider an unstructured mesh in dimension 2. The mesh is defined by a finite number of vertices $\mathbf{x}_r$ and cells Ω_j. We denote $\mathbf{x}_j$ the center of the cell chosen inside Ω_j. We also define the geometric quantity $\mathbf{C}_{jr} = \nabla_{\mathbf{x}_r} \Omega_j$ (Fig. 1).

Definition 1 The classical Lagrange+remap scheme (LP scheme) is

$$\begin{cases} |\,\Omega_j\,|\,\partial_t \rho_j + \frac{1}{\varepsilon}\left(\sum_{R_+}(\mathbf{C}_{jr},\mathbf{u}_r)\rho_j + \sum_{R_-}(\mathbf{C}_{jr},\mathbf{u}_r)\rho_{k(r)}\right) = 0 \\ |\,\Omega_j\,|\,\partial_t \rho_j\mathbf{u}_j + \frac{1}{\varepsilon}\left(\sum_{R_+}(\mathbf{C}_{jr},\mathbf{u}_r)(\rho\mathbf{u})_j + \sum_{R_-}(\mathbf{C}_{jr},\mathbf{u}_r)(\rho\mathbf{u})_{k(r)} + \sum_r \mathbf{G}_{jr}\right) = 0 \\ |\,\Omega_j\,|\,\partial_t \rho_j e_j + \frac{1}{\varepsilon}\left(\sum_{R_+}(\mathbf{C}_{jr},\mathbf{u}_r)(\rho e)_j + \sum_{R_-}(\mathbf{C}_{jr},\mathbf{u}_r)(\rho e)_{k(r)} + \sum_r(\mathbf{G}_{jr},\mathbf{u}_r)\right) = 0 \end{cases} \tag{5}$$

with the fluxes defined by the problem

$$\begin{cases} \mathbf{G}_{jr} = p_j \mathbf{C}_{jr} + c_{jr}\widehat{\alpha}_{jr}(\mathbf{u}_j - \mathbf{u}_r) \\ \sum_j c_{jr}\widehat{\alpha}_{jr}\mathbf{u}_r = \sum_j p_j \mathbf{C}_{jr} + \sum_j c_{jr}\widehat{\alpha}_{jr}\mathbf{u}_j \end{cases} \tag{6}$$

The wave speed is defined by $c_{jr} = \rho_j c_j$. The expression of the flux $\mathbf{u}_r$ comes from a classical relation of the GLACE scheme: $\sum_j \mathbf{G}_{jr} = \mathbf{0}$. For the advection fluxes we define $\mathbf{u}_{jr} = (\mathbf{C}_{jr}, \mathbf{u}_r)$, $R_+ = (r/\mathbf{u}_{jr} > 0)$, $R_- = (r/\mathbf{u}_{jr} < 0)$ and $\rho_{k(r)} = \frac{\sum_{j/\mathbf{u}_{jr}>0} \mathbf{u}_{jr}\rho_j}{\sum_{j/\mathbf{u}_{jr}>0} \mathbf{u}_{jr}}$.

To obtain an AP scheme, we apply the Jin-Levermore procedure [9]. This method consists to incorporate the steady state of the system into the fluxes. The balance equation between source term and hyperbolic part is $\mathrm{div}(\rho\mathbf{u}\otimes\mathbf{u}) + \nabla p = \rho\mathbf{g} - \frac{\sigma}{\varepsilon}\rho\mathbf{u}$. But the proof of the asymptotic limit shows that $\mathrm{div}(\rho\mathbf{u}\otimes\mathbf{u})$ is negligible in the limit. Indeed the previous equation shows that $\mathbf{u} = O(\varepsilon)$, consequently the important relation for the diffusion regime is $\nabla p + O(\varepsilon^2) = \rho\mathbf{g} - \frac{\sigma}{\varepsilon}\rho\mathbf{u}$. To incorporate this relation into the fluxes we use a first order Taylor expansion $p(\mathbf{x}_j) = p(\mathbf{x}_r) + (\mathbf{x}_j - \mathbf{x}_r, \nabla p(\mathbf{x}_r))$. Using the relation between ∇p and the source term we obtain $p(\mathbf{x}_j) = p(\mathbf{x}_r) + \frac{\sigma}{\varepsilon}(\mathbf{x}_j - \mathbf{x}_r, \rho(\mathbf{x}_r)\mathbf{g} - \frac{\sigma}{\varepsilon}\rho(\mathbf{x}_r)\mathbf{u}(\mathbf{x}_r))$. Now we use the the discrete equivalent of the previous equation: $p_j = p_{jr} + (\mathbf{x}_j - \mathbf{x}_r, \rho_r\mathbf{g} - \frac{\sigma}{\varepsilon}\rho_r\mathbf{u}_r)$. If we consider that $\mathbf{G}_{jr}$ is homogeneous to $p_{jr}\mathbf{C}_{jr}$. We obtain $\mathbf{G}_{jr} \approx p_j\mathbf{C}_{jr} + \widehat{\beta}_{jr}\rho_r(\mathbf{g} - \frac{\sigma}{\varepsilon})\mathbf{u}_r$ with $\widehat{\beta}_{jr} = \mathbf{C}_{jr} \otimes (\mathbf{x}_r - \mathbf{x}_j)$ then we obtain the new fluxes, we plug the previous relation in the fluxes (6). To finish we use discretization localized to the interfaces of the cells for the source term. To justify this discretization we use the following identity $\sum_r \widehat{\beta}_{jr} = \mid \Omega_j \mid \widehat{I}_d$ introduced in [2].

Definition 2 The scheme LP-AP is

$$\begin{cases} \mid \Omega_j \mid \partial_t \rho_j + \frac{1}{\varepsilon}\left(\sum_{R_+} \mathbf{u}_{jr}\rho_j + \sum_{R_-} \mathbf{u}_{jr}\rho_{k(r)}\right) = 0 \\ \mid \Omega_j \mid \partial_t \rho_j\mathbf{u}_j + \frac{1}{\varepsilon}\left(\sum_{R_+} \mathbf{u}_{jr}(\rho\mathbf{u})_j + \sum_{R_-} \mathbf{u}_{jr}(\rho\mathbf{u})_{k(r)} + \sum_r \mathbf{G}_{jr}\right) \\ = \frac{1}{\varepsilon}\sum_r \rho_r\widehat{\beta}_{jr}\left(\mathbf{g} - \frac{\sigma}{\varepsilon}\mathbf{u}_r\right) \\ \mid \Omega_j \mid \partial_t \rho_j e_j + \frac{1}{\varepsilon}\left(\sum_{R_+} \mathbf{u}_{jr}(\rho e)_j + \sum_{R_-} \mathbf{u}_{jr}(\rho e)_{k(r)} + \sum_r (\mathbf{G}_{jr}, \mathbf{u}_r)\right) \\ = \frac{1}{\varepsilon}\left(\sum_r \rho_r\widehat{\beta}_{jr}\mathbf{g} - \frac{\sigma}{\varepsilon}\sum_r (\mathbf{u}_r, \widehat{\beta}_{jr}\mathbf{u}_r)\right) \end{cases} \tag{7}$$

with the fluxes

$$\begin{cases} \mathbf{G}_{jr} = p_j\mathbf{C}_{jr} + c_{jr}\widehat{\alpha}_{jr}(\mathbf{u}_j - \mathbf{u}_r) + \rho_r\widehat{\beta}_{jr}(\mathbf{g} - \frac{\sigma}{\varepsilon}\mathbf{u}_r) \\ \left(\sum_j c_{jr}\widehat{\alpha}_{jr} + \frac{\sigma}{\varepsilon}\rho_r\sum_j \widehat{\beta}_{jr}\right)\mathbf{u}_r = \sum_j p_j\mathbf{C}_{jr} + \sum_j c_{jr}\widehat{\alpha}_{jr}\mathbf{u}_j + \rho_r(\sum_j \widehat{\beta}_{jr})\mathbf{g} \end{cases} \tag{8}$$

Proposition 2 *If the local matrices are invertible and the density is positive then the scheme LP-AP tends formally to the following diffusion scheme*

$$\begin{cases} \mid \Omega_j \mid \partial_t \rho_j + \left(\sum_{R_+} \mathbf{u}_{jr} \rho_j + \sum_{R_-} \mathbf{u}_{jr} \rho_{k(r)} \right) = 0 \\ \mid \Omega_j \mid \partial_t \rho_j e_j + \left(\sum_{R_+} \mathbf{u}_{jr} (\rho e)_j + \sum_{R_-} \mathbf{u}_{jr} (\rho e)_{k(r)} + p_j \sum_r (\mathbf{C}_{jr}, \mathbf{u}_r) \right) = 0 \\ \left(\sum_j \sigma_r \rho_r \widehat{\beta}_{jr} \right) \mathbf{u}_r = \sum_j p_j \mathbf{C}_{jr} + \rho_r \left(\sum_j \widehat{\beta}_{jr} \right) \mathbf{g} \end{cases} \tag{9}$$

Proof To obtain this result, we plug the Hilbert expansion in the scheme (7)–(8). We begin by simplify the source terms with the last part of the fluxes $\sum_r \mathbf{G}_{jr}$ and $\sum_r (\mathbf{G}_{jr}, \mathbf{u}_r)$. After we plug these definitions in the model.
The term homogeneous to $\frac{1}{\varepsilon}$ is $\left(\sum_r \sigma_r \rho_r^0 \widehat{\beta}_{jr} \right) \mathbf{u}_r^0 = \mathbf{0}$. Since the density is positive and the matrix is invertible [2] then $\mathbf{u}_r^0 = \mathbf{0}$. The term in the second equation homogeneous to $\frac{1}{\varepsilon}$ is $\sum_r p_j^0 \mathbf{C}_{jr} + \sum_r c_{jr}^0 \widehat{\alpha}_{jr} (\mathbf{u}_j^0 - \mathbf{u}_r^0) = \mathbf{0}$. Using $\mathbf{u}_r = \mathbf{0}$ and since $\sum_r \mathbf{C}_{jr} = \mathbf{0}$ (property of nodal schemes) this term gives $\sum_r c_{jr}^0 \widehat{\alpha}_{jr} \mathbf{u}_j^0 = \mathbf{0}$. The matrix $\sum_j c_{jr} \widehat{\alpha}_{jr}$ is invertible [4] and $\rho_j > 0$ then $\mathbf{u}_j^0 = \mathbf{0}$. To finish we study the terms homogeneous to $\frac{1}{\varepsilon^0}$ using $\mathbf{u}_r^0 = \mathbf{0}$ and $\mathbf{u}_j^0 = \mathbf{0}$:

$$\begin{aligned} &\partial_t \mid \Omega_j \mid \rho_j^0 + \sum_{R_+} \mathbf{u}_{jr}^1 \rho_j^0 + \sum_{R_-} \mathbf{u}_{jr}^1 \rho_{k(r)}^0 = 0 \\ &\partial_t \mid \Omega_j \mid \rho_j^0 e_j^0 + \sum_r \mathbf{C}_{jr} (p_j^0, \mathbf{u}_r^1) + \sum_{R_+} \mathbf{u}_{jr}^1 \rho_j^0 e_j^0 + \sum_{R_-} \mathbf{u}_{jr}^1 (\rho e)_{k(r)}^0 = 0 \end{aligned} \tag{10}$$

and, since $\mathbf{u}_r^0 = \mathbf{0}$ and $\mathbf{u}_j^0 = \mathbf{0}$, we obtain

$$\sigma_r \rho_r^0 \Big(\sum_j \widehat{\beta}_{jr} \Big) \mathbf{u}_r^1 = \sum_j p_j^0 \mathbf{C}_{jr} + \Big(\sum_j \rho_r^0 \widehat{\beta}_{jr} \Big) \mathbf{g} \tag{11}$$

To finish we couple (11) and (10). □

3 Discretization of the Steady States

For some applications as gravitational flows in astrophysics it is very important to treat with a good accuracy the steady states and to initialize the computations with such steady states, otherwise spurious velocity in the hydrostatic equilibrium configuration ($\nabla p = \rho \mathbf{g}$ and $\mathbf{u} = \mathbf{0}$) may disrupt the simulation. For nearly steady flows numerical perturbations dominate the small physical perturbations. In this section we show that the AP scheme is a well-balanced scheme [8] and is more efficient to treat these configurations. For some equation as shallow water equations a well balanced scheme is a method which preserve exactly the steady states. However this definition is not adapted to study the Euler equations with gravity. Indeed the steady state for the shallow water equations the steady states are algebraic unlike the steady states of the Euler equations which are differential.

Definition 3 (*Well-balanced scheme*) We assume that the initial data $(\rho_j, \mathbf{u}_j, e_j)$ satisfy the discrete steady state at the interface ($\nabla_r p = \rho_r \mathbf{g}$ for Euler equations). A scheme is well-balanced if the scheme is exact for the discrete steady state.

For the Shallow water equations the discrete steady state is an exact discretization of the continuous steady states. This is not the case for the Euler equations. Consequently for the Euler equations the numerical error given by a well-balanced scheme come from only to the error between continuous and discrete steady state.

Lemma 1 *Assume the initial data is given by* $\mathbf{u}_j = \mathbf{0}$ *and* $\nabla_r p = \rho_r \mathbf{g}$ *which is equivalent to*

$$\nabla_r p = -\Big(\sum\nolimits_j \widehat{\beta}_{jr}\Big)^{-1} \sum\nolimits_j p_j \mathbf{C}_{jr} = \rho_r \mathbf{g}$$

with ρ_r *a mean of* ρ_j *around* r. *Then the scheme LP-AP is stationary for the hydrostatic equilibrium.*

Proof We write the nodal solver

$$\Big(\sum\nolimits_j c_{jr}\widehat{\alpha}_{jr} + \frac{\sigma_r}{\varepsilon}\rho_r \sum\nolimits_j \widehat{\beta}_{jr}\Big)\mathbf{u}_r = \sum\nolimits_j p_j \mathbf{C}_{jr} + \sum\nolimits_j \widehat{\alpha}_{jr} c_{jr}\mathbf{u}_j + \Big(\sum\nolimits_j \widehat{\beta}_{jr}\Big)\rho_r \mathbf{g}$$

Using the definition of $\mathbf{u}_j$ and p_j, we obtain that the right hand side term is equal to zero. By uniqueness of the solution $\mathbf{u}_r = \mathbf{0}$. Since $\mathbf{u}_r = \mathbf{0}$ and $\mathbf{u}_j = \mathbf{0}$ then $\mathbf{G}_{jr} = p_j \mathbf{C}_{jr} + \rho_r \widehat{\beta}_{jr}\mathbf{g}$, $\partial_t \rho_j = 0$, $\partial_t \rho_j e_j = 0$ and $\partial_t \rho_j \mathbf{u}_j + \frac{1}{\varepsilon}\sum_r \mathbf{G}_{jr} = \frac{1}{\varepsilon}\sum_r \widehat{\beta}_{jr}\rho_r \mathbf{g}$. Next we use the property $\sum_r \mathbf{C}_{jr} = \mathbf{0}$ consequently we obtain that $\sum_r \mathbf{G}_{jr} = \rho_r \sum_r \widehat{\beta}_{jr}\mathbf{g}$ and we conclude that LP-AP scheme is a WB scheme. □

4 Numerical Results

Firstly we study the convergence of the LP and LP-AP schemes for two different steady states where the density is constant or linear. In the two cases we define $\mathbf{g} = (0, -1)$. The initial data for the first test case are defined by $\rho_j = 1, \mathbf{u}_j = \mathbf{0}$ and $e_j = \frac{1}{\gamma - 1}(\mathbf{x}_j, \mathbf{g}) + C$ with C a constant. The initial data for the second test case are defined by $\rho_j(t, \mathbf{x}) = y + b, \mathbf{u}_j = \mathbf{0}$ and $p_j(t, \mathbf{x}) = -(\frac{y^2}{2} + by)g$.

Now we propose two remarks about the numerical results given in Tables 1 and 2. For the constant density case, the AP scheme preserves exactly the steady state unlike the classical scheme which converges with the first order. For the non constant density case, we remark that the AP scheme is more accurate that the classical scheme. Indeed the AP scheme converge with the second order (Table 2).

We also validate the AP property. For this we consider a Sod problem with $\sigma = 1$ and $\varepsilon = 0.005$. We compare the classical scheme on fine grid (480×480 cells) and coarse grid (60×60 cells) and the AP scheme on coarse grid.

Table 1 L^2 error for the first test case (constant density)

Schemes	LP-AP			LP		
Meshes/cells	40	80	160	40	80	160
Cartesian	5.9×10^{-17}	1×10^{-16}	7.1×10^{-17}	0.00470	0.00239	0.00121
Random	1.1×10^{-16}	1.5×10^{-16}	3×10^{-16}	0.01519	0.00947	0.00526
Kershaw	1.4×10^{-16}	2.2×10^{-16}	3.2×10^{-16}	0.08503	0.050	0.02908

Table 2 L^2 error for the second test case (linear density)

Schemes	LP-AP			LP		
Meshes/cells	80	160	320	80	160	320
Cartesian	2.3×10^{-15}	9.4×10^{-15}	3.4×10^{-14}	0.0034068	0.0016984	0.0000848
Random	3.4×10^{-5}	1×10^{-5}	2.8×10^{-6}	0.00967	0.00529	0.002823
Kershaw	1.1×10^{-6}	1.8×10^{-7}	2.6×10^{-8}	0.03687	0.008363	0.00215

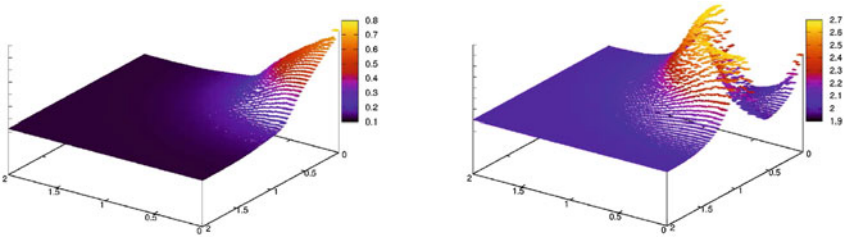


Fig. 2 Density (*left*) and energy (*right*) for the classical LP scheme. Coarse grid

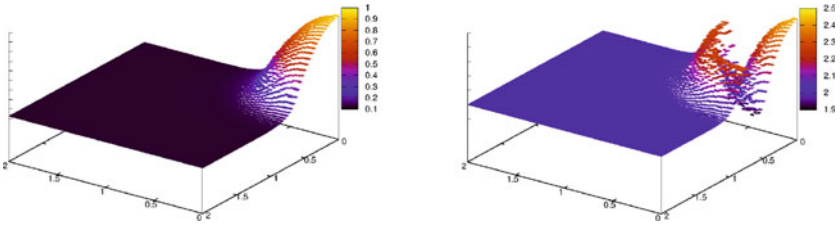


Fig. 3 Density (*left*) and energy (*right*) for the LP AP scheme. Coarse grid

We observe that the AP scheme (Fig. 3) on coarse grid capture correctly the behavior of the solution computed on the fine grid (Fig. 4) at least better than the classical scheme on coarse grid (Fig. 2) which is more diffusive (the numerical viscosity is homogeneous to $\frac{\Delta x}{\varepsilon}$).

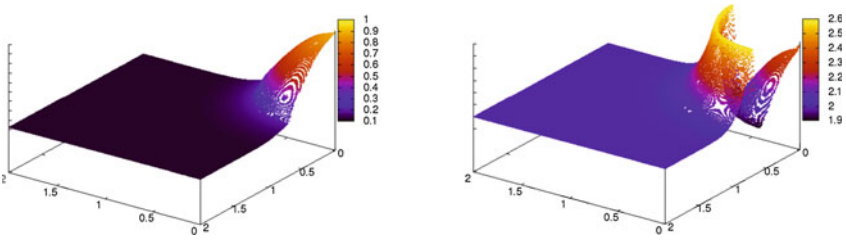


Fig. 4 Density (*left*) and energy (*right*) for the classical LP scheme. Fine grid

5 Conclusion

In this paper we study a modified Lagrange+remap scheme in 2D to capture the behavior induced by the source terms in the Euler equations. We obtain an AP scheme which captures theoretically the diffusion limit independently of the parameter ε. Moreover this scheme preserves experimentally the positivity of ρ and e. To finish, this scheme is well-balanced and converges with the second order for the hydrostatic equilibrium. This new scheme is more accurate than the classical one for these steady states. Contrary to the Shallow water equations where the steady states are algebraic, for the Euler equations the steady states are differential, consequently it is more difficult to obtain a WB scheme exact for all steady states. In the future it will be important to discuss the entropy property and semi-implicit time scheme with a CFL independent of ε.

References

1. Berthon, C., Turpault, R.: Asymptotic preserving hll schemes. Numer. Methods Partial. Diff. Eqn. **27**(6), 1396–1422 (2011)
2. Buet,C., Després, B., Franck, E.: Design of asymptotic preserving schemes for the hyperbolic heat equation on unstructured meshes. Numer. Math. **122**(2), 227–278 (2012)
3. Buet, C., Després, B., Franck, E.: Asymptotic preserving scheme with maximum principle for non linear radiative transfer model on unstructured meshes, C.R. Acad. Sci., Paris, Sér. I, Math., **350**(11–12), 633–638 (2012)
4. Carré, G., Del Pino, S., Després, B., Labourasse, E.: A Cell-centered lagrangian hydrodynamics scheme on general unstructured meshes in arbitrary dimension. JCP. **228**(14), 5160–518 (2009)
5. Chalons, C., Coquel, F., Godlewski, E., Raviart, P-A., Seguin N.: Godunov-type schemes for hyperbolic systems with parameter dependent source. The case of Euler system with friction. M3AS. **20**(11), 2109–2166 2010
6. Chalons C., Girardin M., Kokh S.: Large time step asymptotic preserving numerical schemes for the gas dynamics equations with source terms. SIAM J. Sci. Comput. **35**(6), A2874A2902
7. Gosse, L., Toscani, G.: An asymptotic-preserving well-balanced scheme for the hyperbolic heat equations C. R. Acad. Sci Paris, Ser. I **334**, 337–342 (2002)
8. Greenberg, J., Leroux, A.Y.A.: well balanced scheme for the numerical processing of source terms in hyperbolic equations SIAM J. Numer. Anal **33**(1), 1996
9. Jin S., Levermore D.: Numerical schemes for hyperbolic conservation laws with stiff relaxation terms. JCP. **126**,449–467 (1996)

A Linearity-Preserving Cell-Centered Scheme for the Anisotropic Diffusion Equations

Zhi-Ming Gao and Ji-Ming Wu

Abstract In this paper a cell-centered discretization scheme for the heterogeneous and anisotropic diffusion problems is proposed on general polygonal meshes. The unknowns are the values at the cell center and the scheme relies on linearity-preserving criterion and the use of the harmonic averaging points located at the interface of heterogeneity. Numerical results show that our scheme is robust, and the optimal convergence rates are verified on general distorted meshes in case that the diffusion tensor is taken to be anisotropic, at times discontinuous.

1 Introduction

This paper is contributed to provide a new finite volume scheme for the diffusion problem

$$-\operatorname{div}(\Lambda \nabla u) = f \ \text{ in } \Omega, \tag{1a}$$

$$u = g_D \ \text{ on } \Gamma_D, \tag{1b}$$

$$-\Lambda \nabla u \cdot \boldsymbol{n} = g_N \ \text{ on } \Gamma_N, \tag{1c}$$

where $\Lambda(\boldsymbol{x})$ is a 2×2 diffusion tensor, f is the source term, Ω is an open bounded connected polygonal subset of $\mathbb{R}^2$ with its boundary $\partial\Omega = \bar{\Gamma}_D \cup \bar{\Gamma}_N$, $\boldsymbol{n}$ denotes the outward unit vector normal to the boundary $\partial\Omega$ and g_D, g_N are given scalar functions which are almost everywhere defined on Γ_D, Γ_N respectively.

Z.-M. Gao (✉) · J.-M. Wu
Institute of Applied Physics and Computational Mathematics,
P. O. Box 8009, Beijing 100088, China
e-mail: gao@iapcm.ac.cn

J.-M. Wu
e-mail: wu_jiming@iapcm.ac.cn

J. Fuhrmann et al. (eds.), *Finite Volumes for Complex Applications VII - Methods and Theoretical Aspects*, Springer Proceedings in Mathematics & Statistics 77,
DOI: 10.1007/978-3-319-05684-5_28,

In this paper, we shall relay on the linearity-preserving criterion [2] to derive a new cell-centered finite volume scheme. The linearity-preserving criterion requires that the derivation of a scheme is exact whenever the solution is a linear function and the diffusion coefficient is a constant tensor on each mesh cell. The key point in the construction of our scheme is the discretization of the flux across each cell edge. We first construct the one-sided fluxes on each cell independently and then, integrate the two one-sided fluxes on both edge sides to obtain the unique flux expression. The harmonic averaging point suggested in [1] is another important factor in the construction of the one-sided fluxes. Usually each cell edge has a harmonic averaging point associated with it, which allows our one-side flux expression to process a small stencil involving only the present cell and the cells having a common edge with it. This nature makes it easy to implement our scheme on unstructured polygonal meshes or to extend the scheme to polyhedral meshes.

Our scheme satisfies the following properties:

- it is locally conservative and has a local stencil;
- it allows heterogeneous full diffusion tensors and is reliable on unstructured anisotropic meshes that may be severely distorted;
- it has higher than the first-order accuracy for smooth solutions.

2 Construction of the Scheme

In this paper, a finite volume discretization of Ω is defined as $\mathscr{D} = (\mathscr{M}, \mathscr{E}, \mathscr{O}, \mathscr{P})$, where (1) $\mathscr{M} = \{K\}$ is a finite family of disjoint open polygonal cells in Ω such that $\bar{\Omega} = \cup_{K\in\mathscr{M}} \bar{K}$. (2) $\mathscr{E} = \{\sigma\}$ is a finite family of disjoint edges σ in $\bar{\Omega}$. Let $\mathscr{E}^{int} = \mathscr{E} \cap \Omega$ and $\mathscr{E}^{ext} = \mathscr{E} \cap \partial\Omega$. For $K \in \mathscr{M}$, there exists a subset $\mathscr{E}_K$ of $\mathscr{E}$ such that $\partial K = \cup_{\sigma\in\mathscr{E}_K} \bar{\sigma}$. $\boldsymbol{n}_{K,\sigma}$ denotes the unit vector normal to σ outward to K. (3) $\mathscr{O} = \{\boldsymbol{x}_K, K \in \mathscr{M}\}$ is a set of points, known as cell centers, where $\boldsymbol{x}_K \in K$. (4) $\mathscr{P} = \cup_{K\in\mathscr{M}} \mathscr{P}_K$, where $\mathscr{P}_K = \{\boldsymbol{x}_{K,\sigma}, \sigma \in \mathscr{E}_K\}$ consists of the interpolation points and $\boldsymbol{x}_{K,\sigma}$ is associated with cell K and edge σ.

Approximation of the solution u at the cell center $\boldsymbol{x}_K$ is known as the primary variable and denoted as u_K. The auxiliary variable $u_{K,\sigma}$ is the approximation of u at the interpolation point $\boldsymbol{x}_{K,\sigma}$. $F_{K,\sigma}$ denotes the approximation of $-\int_\sigma (\Lambda_K \nabla u) \cdot \boldsymbol{n}_{K,\sigma}\, \mathrm{d}s$, where we assume that Λ is constant on each cell $K \in \mathscr{M}$ with Λ_K denoting the restriction of Λ on K.

Now we will construct a new cell-centered finite volume scheme, which consists of five steps.

Step 1. Definition of the primary and auxiliary variables

The primary variables are usually defined at the geometric center of K. The auxiliary variables are defined at the interpolation points. For $\sigma \in \mathscr{E}$, we associate it with an interpolation point $\boldsymbol{y}_\sigma$. For the boundary edge $\sigma \in \mathscr{E}^{ext}$, let $\boldsymbol{y}_\sigma$ be the midpoint of σ; for an interior edge $\sigma \in \mathscr{E}_K \cap \mathscr{E}_L$, define

$$\boldsymbol{y}_\sigma = \frac{d_{L,\sigma}\lambda_K^{(n)}\boldsymbol{x}_K + d_{K,\sigma}\lambda_L^{(n)}\boldsymbol{x}_L + d_{K,\sigma}d_{L,\sigma}\left(\Lambda_K^T - \Lambda_L^T\right)\boldsymbol{n}_{K,\sigma}}{d_{L,\sigma}\lambda_K^{(n)} + d_{K,\sigma}\lambda_L^{(n)}}, \tag{2}$$

where $\lambda_K^{(n)} = \boldsymbol{n}_{K,\sigma}^T \Lambda_K \boldsymbol{n}_{K,\sigma}$, $\lambda_L^{(n)} = \boldsymbol{n}_{L,\sigma}^T \Lambda_L \boldsymbol{n}_{L,\sigma}$, and $d_{K,\sigma}$ (resp., $d_{L,\sigma}$) denotes the orthogonal distance from $\boldsymbol{x}_K$ (resp., $\boldsymbol{x}_L$) to σ.

We assume that $(\mathbf{A}_\sigma)$ for any $\sigma \in \mathscr{E}_K \cap \mathscr{E}_L \subset \mathscr{E}$, K (resp., L) is a star-shaped polygonal cell with respect to $\boldsymbol{x}_K$ (resp., $\boldsymbol{x}_L$), and $\boldsymbol{y}_\sigma \in \bar{\sigma}$, then $\boldsymbol{y}_\sigma$ coincides with the *harmonic averaging point* [1, 4] and we have $u(\boldsymbol{y}_\sigma) = \omega_{K,\sigma}u(\boldsymbol{x}_K) + \omega_{L,\sigma}u(\boldsymbol{x}_L)$ with the weights

$$\omega_{K,\sigma} = \frac{d_{L,\sigma}\lambda_K^{(n)}}{d_{K,\sigma}\lambda_L^{(n)} + d_{L,\sigma}\lambda_K^{(n)}}, \quad \omega_{L,\sigma} = \frac{d_{K,\sigma}\lambda_L^{(n)}}{d_{K,\sigma}\lambda_L^{(n)} + d_{L,\sigma}\lambda_K^{(n)}}. \tag{3}$$

Finally, we can choose the set of interpolation points $\mathscr{P}_K = \{\boldsymbol{y}_\sigma, \sigma \in \mathscr{E}_K\}$. Hence we can always write $u_{K,\sigma} = u_{L,\sigma} = u_\sigma$, if $\sigma = \mathscr{E}_K \cap \mathscr{E}_L$, and $u_{K,\sigma} = u_\sigma$, if $\sigma \in \mathscr{E}_K \cap \mathscr{E}^{ext}$.

Remark 1 We point out that when the assumption $(\mathbf{A}_\sigma)$ is violated, the harmonic averaging point $\boldsymbol{y}_\sigma$, defined by (2), is still used as an interpolation point in the present setting. It will be verified in the numerical section.

Step 2. Construction of one-sided flux

For $K \in \mathscr{M}$, once the cell center $\boldsymbol{x}_K$ and the set of interpolation points $\mathscr{P}_K$ are specified, we can establish the one-sided flux expressions through the linearity-preserving approach. For $\sigma \in \mathscr{E}_K$, we choose $\boldsymbol{x}_{K,i(\sigma)}, \boldsymbol{x}_{K,j(\sigma)} \in \mathscr{P}_K$ such that $\Lambda_K^T \boldsymbol{n}_{K,\sigma}$ is located within $\boldsymbol{x}_{K,i(\sigma)} - \boldsymbol{x}_K$ and $\boldsymbol{x}_{K,j(\sigma)} - \boldsymbol{x}_K$. Denote by $\theta_{K,\sigma}^1$ (resp. $\theta_{K,\sigma}^2$) the angle between $\boldsymbol{x}_{K,i(\sigma)} - \boldsymbol{x}_K$ (resp. $\boldsymbol{x}_{K,j(\sigma)} - \boldsymbol{x}_K$) and $\Lambda_K^T \boldsymbol{n}_{K,\sigma}$, we can construct a linearity-preserving one-sided flux of the form [5]

$$F_{K,\sigma} = \alpha_{K,\sigma}(u_K - u_{K,i(\sigma)}) + \beta_{K,\sigma}(u_K - u_{K,j(\sigma)}), \tag{4}$$

where

$$\alpha_{K,\sigma} = \frac{\|\Lambda_K^T \boldsymbol{n}_{K,\sigma}\||\sigma|}{\|\boldsymbol{x}_{K,i(\sigma)} - \boldsymbol{x}_K\|} \frac{\sin\theta_{K,\sigma}^2}{\sin\theta_{K,\sigma}}, \quad \beta_{K,\sigma} = \frac{\|\Lambda_K^T \boldsymbol{n}_{K,\sigma}\||\sigma|}{\|\boldsymbol{x}_{K,j(\sigma)} - \boldsymbol{x}_K\|} \frac{\sin\theta_{K,\sigma}^1}{\sin\theta_{K,\sigma}}.$$

and $\theta_{K,\sigma} = \theta_{K,\sigma}^1 + \theta_{K,\sigma}^2$.

Step 3. A unique definition of edge flux

For an internal edge $\sigma \in \mathscr{E}_K \cap \mathscr{E}_L$, we define the unique edge normal flux

$$\tilde{F}_{K,\sigma} = \omega_{L,\sigma}F_{K,\sigma} - \omega_{K,\sigma}F_{L,\sigma}, \ \tilde{F}_{L,\sigma} = \omega_{K,\sigma}F_{L,\sigma} - \omega_{L,\sigma}F_{K,\sigma}, \tag{5}$$

Table 1 The notations for the schemes used in the numerical computation

Notation	Algorithm description
LPS	The linearity-preserving scheme LPEW2 in [2] with explicit vertex weight
MLPS	A modified LPS obtained by replacing (2.14) in [4] with equal weights in Step 3
SLPS	New linearity-preserving scheme suggested in this paper

where $\omega_{K,\sigma}$ and $\omega_{L,\sigma}$ are given by (3). For $\sigma \in \mathscr{E}_K \cap \mathscr{E}^{ext}$, we simply set $\tilde{F}_{K,\sigma} = F_{K,\sigma}$. Obviously we have the local conservation condition $\tilde{F}_{K,\sigma} + \tilde{F}_{L,\sigma} = 0$.

Step 4. Interpolation of the intermediate variables

To make the finite volume scheme a cell-centered one, we eliminate the auxiliary variables in the flux expressions by an interpolation procedure. Since we have chosen the harmonic averaging points as the interpolation points, we choose

$$u_\sigma = \omega_{K,\sigma} u_K + \omega_{L,\sigma} u_L, \quad \forall\, \sigma \in \mathscr{E}_K \cap \mathscr{E}_L, \tag{6}$$

where $\omega_{K,\sigma}$ and $\omega_{L,\sigma}$ are defined by (3). For $\sigma \in \mathscr{E}_K \cap \mathscr{E}^{ext}$, u_σ is either directly given by Dirichlet boundary data or determined by the expression of $F_{K,\sigma}$ when the flux or mixed boundary condition is imposed.

Step 5. Integration of the finite volume scheme

We formulate the general cell-centered finite volume scheme as follows: find $\{u_K, K \in \mathscr{M}\}$ such that

$$\sum_{\sigma \in \mathscr{E}_K} \tilde{F}_{K,\sigma} = \int_K f(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}, \quad \forall\, K \in \mathscr{M}, \tag{7}$$

where $\tilde{F}_{K,\sigma}$ can be computed from (5), (4) and (6), and it should be noted that the proposed scheme is not symmetric.

3 Numerical Experiments

The notations of the schemes investigated in this section are shown in Table 1.

We use discrete L_2-norm to evaluate discretization errors for the solution:

$$E_u = \left(\sum_{K \in \mathscr{M}} |K| (u(\boldsymbol{x}_K) - u_K)^2 \right)^{\frac{1}{2}}.$$

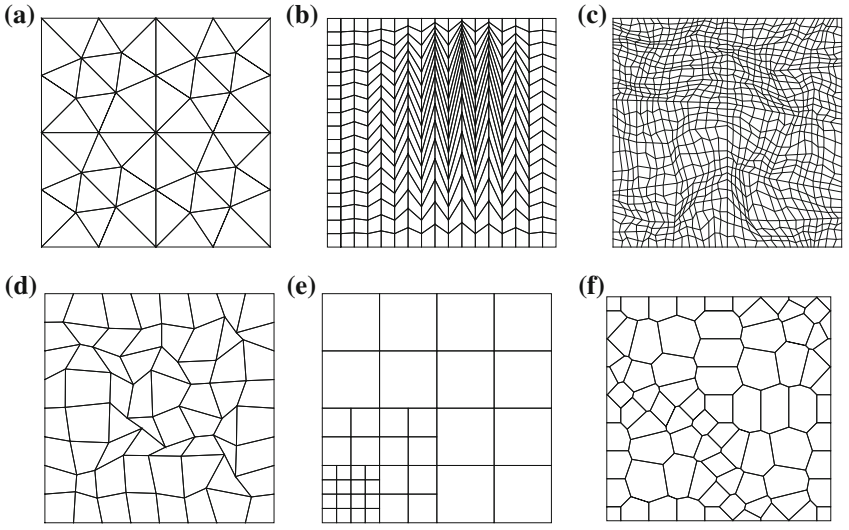


Fig. 1 Samples of the meshes: each mesh was used with 5 successive levels, and the range of associated mesh size h_i $(i = 1, \ldots, 5)$ is shown in the bracket as (h_1, h_5). **a** Mesh1: triangular mesh $(2.5 \times 10^{-1}, 1.56 \times 10^{-2})$. **b** Mesh2: quadrilateral mesh$(3.29 \times 10^{-1}, 6.72 \times 10^{-2})$. **c** Mesh3: Shestakov mesh $(1.70 \times 10^{-1}, 4.27 \times 10^{-2})$. **d** Mesh4: randomly perturbed mesh $(2.85 \times 10^{-1}, 1.91 \times 10^{-2})$. **e** Mesh5: locally refined mesh $(3.54 \times 10^{-1}, 2.21 \times 10^{-2})$. **f** Mesh6: polygonal mesh $(2.29 \times 10^{-1}, 1.49 \times 10^{-2})$

Discrete L_2-norm of the error on the solution gradient can be defined similarly and is denoted by E_q. The rate of convergence R_α $(\alpha = u, q)$ is obtained by a least squares fit on the ones computed on each two successive meshes.

3.1 Test 1: Mild Anisotropy

We consider the problem (1a) with full Dirichlet boundary condition (1b) and $\Omega = [0, 1]^2$. A homogeneous tensor and the exact solution are given below:

$$\Lambda = \begin{pmatrix} 1.5 & 0.5 \\ 0.5 & 1.5 \end{pmatrix}, \quad u(x, y) = \frac{1}{2}\left[\frac{\sin((1-x)(1-y))}{\sin(1)} + (1-x)^3(1-y)^2\right],$$

where the exact solution is located in [0, 1]. This test can be found in FVCA V as a benchmark with a slight modification for the exact solution, and we use a sequence of six mesh types Mesh1–Mesh6 in this numerical test, and the mesh refinement levels are also given in Fig. 1. The numerical results are presented on six mesh types (Mesh1–Mesh6) in Table 2 which shows the following:

Table 2 Results on various meshes Mesh1–Mesh6

Mesh	Scheme	umin	umax	itn	R_u	R_q
Mesh1	LPS	3.99×10^{-3}	0.762	50	1.980	1.378
	MLPS	4.24×10^{-3}	0.763	51	1.977	1.228
	SLPS	3.06×10^{-3}	0.754	43	1.983	1.089
Mesh2	LPS	4.80×10^{-4}	0.911	32	1.887	1.493
	MLPS	4.49×10^{-4}	0.911	32	1.917	1.488
	SLPS	3.56×10^{-4}	0.910	32	1.925	1.496
Mesh3	LPS	1.72×10^{-3}	0.830	39	2.452	1.229
	MLPS	1.76×10^{-3}	0.831	38	1.734	0.405
	SLPS	1.28×10^{-3}	0.827	37	2.484	1.267
Mesh4	LPS	2.63×10^{-3}	0.814	38	2.005	1.070
	MLPS	2.40×10^{-3}	0.813	38	1.957	0.927
	SLPS	1.93×10^{-3}	0.810	37	1.984	1.030
Mesh5	LPS	9.65×10^{-3}	0.906	63	1.852	1.318
	MLPS	9.67×10^{-3}	0.906	66	1.865	1.274
	SLPS	6.66×10^{-3}	0.905	55	2.005	1.502
Mesh6	LPS	7.10×10^{-4}	0.908	36	1.735	1.304
	MLPS	-2.54×10^{-4}	0.907	36	1.749	1.273
	SLPS	4.42×10^{-4}	0.906	36	1.809	1.235

Table 3 Proportion of the edges that violate ($\mathbf{A}_\sigma$) in SLPS

Mesh	h_1 (%)	h_2 (%)	h_3 (%)	h_4 (%)	h_5 (%)
Mesh3	0	0.18	0.19	0.18	0.15
Mesh4	0	1.84	2.79	0.02	0.03
Mesh5	0	0	0	0	0
Mesh6	8.61	10.71	7.56	5.44	3.36

- The minimum and maximum solutions umin and umax are obtained on the coarsest meshes which are exactly those in Fig. 1, and we find that LPS, MLPS and SLPS satisfy the discrete extremum principle except MLPS on Mesh6.
- All schemes show a second order convergence rate with respect to the discrete L_2 norm of the solution on the six types of meshes.
- All schemes have first order convergence rate for the solution gradient with respect to the discrete L_2 norm except schemes MLPS on Mesh3.
- Number of linear iterations itn is presented in the fifth row. Number of linear iterations with respect to three schemes are nearly in the same level.

For the scheme SLPS, the proportions of the edges that violate ($\mathbf{A}_\sigma$) on five successive mesh levels h_1–h_5 are given in Table 3 for Mesh3–Mesh6. It verifies that when the harmonic averaging point does not exist, we can use the harmonic averaging point (2), the resulting scheme can also have expected accuracy.

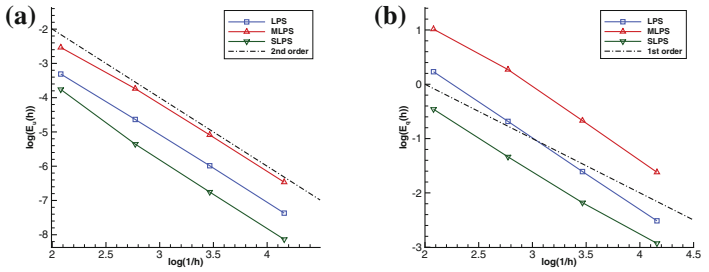


Fig. 2 L_2 errors of the solution and its gradient versus mesh size h on Mesh1. **a** Solution, **b** gradient

3.2 Test 2: Discontinuous Anisotropy

We solve the problem (1a) with the full Dirichlet boundary condition (1b) and $\Omega = [0, 1]^2$. We choose the following diffusion tensor

$$\Lambda = \begin{cases} \begin{pmatrix} 10 & 2 \\ 2 & 5 \end{pmatrix}, & x \leq 0.5, \\ \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, & x > 0.5, \end{cases}$$

and exact solution

$$u(x, y) = \begin{cases} [1 + (x - 0.5)\,(0.1 + 8\pi(y - 0.5))] \exp\left(-20\pi(y - 0.5)^2\right), & x \leq 0.5, \\ \exp(x - 0.5) \exp\left(-20\pi(y - 0.5)^2\right), & x > 0.5. \end{cases}$$

The numerical tests are performed on Mesh1 and Mesh4 with a slight modification that vertices on the line $x = 0.5$ is not disturbed. Firstly in Fig. 2, we notice that on Mesh1, the errors with SLPS are smaller than those with LPS and MLPS. Secondly in Fig. 3, We find that three schemes perform very well on the discontinuous case. In both figures, the expected convergence rates are observed.

3.3 Test 3: Heterogeneous Rotating Anisotropy

In this test, problem (1a)–(1b) is defined on $\Omega = [0, 1]^2$ with the following rotating anisotropic tensor:

$$\Lambda = \frac{1}{x^2 + y^2} \begin{pmatrix} 10^{-3}x^2 + y^2 & (10^{-3} - 1)xy \\ (10^{-3} - 1)xy & x^2 + 10^{-3}y^2 \end{pmatrix},$$

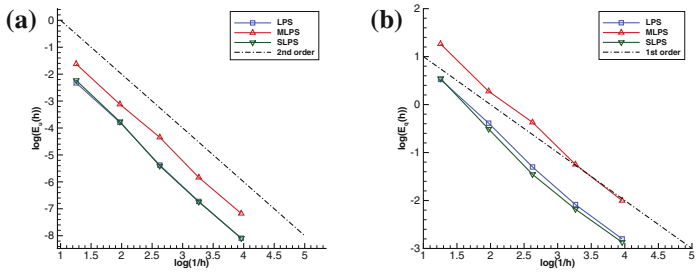


Fig. 3 L_2 errors of the solution and its gradient versus mesh size h on Mesh4. **a** Solution, **b** gradient

Table 4 Numerical results on various meshes

Mesh	Scheme	umin	umax	R_u	R_q
Mesh1	LPS	5.71×10^{-2}	1.002	1.684	0.595
	MLPS	5.37×10^{-2}	1.012	1.663	0.608
	SLPS	4.58×10^{-3}	0.974	1.534	0.545
Mesh5	LPS	2.73×10^{-2}	1.105	1.807	1.383
	MLPS	2.34×10^{-2}	1.064	1.775	1.316
	SLPS	3.95×10^{-2}	0.964	1.371	0.670

and we consider the exact solution $u(x, y) = \sin \pi x \sin \pi y$ in this test. This numerical test can be found in the conference FVCA5 [3].

We present in Table 4 the minimum and maximum solutions umin and umax, the convergence rates for the solution and its gradient on Mesh1 and Mesh5. We find that the scheme SLPS satisfies the extremum principle on the three meshes, but unfortunately, it has a little lower accuracy than LPS and MLPS in this test in case of high anisotropy.

4 Conclusion

We considered a stationary diffusion problem with a full tensor coefficient and suggested a new cell-centered finite volume scheme. The key ingredients in the construction of the scheme are the linearity preserving criterion and the harmonic averaging point. The experiment results on a number of different meshes show that the scheme maintains optimal convergence rates.

Acknowledgments The authors thank the anonymous reviewers for their useful suggestions. This work is supported by the National Natural Science Fundation of China (Nos. 91330107, 61170309, 11135007) and the Science Foundation of China Academy of Engineering Physics (2013B0202034).

References

1. Agelas, L., Eymard, R., Herbin, R.: A nine-point finite volume scheme for the simulation of diffusion in heterogeneous media. CR Acad. Sci. Paris Ser. **I347**, 673–676 (2009)
2. Gao, Z.M., Wu, J.M.: A linearity-preserving cell-centered scheme for the heterogeneous and anisotropic diffusion equations on general meshes. Int. J. Numer. Meth. Fluids **67**(12), 2157–2183 (2011)
3. Herbin, R., Hubert, F.: Benchmark on discretization schemes for anisotropic diffusion problems on general grids. In: Eymard, R., Herard, J.M. (eds.) Finite Volumes for Complex Applications V-Problems and Perspectives, pp. 659–692. Wiley, London (2008)
4. Wu, J.M., Gao, Z.M., Dai, Z.H.: A stabilized linearity-preserving scheme for the heterogeneous and anisotropic diffusion problems on polygonal meshes. J. Comput. Phys. **231**, 7152–7169 (2012)
5. Yuan, G., Sheng, Z.: Monotone finite volume schemes for diffusion equations on polygonal meshes. J. Comput. Phys. **227**(12), 6288–6312 (2008)

Convergence of Finite Volume Scheme for Degenerate Parabolic Problem with Zero Flux Boundary Condition

Boris Andreianov and Mohamed Karimou Gazibo

Abstract This note is devoted to the study of the finite volume methods used in the discretization of degenerate parabolic-hyperbolic equation with zero-flux boundary condition. The notion of an entropy-process solution, successfully used for the Dirichlet problem, is insufficient to obtain a uniqueness and convergence result because of a lack of regularity of solutions on the boundary. We infer the uniqueness of an entropy-process solution using the tool of the nonlinear semigroup theory by passing to the new abstract notion of integral-process solution. Then, we prove that numerical solution converges to the unique entropy solution as the mesh size tends to 0.

1 Introduction

Our goal is to study convergence of a finite volume scheme for a degenerate parabolic equation with zero-flux boundary condition in a regular bounded domain $\Omega \in \mathbb{R}^\ell$ arising, e.g., in sedimentation and traffic models:

$$\begin{cases} u_t + \operatorname{div} f(u) - \Delta\phi(u) = 0 & \text{in} \quad Q = (0,T)\times\Omega, \\ u(0,x) = u_0(x) & \text{in} \quad \Omega, \\ (f(u) - \nabla\phi(u)).\eta = 0 & \text{on} \quad \Sigma = (0,T)\times\partial\Omega. \end{cases} \tag{P}$$

B. Andreianov · M. K. Gazibo (✉)
Laboratoire de Mathématiques CNRS UMR 6623, Université de Franche-Comte,
16 route de Gray, 25030 Besançon, France
e-mail: mgazibok@univ-fcomte.fr

B. Andreianov (✉)
Institut Für Mathematik, Technische Universität Berlin,
Straße des 17. Juni 136, 10623 Berlin, Germany
e-mail: bandreia@univ-fcomte.fr

J. Fuhrmann et al. (eds.), *Finite Volumes for Complex Applications VII - Methods and Theoretical Aspects*, Springer Proceedings in Mathematics & Statistics 77,
DOI: 10.1007/978-3-319-05684-5_29,

Here ϕ is a non-decreasing Lipschitz continuous function, moreover, there exists $u_c \in [0, u_{\max}]$ with $u_{\max} > 0$ such that $\phi|_{[0,u_c]} \equiv 0$ but $\phi'|_{[u_c,u_{\max}]} > 0$. The case $u_c = u_{max}$ was understood in [7]. In the range $[0, u_c]$ of values of u, (P) degenerates into a hyperbolic problem, and admissibility criteria of Kruzhkov type are needed to single out the unique and physically motivated weak solution (see, e.g., [7, 13]). We require that the flux function f is Lipschitz, genuinely nonlinear on $[0, u_c]$; moreover, $[0, u_{\max}]$ is an invariant domain for the evolution of (P) due to assumption

$$f(0) = f(u_{\max}) = 0, \quad u_0 \in L^\infty(\Omega; [0, u_{\max}]) \tag{H1}$$

(the latter means the space of measurable on Ω functions with values in $[0, u_{\max}]$). In the work [4], inspired by [7] we proposed a new entropy formulation of (P) saying that $u \in L^\infty(Q; [0, u_{\max}])$ is an entropy solution of (P) if $u \in C([0, T]; L^1(\Omega))$ with $u(0) = u_0$, $\phi(u) \in L^2(0, T; H^1(\Omega))$ and $\forall k \in [0, u_{\max}]$

$$|u - k|_t + \operatorname{div}\left(sign(u-k)\big[f(u) - f(k) - \nabla\phi(u)\big]\right) \le |f(k).\eta|\, d\mathscr{H}^\ell \tag{1}$$

in $\mathscr{D}'((0, T) \times \overline{\Omega})$, where η is the exterior unit normal vector to the boundary $\Sigma = (0, T) \times \partial\Omega$ and the last term is taken with respect to the Hausdorff measure $\mathscr{H}^\ell$ on Σ. Contrary to the Dirichlet case (cf. [9]) where the boundary condition is relaxed, (1) implies that zero-flux condition in (P) holds in the weak sense.

Existence of an entropy solution to (P) can be obtained by standard vanishing viscosity method, relying in particular on the *strong compactness* arguments derived from genuine nonlinearity of $f|_{[0,u_c]}$ and non-degeneracy of $\phi|_{[u_c,u_{max}]}$, see [12]. But in order to prove uniqueness, one faces a serious difficulty (not relevant in the case $u_c = u_{max}$, [7]) related to the lack of regularity of the flux $\mathscr{F}[u] := f(u) - \nabla\phi(u)$ and specifically, to the weak sense in which the normal component $\mathscr{F}[u].\eta$ of the flux annulates on Σ. Techniques of nonlinear semigroup theory (see, e.g., [5, 6]) can be used to circumvent this regularity problem in some cases (see [3, 4]) and to prove well-posedness for (P) in the sense (1). Let us present the key arguments: indeed, they are also important for study of convergence of the Finite Volume scheme for (P), which is the goal of this note. The standard doubling of variables method based upon formulation (1) readily leads to the uniqueness and L^1 contraction property

$$\forall t \in [0, T] \quad \|u(t, \cdot) - \hat{u}(t, \cdot)\|_{L^1} \le \|u_0 - \hat{u}_0\|_{L^1} \tag{2}$$

if we compare two solutions u, $\hat{u}$ such that the strong (in the sense of L^1 convergence, see [11, 13]) trace of the normal flux $\mathscr{F}[u].\eta$ at the boundary exists. In the sequel, we call such solutions *trace-regular*. Every entropy solution is a trace-regular in the case of the pure hyperbolic problem (case $u_c = u_{\max}$, see [7, 11, 13]). The idea of symmetry breaking in the doubling of variables (see [3]) permits an extension of (2) to a kind of weak-strong comparison principle where u is a general solution and $\hat{u}$ is a trace-regular solution. When a sufficiently large family of trace-regular solutions is available, uniqueness of a general solution and principle (2) may follow

by density arguments. A closely related technique consists in exploiting the above weak-strong comparison arguments using the idea of integral solution and somewhat stronger regularity properties of *stationary solutions*. E.g., for the pure parabolic one ($u_c = 0$, see [3]) every entropy solution of the stationary problem

$$\hat{u} + \operatorname{div} f(\hat{u}) - \Delta\phi(\hat{u}) = g \ \text{ in } \Omega, \quad (f(\hat{u}) - \nabla\phi(\hat{u})).\eta = 0 \text{ on } \partial\Omega \qquad \text{(S)}$$

with $g \in L^\infty(\Omega)$ is trace-regular if $f \circ \phi^{-1} \in \mathbf{C}^{0,\gamma}, \gamma > 0$ (see [3]). This observation, in conjunction with the use of integral solutions [6] of abstract evolution problem

$$u' + Au \ni h, \quad u(0) = u_0 \qquad (3)$$

for suitably defined operator $A = A_{f,\phi}$ (problem (S) taking the form $(\mathrm{Id} + A_{f,\phi})$ $u \ni g$) permits to get uniqueness of entropy solution in [3], for the parabolic case $u_c = 0$. Let us stress that the question of uniqueness for (P) with $u_c \notin \{0, u_{max}\}$ and $\ell > 1$ remains open. The one-dimensional hyperbolic-parabolic case ($\ell = 1$, $\Omega = (a, b)$ with arbitrary $u_c \in [0, u_{\max}]$) has been treated by the authors in [4], using the above abstract approach along with the elementary observation that yields trace-regularity:

$$\big(f(\hat{u}) - \phi(\hat{u})_x\big)_x = g - u \in L^\infty((a,b)) \ \Rightarrow \ \mathscr{F}[u] = \big(f(\hat{u}) - \phi(\hat{u})_x\big) \in \mathbf{C}([a,b]).$$

Another essential aspect of the study of (P) is to justify convergence of numerical approximations. The difference with the existence proof is that, for numerical approximations, the use of *strong compactness* arguments is very technical, and *weak compactness* methods are often preferred. Such study relying on *nonlinear weak-∗ compactness* technique of [8, 9] is our goal in this note. We study a finite volume scheme discretization in the spirit of [9] for (P) on a family of admissible meshes $(\mathscr{O}_h)_h$ with implicit time stepping. According to the standard weak compactness estimates, as for the Dirichlet problem [9] approximate solutions $u^h := u_{\mathscr{O}_h, \delta t_h}$ converge up to a subsequence, as the discretization size h goes to zero, towards an *entropy-process solution* ν. This notion closely related to Young measures' techniques (see [8] and references therein) incorporates dependence on an additional variable $\alpha \in [0, 1]$ which may represent oscillations in the family $(u^h)_h$. It remains to prove the uniqueness of an entropy-process solution which implies the independence of $\nu(t, x, \alpha)$ on α so that $u(t, x) \equiv \nu(t, x, \alpha)$ is an entropy solution of (P). As for the proof of uniqueness of an entropy solution discussed above, we face the major difficulty due to the lack of regularity of $\mathscr{F}[u].\eta$. Hence, we found it useful to define the new notion of *integral-process solution* in the framework of abstract problem (3). Following the pattern of the uniqueness proofs in [3, 4], we compare an entropy-process solution of (P) and a trace regular solution of (S), then we prove that an entropy-process solution of (P) is an integral-process solution of (3) defined for an appropriate m-accretive operator $A_{f,\phi}$. The convergence result holds due to the fact that the integral-process solution coincides with the unique integral solution

of (3); and the latter one coincides with the unique entropy solution of (P) in the sense (1).

The remainder of this note is organized as follows. In Sect. 2 we present our scheme. In Sect. 3 we present the standard steps of convergence arguments for the problem (P), obtained as for Dirichlet problem [9]. In Sect. 4, we achieve the convergence result using classical and new tools of the nonlinear semigroup theory. In Remark 1, we sketch a convergence argument for Finite Volume schemes based upon a direct use of integral-process solutions, bypassing the entropy-process ones.

2 Description of the Finite Volume Scheme for (P)

Let us begin with considering an admissible mesh $\mathscr{O}$ of Ω (see [8, 9]) for space discretization and using the conventional notation present in the main literature. Because we consider the zero-flux boundary condition, we don't need to distinguish between interior and exterior control volumes K, only inner interfaces σ between volumes are needed in order to formulate the scheme. For $K \in \mathscr{O}$ and $\sigma \in \varepsilon_K$, we denote by $\tau_{K,\sigma}$ the transmissivity coefficient. For the approximation of the convective term, we consider the numerical convection fluxes $F_{K,\sigma} : \mathbb{R}^2 \longrightarrow \mathbb{R}$ that are consistent with f, monotone, Lipschitz regular, and conservative (see [8, 9]).

The values of the discrete unknowns u_K^{n+1} for all control volume $K \in \mathscr{O}$, and $n \in \mathbb{N}$ are defined thanks to the following relations: first we initialize the scheme by

$$u_K^0 = \frac{1}{m(K)} \int_K u_0(x)dx \quad \forall K \in \mathscr{O}, \tag{4}$$

then, we use the implicit scheme for the discretization of problem (P):
$\forall n > 0,\ \forall K \in \mathscr{O},$

$$m(K)\frac{u_K^{n+1} - u_K^n}{\delta t} + \sum_{\sigma \in \varepsilon_K} \left(F_{K,\sigma}(u_K^{n+1}, u_{K,\sigma}^{n+1}) - \tau_{K,\sigma}\left(\phi(u_{K,\sigma}^{n+1}) - \phi(u_K^{n+1})\right)\right) = 0. \tag{5}$$

If the scheme has a solution $(u_K^n)_{K,n}$, we will say that the approximate solution to (P) is the piecewise constant function $u_{\mathscr{O},\delta t}(t, x)$ defined by:

$$u_{\mathscr{O},\delta t}(t, x) = u_K^{n+1} \text{ for } x \in K \text{ and } t \in (n\delta t, (n+1)\delta t]. \tag{6}$$

A weakly consistent discrete gradient $\nabla_{\mathscr{O}}\phi(u_{\mathscr{O},\delta t})$ is defined "per diamond"; we refer to [10] for details. Let us stress that the zero-flux boundary condition is included in the scheme, since the flux terms on $\partial K \cap \partial\Omega$ are set to be zero in Eq. (5).

3 Analysis of the Approximate Solution: Classical Arguments

Following the guidelines of [8, 9], we can justify uniqueness of discrete solutions, obtain several uniform estimates (confinement of values of $u_{\mathscr{O},\delta t}$ in $[0, u_{max}]$, weak BV estimate for $u_{\mathscr{O},\delta t}$, discrete $L^2(0,T;H^1(\Omega))$ estimate of $\phi(u_{\mathscr{O},\delta t})$), and derive existence of $u_{\mathscr{O},\delta t}$. We refer to the PhD thesis [10] of the second author for details, with a particular emphasis on the treatment of boundary volumes. It follows that the discrete solution $u_{\mathscr{O},\delta t}$ satisfies the approximate continuous entropy formulation.

Theorem 1 *Let $u_{\mathscr{O},\delta t}$ be the approximate solution of the problem (P) defined by (4),(5),(6). Then the following approximate entropy inequalities hold:*
for all $k \in [0, u_{\max}]$, for all $\xi \in \mathscr{C}^\infty([0,T)\times\mathbb{R}^\ell)$, $\xi \geq 0$,

$$\int_0^T\int_\Omega \left\{|u_{\mathscr{O},\delta t}-k|\xi_t + sign(u_{\mathscr{O},\delta t}-k)\Big[f(u_{\mathscr{O},\delta t})-f(k)-\nabla_{\mathscr{O}}\phi(u_{\mathscr{O},\delta t})\Big].\nabla\xi\right\}dxdt$$
$$+\int_0^T\int_{\partial\Omega}|f(k).\eta(x)|\,\xi(t,x)d\mathscr{H}^{\ell-1}(x)dt+\int_\Omega|u_0-k|\xi(0,x)dx \geq -\upsilon_{\mathscr{O},\delta t}(\xi), \tag{7}$$

where $\forall\xi\in\mathscr{C}^\infty([0,T)\times\mathbb{R}^\ell)$, $\upsilon_{\mathscr{O},\delta t}(\xi)\to 0$ when $h\to 0$.

In order to pass to the limit in (7) using only the L^∞ bound on $u_{\mathscr{O},\delta t}$, one can adapt the notion of an entropy-process solution to problem (P) in the entropy sense (1).

Definition 1 Let $\mu \in L^\infty(Q\times(0,1))$. The function $\mu=\mu(t,x,\alpha)$ is called an entropy-process solution to the problem (P) if $\forall k\in[0,u_{\max}]$, $\forall\xi\in\mathscr{C}^\infty([0,T)\times\mathbb{R}^\ell)$, with $\xi\geq 0$, the following inequalities hold:

$$\int_0^T\int_\Omega\int_0^1\left\{|\mu-k|\xi_t+sign(\mu-k)\Big[f(\mu)-f(k)\Big].\nabla\xi\right\}dxdtd\alpha$$
$$-\int_0^T\int_\Omega\nabla|\phi(u)-\phi(k)|.\nabla\xi dxdt+\int_0^T\int_{\partial\Omega}|f(k).\eta(x)|\,\xi(t,x)d\mathscr{H}^{\ell-1}(x)dt$$
$$+\int_\Omega|u_0-k|\xi(0,x)dx\geq 0,\quad \text{where } u(t,x):=\int_0^1\mu(t,x,\alpha)d\alpha.$$

From Theorem 1 we derive the following result which, however, will not be conclusive. In the sequel, we will upgrade (or circumvent, see Remark 1) this claim.

Proposition 1 *Let $u_{\mathscr{O},\delta t}$ be the approximate solution of the problem (P) defined by (4), (5). There exists an entropy-process solution μ of (P) in the sense of Definition 1 and a subsequence of $(u_{\mathscr{O},\delta t})_{\mathscr{O},\delta t}$, such that:*
• *The sequence $(u_{\mathscr{O},\delta t})_{\mathscr{O},\delta t}$ converges to μ in the nonlinear weak-$*$ sense.*

- *Moreover,* $(\phi(u_{\mathscr{O},\delta t}))_{\mathscr{O},\delta t}$ *converges strongly in* $L^2(Q)$ *to* $\phi(u)$, $u = \int_0^1 \mu(t,x,\alpha)\, d\alpha$, *and* $(\nabla_{\mathscr{O}}\phi(u_{\mathscr{O},\delta t}))_{\mathscr{O},\delta t} \rightharpoonup \nabla\phi(u)$ *in* $(L^2(Q))^\ell$ *weakly, as* $h, \delta t \to 0$.

Proof The proof is essentially the same as in main reference papers dealing with finite volume scheme for degenerate parabolic equations (see [2, 9]). □

4 Reduction of Entropy-Process Solution: Semigroup Arguments

In the context of the Dirichlet problem (see [8, 9]) there holds the uniqueness and reduction result stating that an entropy-process solution μ is α-independent, so that it reduces to an entropy solution. The lack of regularity of the fluxes at the boundary makes it difficult to prove the analogous result with zero-flux conditions. Here, we show how this difficulty can be bypassed, using classical tools and a new notion of *integral-process solution* in the abstract context of nonlinear semigroup theory [6].

4.1 Notion of Integral-Process Solution and Equivalence Result

Given a Banach space X and an accretive operator $A \subset X \times X$, $u \in C([0,T];X)$ is called integral solution (see Bénilan et al. [5, 6]) of the abstract evolution problem (3) if, $\|\cdot\|$ being the norm and $[u,v] := \lim_{\lambda\downarrow 0} \frac{\|u+\lambda v\|-\|u\|}{\lambda}$ the bracket on X, one has $u(0) = u_0$ and the following family of inequalities holds:

$$\forall (\hat{u},\hat{z}) \in A \quad \|u(t)-\hat{u}\| - \|u(s)-\hat{u}\| \le \int_s^t [u(\tau)-\hat{u}, h(\tau)-\hat{z}], \quad 0 \le s \le t \le T.$$

For m-accretive operators the classical in the nonlinear semigroup theory notion of mild solution coincides with the notion of integral solution, so that we have

Proposition 2 *Assume that* A *is* m*-accretive, with* $\overline{Dom(A)}^{\|\cdot\|_X} = X$. *Then for any* $h \in L^1((0,T);X)$, $u_0 \in X$ *there exists a unique integral solution of* (3).

We refer to [6] for the proof of uniqueness of an integral solution and to [5] for a generalization relevant to our case: continuity of $u : [0,T] \to X$ can be relaxed, cf. (9). We propose a variant of the above notion that we call *integral-process solution*. This notion is motivated by an application in the setting where X is a Lebesgue space on $\Omega \subset \mathbb{R}^\ell$ and v is a *nonlinear weak-∗ limit* (see [8]) of approximate solutions.

Definition 2 Let A be an accretive operator on X, $h \in L^1(0,T;X)$ and $u_0 \in X$. An X-valued function v of $(t,\alpha) \in [0,T] \times [0,1]$ is an integral-process solution of abstract problem $u' + Au \ni h$ on $[0,T]$ with datum $v(0,\cdot,\alpha) \equiv u_0(\cdot)$, if for all $(\hat{u},\hat{z}) \in A$

$$\int_0^1 \Big(\|v(t,\alpha) - \hat{u}\| - \|v(s,\alpha) - \hat{u}\| \Big) d\alpha \leq \int_0^1 \int_s^t \Big[v(\tau,\alpha) - \hat{u}, h(\tau) - \hat{z} \Big] d\tau d\alpha \quad (8)$$

for $0 < s \leq t \leq T$ and the initial condition is satisfied in the sense

$$\text{ess-}\lim_{t \downarrow 0} \int_0^1 \|v(t,\alpha) - u_0\| d\alpha = 0. \quad (9)$$

The main fact concerning integral-process solutions is the following result [10].

Theorem 2 *Assume that A is m-accretive in X and $u_0 \in \overline{D(A)}$. Then v is an integral-process solution of (3) if and only if v is independent on α and for all α, $v(.,\alpha)$ coincides with the unique integral and mild solution $u(\cdot)$ of (3).*

4.2 Convergence of the Scheme

Let us define the operator $A_{f,\phi}$ on $L^1(\Omega; [0, u_{\max}]) \subset X = L^1(\Omega)$ endowed with $\|\cdot\|_1$:

$$(v,z) \in A_{f,\phi} = \Big\{ v \text{ such that } v \text{ is a } trace\ regular \text{ solution of (S), with } g = v + z \Big\}$$

(instead of $L^1(\Omega)$ we can work in $L^1(\Omega; [0, u_{\max}])$ due to the confinement principle for solutions of (S)). The main result of this paper is the following theorem.

Theorem 3 *Assume operator $A_{f,\phi}$ on $L^1(\Omega; [0, u_{\max}])$ is m-accretive densely defined, then any entropy-process-solution of (P) is its unique entropy solution. In particular, the scheme (4),(5) for discretization of (P) in the sense (1) is convergent:*

$$\forall p \in [1, +\infty) \quad u_{\mathscr{O},\delta t} \longrightarrow u \ \text{ in } L^p(0, T \times \Omega) \ \text{ as } \ \max(\delta t, h) \longrightarrow 0.$$

Proof First, in Proposition 1 we prove that the approximate solutions $u_{\mathscr{O},\delta t}$ converge towards an entropy-process solution μ. Then, with the technique of [3, 4] we compare the entropy-process solution μ and a trace-regular solution $\hat{u}$ of stationary problem (S). We find that μ is also an integral-process solution. By Theorem 2, μ is independent on the variable α. Therefore $\mu(\cdot,\alpha)$ coincides with the unique integral solution of the abstract evolution problem (3) governed by operator $A_{f,\phi}$; we know from the analysis of [3, 4] that it is also the unique entropy solution of (P). □

Theorem 3 is applicable in the following three cases where trace-regularity for the solutions of (S) can be justified, at least for a dense set of source terms.

Proposition 3 *Assume that $\ell \geq 1$, and $u_c = u_{\max}$ (i.e., (P) is purely hyperbolic). Then $A_{f,\phi}$ is m-accretive densely defined on $L^1(\Omega; [0, u_{\max}])$.*

Proposition 4 *Assume that $\ell \geq 1$ and $u_c = 0$ (i.e. (P) is non-degenerate parabolic). Then $A_{f,\phi}$ is m-accretive densely defined on $L^1(\Omega; [0, u_{\max}])$ if $f \circ \phi^{-1} \in \mathscr{C}^{0,\gamma}$, $\gamma > 0$.*

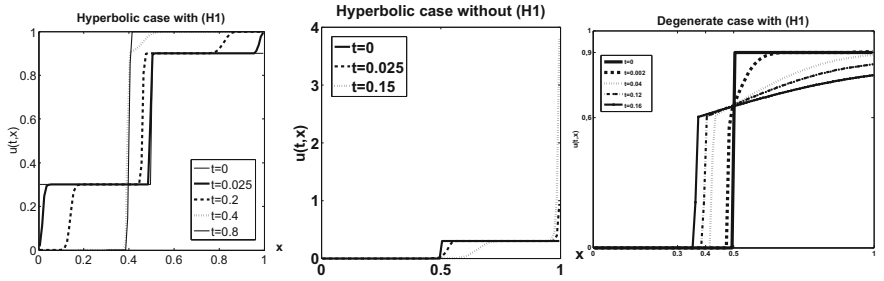


Fig. 1 **a** $f(u) = u(1-u)$, $\phi \equiv 0$, **b** $f(u) = \frac{u^2}{2}$, $\phi \equiv 0$, **c** $f(u) = u(1-u)$, $\phi(u) = (u-0.6)^+$

Proposition 5 *Assume that* $\Omega = (a, b)$ *(thus,* $\ell = 1$*). Then* $A_{f,\phi}$ *is m-accretive densely defined on* $L^1(\Omega; [0, u_{\max}])$.

Prop. 3 follows by the strong trace results of [11, 13] (cf. [7]), Prop. 4 is justified like in [3], while Prop. 5 was an ingredient of the uniqueness proof in [4].

Remark 1 Actually, the use of entropy-process solutions can be circumvented. Observe that the stationary problem (S) can be discretized with the scheme analogous to the time-implicit scheme used for the evolution problem (P). Consider the situation where strong compactness (and convergence to $\hat{u} \in \mathrm{Dom}(A_{f,\phi})$) can be proved for approximate solutions $\hat{u}_{\mathscr{O}}$ of (S) but only nonlinear weak-$*$ compactness for approximate solutions $u_{\mathscr{O},\delta t}$ of (P) is known (this occurs when $\ell = 1$, where compactness of $\hat{u}_{\mathscr{O}}(x_i)$, for all $x_i \in \mathbb{Q}$, is immediate: see the arguments developed in [1]). Then convergence of the stationary scheme is easily proved, moreover, one infers inequalities (8) for the limit $\nu(\cdot, \alpha)$ of $u_{\mathscr{O},\delta t}$. Then, the result of Theorem 2 proves convergence of the scheme for the evolution problem. In a future work, this argument will be applied to a large variety of one-dimensional degenerate parabolic conservation laws with boundary conditions or interface coupling conditions.

5 Numerical Experiments

We conclude with $1D$ numerical illustrations presented in Fig. 1a, c obtained with the explicit analogue of the scheme (4),(5) under the *ad hoc* CFL restrictions. On this occasion, we use the scheme to highlight the importance of hypothesis (H1). In the test of Fig. 1b assumption (H1) fails, and a boundary layer appears. If one refines the mesh one observes convergence of $u_{\mathscr{O}_h,\delta t_h}$ towards a function bounded by $\|u_0\|_\infty$ while the sequence $(u_{\mathscr{O}_h,\delta t_h})_h$ seems unbounded. However, the condition of zero flux imposed in (5) is relaxed in the limit, making formulation (1) inappropriate outside the framework (H1). Introduction of appropriate boundary formulation satisfied by the limit of the scheme, in absence of (H1), is postponed to future work.

Acknowledgments This work has been supported by the French ANR project CoToCoLa.

References

1. Andreianov, B.: In: H.H. Chen G.-Q., K. Karlsen (eds.) Hyperbolic Conservation Laws and Related Analysis with Applications, Springer Proceedings in Mathematics and Statistics, vol. 29, pp. 1–22
2. Andreianov, B., Bendahmane, M., Karlsen, K.: Discrete duality finite volume schemes for doubly nonlinear degenerate hyperbolic-parabolic equations. J. Hyperb. Diff. Eq. 7, 1–67 (2010)
3. Andreianov, B., Bouhsiss, F.: Uniqueness for an elliptic-parabolic problem with Neumann boundary condition. J. Evol. Eq. **4**, 273–295 (2004)
4. Andreianov, B.: Gazibo Karimou, M.: Entropy formulation of degenerate parabolic equation with zero-flux boundary condition. Z. Angew. Math. Phys. **64**(5), 1471–1491 (2013)
5. Barthélémy, L., Bénilan, P.: Subsolutions for abstract evolution equations. Potential Anal. **1**(1), 93–113 (1992)
6. Bénilan, P., Crandall, M.G., Pazy, A.: Nonlinear evolution equations in Banach spaces. (Preprint book)
7. Bürger, R., Frid, H., Karlsen, K.H.: On the well-posedness of entropy solutions to conservation laws with a zero-flux boundary condition. J. Math. Anal. Appl. **326**(1), 108–120 (2007)
8. Eymard, R., Gallouët, T., Herbin, R.: Finite volume methods. Handb. Numer. Anal. **7**, 713–1018 (2000)
9. Eymard, R., Gallouët, T., Herbin, R., Michel, A.: Convergence of a finite volume scheme for nonlinear degenerate parabolic equations. Numer. Math. **92**(1), 41–82 (2002)
10. Gazibo Karimou, M.: Etudes mathématiques et numériques des problèmes paraboliques avec des conditions aux limites. Thèse de Doctorat Besançon (2013)
11. Panov, E.Y.: Existence of strong traces for quasi-solutions of multidimensional conservation laws. J. Hyperb. Diff. Eq. **4**(4), 729–770 (2007)
12. Panov, E.Y.: On the strong pre-compactness property for entropy solutions of a degenerate elliptic equation with discontinuous flux. J. Differ. Eq. **247**(10), 2821–2870 (2009)
13. Vasseur, A.: Strong traces for solutions of multidimensional scalar conservation laws. Arch. Ration. Mech. Anal. **160**(3), 181–193 (2001)

On Aposteriori Error Analysis of DG Schemes Approximating Hyperbolic Conservation Laws

Jan Giesselmann and Tristan Pryer

Abstract This contribution is concerned with aposteriori error analysis of discontinuous Galerkin (dG) schemes approximating hyperbolic conservation laws. In the scalar case the aposteriori analysis is based on the L^1 contraction property and the doubling of variables technique. In the system case the appropriate stability framework is in L^2, based on relative entropies. It is only applicable if one of the solutions, which are compared to each other, is Lipschitz. For dG schemes approximating hyperbolic conservation laws neither the entropy solution nor the numerical solution need to be Lipschitz. We explain how this obstacle can be overcome using a reconstruction approach which leads to an aposteriori error estimate.

1 Introduction

We investigate numerical approximations of systems of hyperbolic conservation laws. The problem has the general form

$$\mathbf{u}_t + \operatorname{div}(\mathbf{f}(\mathbf{u})) = \mathbf{0}, \tag{1}$$

where $\mathbf{u}(t, x) \in U$, for some state space $U \subset \mathbb{R}^d$ and we assume the flux function satisfies $\mathbf{f} \in C^2(U, \mathbb{R}^d)$. We study semi-discretisations of (1) by the discontinuous Galerkin (dG) method and derive an aposteriori error estimate. The discretisation of

J. Giesselmann (✉)
Institute for Applied Analysis and Numerical Simulation, University of Stuttgart, Pfaffenwaldring 57, 70569 Stuttgart, Germany
e-mail: giesselmann@ians.uni-stuttgart.de

T. Pryer
Department of Mathematics and Statistics, Whiteknights, University of Reading, Reading PO Box 220, GB-RG6 6AX, UK
e-mail: t.pryer@reading.ac.uk

J. Fuhrmann et al. (eds.), *Finite Volumes for Complex Applications VII - Methods and Theoretical Aspects*, Springer Proceedings in Mathematics & Statistics 77,
DOI: 10.1007/978-3-319-05684-5_30,

(1) by finite volume and dG schemes is standard, as it is well known that solutions may develop discontinuities in finite time. However, the aposteriori analysis has been only developed for special cases. In [9] aposteriori error estimates (in L^1) are derived in the scalar case. These arguments were generalized to fully applicable Runge–Kutta dG schemes in the scalar case in [4]. As pointed out in [12] these estimates are based on exploiting the L^1-contraction property of scalar hyperbolic conservation laws and the doubling of variables technique. The work [8], which establishes aposteriori error estimates for Friedrichs systems, is in the same spirit, but replaces the L^1 contraction framework by the relative entropy technique, which dates back to [2, 5].

A different approach is the construction of localized aposteriori error estimates via adjoint problems for space-time dG schemes in [7]. Nodal super-convergence of dG schemes was investigated in a sequence of works by Adjerid and coworkers, see [1] and references therein.

All the estimates mentioned before restrict themselves to one of the following two cases:

1. Equation (1) is required to be a scalar equation or a Friedrichs system.
2. Only continuous solutions $\mathbf{u}$ of (1) are considered.

In case of the estimates using adjoint problems the latter restriction is introduced via the stability assumptions on the solutions of the adjoint problems.

The main difficulty in constructing error estimates in the spirit of [4, 8, 9] for (multidimensional) systems of hyperbolic conservation laws without assuming (Lipschitz) continuity of solutions is encapsulated by the following: The appropriate stability theory for this class of PDE is the relative entropy technique. It has certain features (in contrast to the L^1-contraction stability theory available for scalar conservation laws, see [3, Chap. 6.2]) which make its use for constructing aposteriori error estimates more difficult:

1. It cannot be used to compare two discontinuous solutions but it can only compare a Lipschitz continuous solution to another (possibly discontinuous) one. At the same time the numerical solution obtained from a finite volume or dG scheme will be discontinuous and the exact (entropy) solution might also be discontinuous, even for smooth initial data.
2. It leads to an L^2-stability framework which is difficult to use with measure valued residuals, which may not belong to L^2.

We will sketch how to overcome these difficulties for dG spatial discretisations in one space dimension by a reconstruction technique. The details of our arguments, in particular the proofs of Lemma 1 and Theorem 1, can be found in [6]. Our error estimate is expected to be of optimal order (determined by the order of the dG scheme and the regularity of the solution) in the case the entropy solution is Lipschitz continuous. In the case the entropy solution is not Lipschitz the error estimate is not expected to converge, see Remark 7, but in that case uniqueness of the entropy solution cannot be guaranteed anyway.

To avoid any difficulties introduced by boundary conditions, we consider the following version of (1):

$$\mathbf{u}_t + (\mathbf{f}(\mathbf{u}))_x = \mathbf{0} \text{ in } (0,\infty) \times S^1, \tag{2}$$

for some initial data $\mathbf{u}_0 \in L^\infty(S^1, U)$, where S^1 denotes the periodic unit interval with the endpoints being identified with each other. We will assume that (2) is endowed with (at least) one convex entropy/entropy flux pair (η, q), i.e. $q, \eta \in C^1(U, \mathbb{R})$, η is strictly convex and

$$\mathrm{D}\eta\,\mathrm{D}\mathbf{f} = \mathrm{D}q, \tag{3}$$

where D denotes the Jacobian/gradient. For systems of hyperbolic conservation laws there is (usually) only one (physical) entropy/entropy flux pair, while in the scalar case every convex function is an entropy. Equation (3) gives rise to the additional conservation law

$$\eta(\mathbf{u})_t + q(\mathbf{u})_x = 0 \text{ in } (0,\infty) \times S^1, \tag{4}$$

for every strong solution $\mathbf{u}$ of (2). This is crucial for defining entropy solutions:

Definition 1 (*entropy solution*) A function $\mathbf{u} \in L^\infty([0,\infty) \times S^1, U)$ is called an *entropy solution* of the initial boundary value problem (2), with respect to the entropy/entropy-flux pair (η, q), if

$$\int_0^\infty \int_{S^1} \mathbf{u}\cdot\boldsymbol{\phi}_t + \mathbf{f}(\mathbf{u})\cdot\boldsymbol{\phi}_x \,\mathrm{d}x\,\mathrm{d}t + \int_{S^1} \mathbf{u}_0\cdot\boldsymbol{\phi}(0,\cdot)\,\mathrm{d}x = 0 \,\forall\, \boldsymbol{\phi} \in C_c^\infty([0,\infty), \mathbb{R}^d) \tag{5}$$

and

$$\int_0^\infty \int_{S^1} \eta(\mathbf{u})\phi_t + q(\mathbf{u})\phi_x \,\mathrm{d}x\,\mathrm{d}t + \int_{S^1} \eta(\mathbf{u}_0)\phi(0,\cdot)\,\mathrm{d}x \geq 0 \quad \forall \phi \in C_c^\infty([0,\infty) \times S^1, [0,\infty)). \tag{6}$$

We consider approximations of (2) by a class of semi-discrete dG schemes using Godunov type numerical fluxes in Sect. 2. In Sect. 3 we will introduce an explicitly computable reconstruction $\hat{\mathbf{u}}$ of the numerical solution $\mathbf{u}_h$. The reconstruction $\hat{\mathbf{u}}$ is continuous and satisfies a perturbed version of (2) with residuals in L^2. We state an aposteriori estimate, based on the relative entropy framework, of the error between the exact solution $\mathbf{u}$ and the reconstruction $\hat{\mathbf{u}}$ in Sect. 4. This implies an explicitly computable estimate for the difference of the entropy solution $\mathbf{u}$ and the numerical solution $\mathbf{u}_h$, see Theorem 1. We will compare this result to the result from [4] in the scalar case.

2 Semi-Discrete Discontinuous Galerkin Schemes

Before we state the dG schemes under consideration, let us fix some notation. We will discretise (2) in space using a consistent dG scheme. Let $I := [0, 1]$ be the unit interval and choose $0 = x_0 < x_1 < \cdots < x_N = 1$. By $I_n = [x_n, x_{n+1}]$ we denote the n–th sub-interval and by $h_n := x_{n+1} - x_n$ its size. Let $\mathbb{P}_p(I)$ be the space of polynomials of degree less than or equal to p on I, then we denote

$$\mathbb{V}_p := \left\{ \mathbf{g} : I \to \mathbb{R}^d \ : \ (g_i)|_{I_n} \in \mathbb{P}_p(I_n) \text{ for } i = 1, \dots, d, \ n = 0, \dots, N-1 \right\}, \tag{7}$$

where $\mathbf{g} = (g_1, \dots, g_d)^T$, is the usual space of piecewise p–th degree polynomials for vector valued functions over I. In addition, we define jump operators such that

$$[\mathbf{g}]_n := \mathbf{g}(x_n^-) - \mathbf{g}(x_n^+) := \lim_{s \searrow 0} \mathbf{g}(x_n - s) - \lim_{s \searrow 0} \mathbf{g}(x_n + s). \tag{8}$$

We will examine the following class of semi-discrete numerical schemes where $\mathbf{u}_h \in C^1([0, T), \mathbb{V}_p)$ is determined such that

$$0 = \sum_{n=0}^{N-1} \int_{I_n} ((\mathbf{u}_h)_t \cdot \boldsymbol{\phi} - \mathbf{f}(\mathbf{u}_h) \cdot \boldsymbol{\phi}_x) \, \mathrm{d}\, x + \sum_{n=0}^{N-1} \mathbf{F}(\mathbf{u}_h(x_n^-), \mathbf{u}_h(x_n^+)) \cdot [\boldsymbol{\phi}]_n \ \forall \boldsymbol{\phi} \in \mathbb{V}_p. \tag{9}$$

In the sequel we will assume that (9) has a solution and, in particular, that $\mathbf{u}_h$ takes values in U. We also set

$$[\mathbf{u}_h]_0 := \mathbf{u}_h(x_N^-) - \mathbf{u}_h(x_0^+) \tag{10}$$

to account for the periodic boundary conditions. In (9) $\mathbf{F} : U^2 \subset \mathbb{R}^{2d} \to \mathbb{R}^d$ is a numerical flux function. We restrict our attention to a certain class of numerical flux functions. We impose that there exists a function

$$\mathbf{w} : U \times U \to U \text{ such that } \mathbf{F}(\mathbf{u}, \mathbf{v}) = \mathbf{f}(\mathbf{w}(\mathbf{u}, \mathbf{v})) \tag{11}$$

and that there exists a constant $L > 0$ such that $\mathbf{w}$ satisfies

$$|\mathbf{w}(\mathbf{u}, \mathbf{v}) - \mathbf{u}| \leq L|\mathbf{u} - \mathbf{v}|, \qquad |\mathbf{w}(\mathbf{u}, \mathbf{v}) - \mathbf{v}| \leq L|\mathbf{u} - \mathbf{v}| \quad \forall \, \mathbf{u}, \mathbf{v} \in \mathbb{R}^d. \tag{12}$$

Remark 1 The restriction of the flux functions, in general, restricts our analysis to fluxes of Godunov type. Still, fluxes of Roe or Osher-Solomon type fall into this framework in some situations. We need this restriction in order to define the reconstructions in Sect. 3. If we do not have this restriction we may still define reconstructions but the error estimate will no longer be of optimal order for smooth solutions of (2).

3 Reconstructions

In order to derive error estimates for the scheme (9) we introduce reconstructions, which are similar to those used for dG schemes in time in [11], denoted by $\hat{\mathbf{u}}$ and $\hat{\mathbf{f}}$. For brevity we will omit the time dependency of all quantities in this section.

Definition 2 (*Reconstruction of* u_h) The reconstruction $\hat{\mathbf{u}}$ is the unique element of $\mathbb{V}_{p+1}$ such that

$$\sum_{n=0}^{N-1} \int_{I_n} \hat{\mathbf{u}} \cdot \boldsymbol{\phi} \, \mathrm{d}\, x = \sum_{n=0}^{N-1} \int_{I_n} \mathbf{u}_h \cdot \boldsymbol{\phi} \, \mathrm{d}\, x \; \forall \, \boldsymbol{\phi} \in \mathbb{V}_{p-1} \tag{13}$$

and

$$\begin{aligned} \hat{\mathbf{u}}(x_n^+) &= \mathbf{w}(\mathbf{u}_h(x_n^-), \mathbf{u}_h(x_n^+)) \\ \hat{\mathbf{u}}(x_{n+1}^-) &= \mathbf{w}(\mathbf{u}_h(x_{n+1}^-), \mathbf{u}_h(x_{n+1}^+)) \end{aligned} \quad \forall \, n \in \{0, \ldots, N-1\} \tag{14}$$

recalling that $\mathbf{u}_h(x_0^-) := \mathbf{u}_h(x_N^-)$, and $\mathbf{u}_h(x_N^+) := \mathbf{u}_h(x_0^+)$.

Definition 3 (*Reconstruction of* $f(u_h)$) The reconstruction $\hat{\mathbf{f}}$ is the unique element of $\mathbb{V}_{p+1}$ such that

$$\begin{aligned} \sum_{n=0}^{N-1} \int_{I_n} \hat{\mathbf{f}}_x \cdot \boldsymbol{\phi} \, \mathrm{d}\, x = &- \sum_{n=0}^{N-1} \int_{I_n} \mathbf{f}(\mathbf{u}_h) \cdot \boldsymbol{\phi}_x \, \mathrm{d}\, x \\ &+ \sum_{n=0}^{N-1} \mathbf{f}(\mathbf{w}(\mathbf{u}_h(x_n^-), \mathbf{u}_h(x_n^+))) \cdot [\boldsymbol{\phi}]_n \; \forall \, \boldsymbol{\phi} \in \mathbb{V}_p \end{aligned} \tag{15}$$

and

$$\hat{\mathbf{f}}(x_n^+) = \mathbf{f}(\mathbf{w}(\mathbf{u}_h(x_n^-), \mathbf{u}_h(x_n^+))) \; \forall \, n \in \{0, \ldots, N-1\}. \tag{16}$$

Lemma 1 (Properties of the reconstruction) *The reconstructions* $\hat{\mathbf{u}}$ *and* $\hat{\mathbf{f}}$ *are uniquely defined and continuous. Moreover, the reconstructions are explicitly and locally computable.*

Proof The proof of uniqueness and continuity of $\hat{\mathbf{u}}$ is straightforward. To assert the continuity of $\hat{\mathbf{f}}$, we use an analagous argument to that of [11, Lemma 2.1] by testing (15) with piecewise constant functions.

Using the specific reconstruction (15) and (9) we see that

$$0 = \sum_{n=0}^{N-1} \int_{I_n} ((\mathbf{u}_h)_t \cdot \boldsymbol{\phi} - \hat{\mathbf{f}}_x \cdot \boldsymbol{\phi}) \, \mathrm{d}\, x \; \forall \, \boldsymbol{\phi} \in \mathbb{V}_p. \tag{17}$$

As $(\mathbf{u}_h)_t$ and $\hat{\mathbf{f}}_x$ are piecewise polynomials of degree p this implies the *pointwise* equation

$$(\mathbf{u}_h)_t + \hat{\mathbf{f}}_x = \mathbf{0} \quad \text{a.e. in } S^1 \tag{18}$$

which is equivalent to

$$\hat{\mathbf{u}}_t + \mathbf{f}(\hat{\mathbf{u}})_x = \mathbf{R}_h := \hat{\mathbf{u}}_t - (\mathbf{u}_h)_t + \mathbf{f}(\hat{\mathbf{u}})_x - \hat{\mathbf{f}}_x. \tag{19}$$

Remark 2 Equation (19) shows that $\hat{\mathbf{u}}$ solves a perturbed version of (1). As $\mathbf{f}(\hat{\mathbf{u}})$ and $\hat{\mathbf{f}}$ are continuous and piecewise polynomial, $\mathbf{R}_h(t,\cdot) \in L^2(S^1)$ for all $t > 0$. In addition $\mathbf{R}_h$ is explicitly computable.

Remark 3 The reconstruction $\hat{\mathbf{u}}$ is Lipschitz continuous because it is piecewise polynomial and continuous. However, it must be noted that it is not clear whether the Lipschitz constant of $\hat{\mathbf{u}}$ is uniformly bounded if h goes to zero.

Remark 4 It might be expected that the x-derivatives appearing in the definition of $\mathbf{R}_h$ in (19) might lead to a suboptimal order of the error estimate. This is precisely the point at which we need assumption (11) in order to obtain an error estimate of optimal order. For details we refer to [6].

Due to Remarks 2 and 3 the relative entropy framework can be used to estimate the difference between $\hat{\mathbf{u}}$ and the entropy solution $\mathbf{u}$ in terms of $\mathbf{R}_h$ and $\hat{\mathbf{u}}$ even if $\mathbf{u}$ is discontinuous. Once we obtained such an estimate we can estimate the error of the numerical scheme by

$$\|\mathbf{u} - \mathbf{u}_h\|_{L^\infty(0,T;L^2(S^1))} \leq \|\mathbf{u} - \hat{\mathbf{u}}\|_{L^\infty(0,T;L^2(S^1))} + \|\hat{\mathbf{u}} - \mathbf{u}_h\|_{L^\infty(0,T;L^2(S^1))}. \tag{20}$$

4 The Aposteriori Error Estimate

In the remainder of this paper we make the following assumption on the flux and the entropy which is standard in relative entropy arguments. We will assume that there are constants $0 < C_{\overline{\mathbf{f}}} < \infty$ and $0 < C_{\underline{\eta}} < C_{\overline{\eta}} < \infty$ such that

$$|\mathbf{v}^T \mathrm{H}[\mathbf{f}(\mathbf{u})]\mathbf{v}| \leq C_{\overline{\mathbf{f}}}|\mathbf{v}|^2, \quad C_{\underline{\eta}}|\mathbf{v}|^2 \leq \mathbf{v}^T \mathrm{H}[\eta(\mathbf{u})]\mathbf{v} \leq C_{\overline{\eta}}|\mathbf{v}|^2 \quad \forall\, \mathbf{v} \in \mathbb{R}^d, \mathbf{u} \in U, \tag{21}$$

where $|\cdot|$ is the Euclidean norm for vectors, and $\mathrm{H}[\cdot]$ denotes the Hessian of a function or vector field.

Using an analogous argument to [3, Theorem: 5.3.1] we infer

Theorem 1 (Aposteriori error estimate) *Let* $\mathbf{f} \in W_2^\infty(U, \mathbb{R}^d)$ *satisfy* (21). *Let* $\mathbf{u}$ *be an entropy solution of* (2) *with periodic boundary conditions. Then, for* $0 \leq t \leq T$ *the error between the numerical solution* $\mathbf{u}_h$, *given by* (9), *and* $\mathbf{u}$ *satisfies*

$$\begin{aligned}\|\mathbf{u}(t,\cdot)-\mathbf{u}_h(t,\cdot)\|^2_{L^2(S^1)} &\le \|\hat{\mathbf{u}}(t,\cdot)-\mathbf{u}_h(t,\cdot)\|^2_{L^2(S^1)}\\ &\quad + C_{\underline{\eta}}^{-1}\Big(\|\mathbf{R}_h\|^2_{L^2((0,t)\times S^1)} + C_{\overline{\eta}}\|\mathbf{u}_0-\hat{\mathbf{u}}_0\|^2_{L^2(S^1)}\Big)\\ &\quad\times \exp\Big(\int_0^t \frac{C_{\overline{\eta}}C_{\overline{\mathbf{f}}}\|\hat{\mathbf{u}}_x(s,\cdot)\|_{L^\infty(S^1)}+C^2_{\overline{\eta}}}{C_{\underline{\eta}}}\,\mathrm{d}\,s\Big),\end{aligned} \tag{22}$$

where $\hat{\mathbf{u}}$ *is the reconstruction of* $\mathbf{u}_h$ *given in Definition (2) and* $\mathbf{R}_h$ *is defined in* (19).

Remark 5 Note that all the terms on the right hand side of (22) are explicitly computable. Provided $\|\hat{\mathbf{u}}_x(s,\cdot)\|_{L^\infty(S^1)}$ is uniformly bounded in h the right hand side of (22) is expected to be of optimal order. This is expected in case of an at least Lipschitz continuous entropy solution. This is also confirmed by numerical experiments, see [6].

Remark 6 In [6] it is shown that $\|\mathbf{R}_h\|^2_{L^2((0,t)\times S^1)}$ can be estimated without explicitly computing $\hat{\mathbf{u}}$. The estimate [6, Lemma 5.6] is rather technical but it shows that for every $t\in(0,T)$

$$\begin{aligned}\|\mathbf{R}_h\|_{L^2(S^1)} \lesssim C\sum_{n=0}^{N-1} h_n\Big(\big|[\mathbf{u}_h]_n\big|^2+\big|[\mathbf{u}_h]_{n+1}\big|^2\Big)\\ \times\Big(\frac{\big|[\mathbf{u}_h]_n\big|+\big|[\mathbf{u}_h]_{n+1}\big|}{h_n}+\|(\mathbf{u}_h)_x\|_{L^\infty(I_n)}\Big),\end{aligned} \tag{23}$$

where $C>0$ is a computable constant and the "$\lesssim$" in (23) should indicate that there are additional terms needed to estimate $\|\mathbf{R}_h\|_{L^2(S^1)}$ which are of the same order as the right hand side of (23).

Remark 7 Let us compare the estimate which is obtained by combining Theorem 1 and (23) to the estimate which is obtained if the arguments from [4] are applied to our scheme (9) in the scalar case:

$$\begin{aligned}\|u(t,\cdot)-u_h(t,\cdot)\|_{L^1(S^1)} &\le \|u(0,\cdot)-u_h(0,\cdot)\|_{L^1(S^1)}\\ &\quad+\sqrt{K_1\int_0^t\sum_n\big(h_nR_n+h_{n+\frac12}R_{n+\frac12}\big)}\\ &\quad+\sqrt{K_2\int_0^t\sum_n\big(\|\bar u_h-u_h\|_{L^\infty(I_n)}R_n+\big|\bar u_h(x_n^+)-u_h(x_n^+)\big|R_{n+\frac12}\big)},\end{aligned} \tag{24}$$

where $\bar u_h$ is the intervalwise mean of u_h, $h_{n+\frac12}=\frac12(h_n+h_{n+1})$ and

$$\begin{aligned}R_n(t) &:= \textstyle\int_{I_n}|(u_h)_t(t,x)+f(u_h)_x(t,x)|\,\mathrm{d}\,x\\ R_{n+\frac12} &:= \big|[u_h(t,\cdot)]_n+[u_h(t,\cdot)]_{n+1}\big|\end{aligned} \tag{25}$$

where K_1, K_2 are computable constants. In this comparison obviously the two estimators are rather different. The two most important differences making (24) preferable in the scalar case are the following:

1. The estimator in (24) is proportional to $\sqrt{t}$ while the estimator in (22) depends exponentially on time.
2. The estimator in (24) is expected to converge even for discontinuous entropy solutions. In contrast the estimator in (22) depends exponentially on $\|\hat{\mathbf{u}}_x\|_{L^\infty([0,t]\times S^1)}$. For discontinuous entropy solutions $\|\hat{\mathbf{u}}_x\|_{L^\infty([0,t]\times S^1)}$ will be of order h^{-1}. Thus, for discontinuous entropy solutions, the right hand of side (22) will (at best) behave like $h^{2p+2}e^{-1/h}$ which blows up for $h \to 0$. Therefore, while (22) indeed holds even for discontinuous $\mathbf{u}$, the estimator will not converge for $h \to 0$ and its practical use is limited in that case.

Both of these observations are consequences of the use of the relative entropy. We see that it provides a much weaker kind of stability that the L^1-contraction property does for scalar equations.

Remark 8 (Higher space dimensions and time-discretisation) There are two immediate directions for generalisation of the results stated here. The first direction is to derive aposteriori error estimates for fully discrete Runge-Kutta-discontinuous Galerkin schemes. We are optimistic that similar methods to those used in [10] will permit us to obtain such estimates. A special emphasis in this analysis should be put on considering explicit discretisations in time as they are most commonly used in practice.

The second direction is the generalisation to several space dimensions. The crucial issue there is to find appropriate reconstructions of the numerical solution as well as of the numerical fluxes. This is the subject of ongoing research. Once such reconstructions are determined the other arguments presented here can immediately be applied, as they are by no means restricted to the one dimensional case.

References

1. Baccouch, M., Adjerid, S.: Discontinuous Galerkin error estimation for hyperbolic problems on unstructured triangular meshes. Comput. Methods Appl. Mech. Eng. **200**(1–4), 162–177 (2011)
2. Dafermos, C.M.: The second law of thermodynamics and stability. Arch. Ration. Mech. Anal. **70**(2), 167–179 (1979)
3. Dafermos, C.M.: Hyperbolic conservation laws in continuum physics, Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], vol. 325. Springer, Berlin (2010)
4. Dedner, A., Makridakis, C., Ohlberger, M.: Error control for a class of Runge-Kutta discontinuous Galerkin methods for nonlinear conservation laws. SIAM J. Numer. Anal. **45**(2), 514–538 (2007)
5. DiPerna, R.J.: Uniqueness of solutions to hyperbolic conservation laws. Indiana Univ. Math. J. **28**(1), 137–188 (1979)
6. Giesselmann, J., Makridakis, C., Pryer, T.: Aposteriori analysis of discontinuous Galkerin schemes for systems of hyperbolic conservation laws. In preparation
7. Hartmann, R., Houston, P.: Adaptive discontinuous Galerkin finite element methods for nonlinear hyperbolic conservation laws. SIAM J. Sci. Comput. **24**(3), 979–1004 (electronic) (2002)

8. Jovanović, V., Rohde, C.: Finite-volume schemes for Friedrichs systems in multiple space dimensions: a priori and a posteriori error estimates. Numer. Methods Partial Differ. Equ. **21**(1), 104–131 (2005)
9. Kröner, D., Ohlberger, M.: A posteriori error estimates for upwind finite volume schemes for nonlinear conservation laws in multidimensions. Math. Comp. **69**(229), 25–39 (2000)
10. Makridakis, C.: Space and time reconstructions in a posteriori analysis of evolution problems. In: ESAIM proceedings, vol. 21, pp. 31–44. (2007) [Journées d'Analyse Fonctionnelle et Numérique en l'honneur de Michel Crouzeix], (EDP Sci., Les Ulis (2007))
11. Makridakis, C., Nochetto, R.H.: A posteriori error analysis for higher order dissipative methods for evolution problems. Numer. Math. **104**(4), 489–514 (2006)
12. Ohlberger, M.: A review of a posteriori error control and adaptivity for approximations of non-linear conservation laws. Internat. J. Numer. Methods Fluids **59**(3), 333–354 (2009)

Estimating the Geometric Error of Finite Volume Schemes for Conservation Laws on Surfaces for Generic Numerical Flux Functions

Jan Giesselmann and Thomas Müller

Abstract This contribution is concerned with finite volume schemes approximating scalar hyperbolic conservation laws on evolving hypersurfaces of $\mathbb{R}^3$. Theoretical schemes assuming knowledge of all geometric quantities are compared to (practical) schemes defined on moving polyhedra approximating the surface. For the former schemes error estimates have already been proven, but the implementation of such schemes is not feasible for complex geometries. The latter schemes, in contrast, only require (easily) computable geometric quantities and are thus more useful for practical computations. In (Giesselmann and Müller Number. Math. 2014, doi:10.1007/s00211-014-0621-5) an estimate for the difference between solutions of both classes of schemes is proven. This estimate relies on an estimate for the geometric error of the numerical fluxes, which will be investigated in more detail in this contribution.

1 Introduction

Hyperbolic conservation laws serve as models for a wide variety of applications in continuum dynamics. In many applications the problems are posed on (moving) hypersurfaces. Examples include geophysical flows [16], transport processes on cell surfaces [14], surfactant flow on interfaces in multiphase flow [3] and petrol flow on a time dependent water surface. The numerical approximation of such problems was investigated by many groups in recent years, the shallow water equations on a rotating

J. Giesselmann
Institute for Applied Analysis and Numerical Simulation, University of Stuttgart,
Pfaffenwaldring 57, 70569 Stuttgart, Germany
e-mail: jan.giesselmann@mathematik.uni-stuttgart.de

T. Müller (✉)
Abteilung für Angewandte Mathematik, Universität Freiburg,
Hermann-Herder-Street 10, 79104 Freiburg, Germany
e-mail: mueller@mathematik.uni-freiburg.de

J. Fuhrmann et al. (eds.), *Finite Volumes for Complex Applications VII - Methods and Theoretical Aspects,* Springer Proceedings in Mathematics & Statistics 77,
DOI: 10.1007/978-3-319-05684-5_31,

sphere for example were simulated in [4, 10, 15]. As we are interested in numerical analysis we restrict ourselves to the scalar case as a model problem. Well-posedness analysis can be found in [2, 6, 12] and the convergence of appropriate finite volume schemes was investigated in [1, 7, 9, 11].

The hitherto error analysis studied schemes defined on the curved surface assuming exact knowledge of all geometric quantities, e.g. areas and conormals. For engineering applications posed on hypersurfaces of $\mathbb{R}^3$ the geometric quantities are usually not known exactly but need to be approximated. In particular for moving surfaces for which the geometric quantities need to be computed in each time step it is desirable to reduce the computational effort needed to compute the geometric quantities.

In this situation it is important to know to which extent an approximation of the geometry influences the order of convergence.

We consider the following initial value problem, posed on a family of closed, smooth hypersurfaces $\Gamma = \Gamma(t) \subset \mathbb{R}^3$. For a derivation cf. e.g. [6]. For some $T > 0$, find $u : G_T := \bigcup_{t\in[0,T]} \Gamma(t) \times \{t\} \to \mathbb{R}$ with

$$\dot{u} + u\nabla_\Gamma \cdot v + \nabla_\Gamma \cdot f(u, \cdot, \cdot) = 0 \text{ in } G_T, \tag{1}$$

$$u(\cdot, 0) = u_0 \text{ on } \Gamma(0), \tag{2}$$

where v is the velocity of the material points of the surface and $u_0 : \Gamma(0) \to \mathbb{R}$ are initial data. For every $\bar{u} \in \mathbb{R}$, $t \in [0, T]$ the flux $f(\bar{u}, \cdot, t)$ is a smooth vector field tangential to $\Gamma(t)$, which depends Lipschitz on $\bar{u}$ and smoothly on t. We impose the following growth condition

$$|\nabla_\Gamma \cdot f(\bar{u}, x, t)| \leq c + c|\bar{u}| \quad \forall\, \bar{u} \in \mathbb{R}, (x, t) \in G_T \tag{3}$$

for some constant $c > 0$. By $\dot{u}$ we denote the material derivative of u, given by

$$\dot{u}(\Phi_t(x), t) := \frac{d}{dt} u(\Phi_t(x), t),$$

where $\Phi_t : \Gamma(0) \to \Gamma(t)$ is a family of diffeomorphisms depending smoothly on t, such that Φ_0 is the identity on $\Gamma(0)$. Obviously this excludes changes of the topology of Γ. We will assume that the movement of the surface and also the family Φ_t is prescribed. In [8] two approximations of u are considered. They are called the flat approximate and the curved approximate solution, respectively. The curved approximate solution is determined by a finite volume scheme defined on the curved surface, while the flat approximate solution is determined by a finite volume scheme defined on a polyhedron approximating the surface. We will explain these definitions in more detail in Sect. 2. In [8] an estimate for the difference of the curved and the flat approximate solution was obtained. For completeness we will state it as Theorem 1. In [8] it turned out that while the numerical fluxes of the curved scheme need to satisfy the classical consistency, conservation and monotonicity conditions, the fluxes of the flat scheme need to satisfy a geometric error estimate, cf. (9). The main contribution

of this work is a rather generic framework showing that standard numerical fluxes satisfy this condition in Sect. 4.

2 The Finite Volume Schemes

For our analysis the family of triangulations $\mathscr{T}_h(t)$ of the surfaces needs to be suitably linked to polyhedral approximations $\Gamma_h(t)$ of the surfaces.

The triangulation and the definition of the finite volume scheme on Γ_h are in the same spirit as the one in [13], developed for the diffusion equation on evolving surfaces. They are detailed in [8]. Let us simply mention that a triangulation $\bar{\mathscr{T}}_h(t)$ of the polyhedral $\Gamma_h(t)$ is given by its decomposition into faces. Note that in what follows we will consider all faces and edges to be closed sets. We define the triangulation $\mathscr{T}_h(t)$ on $\Gamma(t)$ as the image of $\bar{\mathscr{T}}_h(t)$ under a projection in normal direction from $\Gamma_h(t)$ to $\Gamma(t)$. We denote curved cells with $K(t)$ and the curved faces with $e(t)$. A flat quantity corresponding to some curved quantity is denoted by the same letter and a bar, e.g. let $K(t) \subset \Gamma(t)$ be a curved cell then $\bar{K}(t)$ is the corresponding flat cell. In order to reflect the fact that all triangulations share the same grid topology we introduce the following notation. We denote by K the family of all curved triangles relating to the same triangle $\bar{K}(0)$ on $\Gamma_h(0)$. We do the same for e, $\bar{K}$, $\bar{e}$. Analogously by $\mathscr{T}_h$ we denote the family of such families of triangles K.

We will use the following notation. By $h_{K(t)} := \operatorname{diam}(K(t))$ we denote the diameter of each cell, furthermore $h := \max_{t\in[0,T]} \max_{K(t)} h_{K(t)}$ and $|K(t)|$, $|\partial K(t)|$ are the Hausdorff measures of $K(t)$ and the boundary of $K(t)$ respectively. When we write $e(t) \subset \partial K(t)$ we mean $e(t)$ to be a face of $K(t)$.

In addition, we need to impose the following assumption uniformly on all flat triangulations $\bar{\mathscr{T}}_h(t)$. There is a constant number $\alpha > 0$ such that for each flat cell $\bar{K}(t) \in \bar{\mathscr{T}}_h(t)$ we have

$$\alpha h^2 \leq |\bar{K}(t)| \quad \text{and} \quad \alpha |\partial \bar{K}(t)| \leq h. \tag{4}$$

In [8] it was shown that this implies the respective estimate for the curved triangulation.

2.1 The Finite Volume Scheme on Curved Elements

Let us briefly review the notion of finite volume schemes on moving curved surfaces. We consider a sequence of times $0 = t_0 < t_1 < t_2 < \dots$ and set $I_n := [t_n, t_{n+1}]$. We assign to each $n \in \mathbb{N}$ and $K \in \mathscr{T}_h$ the term u_K^n approximating the mean value of u on $\bigcup_{t\in I_n} K(t) \times \{t\}$ and to each $K \in \mathscr{T}_h$ and face $e \subset \partial K$ a numerical flux function $f_{K,e}^n : \mathbb{R}^2 \to \mathbb{R}$, which should approximate

$$\frac{1}{|I_n|\,|e(t_n)|}\int_{I_n}\int_{e(t)} \langle f(u(x,t),x,t),\mu_{K(t),e(t)}(x)\rangle\, de(t)\,dt, \tag{5}$$

where $de(t)$ is the line element, $\mu_{K(t),e(t)}(x)$ is the unit conormal to $e(t)$ pointing outwards from $K(t)$ and $\langle\cdot,\cdot\rangle$ is the standard Euclidean inner product. Please note that $\mu_{K(t),e(t)}(x)$ is tangential to $\Gamma(t)$. Then the "curved" finite volume scheme is given by

$$\begin{aligned} u_K^0 &:= \int_{K(0)} u_0(x)d\Gamma(0),\\ u_K^{n+1} &:= \frac{|K(t_n)|}{|K(t_{n+1})|}u_K^n - \frac{|I_n|}{|K(t_{n+1})|}\sum_{e\subset\partial K}|e(t_n)|f_{K,e}^n(u_K^n,u_{K_e}^n),\\ u^h(x,t) &:= u_K^n \quad \text{for } t\in[t_n,t_{n+1}),\ x\in K(t), \end{aligned} \tag{6}$$

where K_e denotes the cell sharing face e with K and $d\Gamma(0)$ is the surface element. We assume the numerical fluxes to be consistent, conservative, monotone, and uniformly Lipschitz continuous, which is standard in the error analysis of the curved schemes. Let L denote the Lipschitz constant of the numerical fluxes, then additionally the CFL condition

$$t_{n+1}-t_n \le \frac{\alpha^2 h}{8L} \tag{7}$$

has to be imposed to ensure stability. Note, that the right hand side of (7) is related to the minimal diameter of inner circles of cells through the constant α^2.

2.2 The Finite Volume Scheme on Flat Elements

In this section we define finite volume schemes on $\bar{\mathcal{T}}_h$ which are in the spirit of (6) but only rely on easily accessible geometrical information. We like to point out that the calculation of areas and lengths is straightforward for flat elements. As well, the approximation of integrals can be achieved using quadrature formulas by mapping cells and edges to a standard triangle and the unit interval, respectively, using affine linear maps. In this fashion we obtain for every time $t\in[0,T]$ quadrature operators $Q_{\bar{K}(t)}$ on flat cells and $Q_{\bar{e}(t)}$ on flat edges of order $p_1, p_2 \ge 1$, respectively. In addition for any compact interval $I\subset[0,T]$ the term Q_I denotes a quadrature operator of order $p_3\ge 1$ on I. For Lipschitz continuous numerical flux functions $\bar{f}^n_{\bar{K},\bar{e}}:\mathbb{R}\times\mathbb{R}\to\mathbb{R}$ we define the finite volume scheme on flat elements by

$$\begin{aligned}
\bar{u}^0_{\bar{K}} &:= \frac{1}{|\bar{K}(0)|} Q_{\bar{K}(0)}(u_0(a(\cdot,0))),\\
\bar{u}^{n+1}_{\bar{K}} &:= \frac{|\bar{K}(t_n)|}{|\bar{K}(t_{n+1})|}\bar{u}^n_{\bar{K}} - \frac{|I_n|}{|\bar{K}(t_{n+1})|}\sum_{\bar{e}\subset\partial\bar{K}} |\bar{e}(t_n)|\bar{f}^n_{\bar{K},\bar{e}}(\bar{u}^n_{\bar{K}}, \bar{u}^n_{\bar{K}_{\bar{e}}}), \qquad (8)\\
\bar{u}^h(x,t) &:= \bar{u}^n_{\bar{K}}, \quad \text{for } t\in[t_n,t_{n+1}),\ x\in K(t),
\end{aligned}$$

where a in $(8)_1$ is the projection map in normal direction from Γ_h to Γ, see [5, e.g.]. Note that by $(8)_3$ the function $\bar{u}^h$ is defined on G_T. For the numerical analysis in [8] the following estimate for the (geometric) error between the numerical fluxes $f^n_{K,e}$ and $\bar{f}^n_{\bar{K},\bar{e}}$ ist crucial:

$$\left| f^n_{K,e}(u,v) - \bar{f}^n_{\bar{K},\bar{e}}(u,v)\right| \le Ch^2 \qquad \forall\,(u,v)\in\mathscr{K},\ K\in\mathscr{T}_h,\ e\subset\partial K, \qquad (9)$$

where $\mathscr{K}$ is a compact subset of $\mathbb{R}^2$ and C a constant depending only on G_T, f and $\mathscr{K}$. In particular, C depends on the curvature of the surface, which is bounded, as we consider closed surfaces and finite times. The dependence on the curvature is rather complex and is related to the approximation of all geometric quantities, see [8, Lemma 2, 3 and 4] for details. The compact set $\mathscr{K}$ is due to L^∞-estimates for the finite volume schemes, cf. [8, Lemma 7 and 9], and allows the control of f and its derivatives. It was shown in [8] that (9) holds for the Lax-Friedrichs flux, in case the flat and the curved scheme use the same amount of numerical viscosity.

3 Error Estimate

The main upshot of [8] is the following theorem.

Theorem 1 *For initial data $u_0\in L^\infty(\Gamma(0))$, let u^h denote the solution of the curved finite volume scheme (6) and let $\bar{u}^h$ denote the solution of the flat finite volume scheme (8). Let the quadrature operators $Q_{\bar{K}}(0)$ and the initial data u_0 be such that*

$$\|u^h(0) - \bar{u}^h(0)\|_{L^1(\Gamma(0))} \le C\,h \qquad (10)$$

for some constant C. Let the curved numerical flux functions be consistent, conservative and monotone and let the time step satisfy (7). Let, additionally, (9) hold for the flat numerical flux functions. Then, for fixed $T>0$, the difference between u^h and $\bar{u}^h$ satisfies

$$\|u^h(T) - \bar{u}^h(T)\|_{L^1(\Gamma(T))} \le C\,h, \qquad (11)$$

for some constant C depending on G_T, f, u_0.

Remark 1 Note that the arising geometry errors can be neglected compared to the error between the curved approximate solution and the exact solution, i.e. both

approximate solutions converge to the entropy solution with the same convergence rate $\mathcal{O}(h^{1/4})$, see [9, 11]. Numerical examples in [8] show that the proven convergence rate (11) is optimal under the assumptions for the numerical analysis. However, for most numerical experiments higher orders of convergence are observed.

The analysis indicates that the geometry error poses an obstacle to the construction of higher order schemes. Numerical experiments in [8] show that this is indeed the case. Therefore, to obtain higher order convergence, also the geometry of the curved surface has to be approximated sufficiently accurate. One could think of extending this work to the case of higher order schemes, Discontinuous Galerkin or higher order finite volume schemes e.g., defined on higher order approximations of the moving surface.

In what follows we will present a generic framework to investigate whether flux functions satisfy (9).

4 Determining Which Numerical Fluxes Satisfy (9)

In order to construct numerical fluxes for the curved scheme from standard numerical flux functions, originally developed for equations posed in $\mathbb{R}$, we define, for every $n \in \mathbb{N}$, $K \in \mathcal{T}_h$, $e \in \partial K$, the function

$$c^n_{K,e}(u) := \frac{1}{|I_n|\,|e(t_n)|} \int_{I_n} \int_{e(t)} \langle f(u,x,t), \mu_{K(t),e(t)}(x)\rangle de(t)dt \quad \forall u \in \mathbb{R},$$

that can be interpreted as an approximation of the flux accross the edge e during the time interval I_n.

For a function $\tilde{f} : \mathbb{R} \to \mathbb{R}$, let $G[\tilde{f}] : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ be a one-dimensional numerical flux function that is consistent with $\tilde{f}$, monotone, conservative, i.e.

$$G[\tilde{f}](u,v) = -G[-\tilde{f}](v,u) \quad \forall u, v \in \mathbb{R},$$

and Lipschitz-continuous. From $G : C^1(\mathbb{R}) \to C^1(\mathbb{R}^2)$ we derive a generic family of corresponding numerical flux functions

$$f^n_{K,e}(u,v) := G[c^n_{K,e}](u,v) \text{ for all } u, v \in \mathbb{R} \tag{12}$$

for the curved scheme. Note that monotonicity, conservation, Lipschitz-continuity and consistency for the surface numerical fluxes are inherited from the 1-D numerical fluxes. The Lax-Friedrichs flux from [8] for the curved scheme on moving surfaces can be recovered by ${}^{\mathrm{LF}}f^n_{K,e}(u,v) = {}^{\mathrm{LF}}G[c^n_{K,e}](u,v)$ with the one-dimensional Lax-Friedrichs flux

$${}^{\mathrm{LF}}G[\tilde{f}](u,v) := \frac{1}{2}\left(\tilde{f}(u) + \tilde{f}(v)\right) + \lambda(u-v),$$

for a sufficiently large diffusion coefficient $\lambda \geq 0$. The flat scheme from [8] can be recovered by replacing $c^n_{K,e}$ accordingly, i.e.

$$^{\mathrm{LF}}\bar{f}^n_{\bar{K},\bar{e}}(u,v) = {}^{\mathrm{LF}}G[\bar{c}^n_{\bar{K},\bar{e}}](u,v)$$

with

$$\bar{c}^n_{\bar{K},\bar{e}}(u) := \frac{1}{|I_n|}\mathcal{Q}_{I_n}\left[\frac{1}{|\bar{e}(t_n)|}\mathcal{Q}_{\bar{e}(\cdot)}\left(\langle f(u,\cdot,\cdot),\bar{\mu}_{\bar{K}(\cdot),\bar{e}(\cdot)}\rangle\right)\right].$$

In this manner we can also construct Godunov numerical fluxes on moving surfaces with the help of the one-dimensional Godunov numerical flux functions

$$^{\mathrm{GV}}G[\tilde{f}](u,v) := \begin{cases} \min_{w\in I(u,v)} \tilde{f}(w) & \text{if } u \leq v, \\ \max_{w\in I(u,v)} \tilde{f}(w) & \text{if } v \leq u. \end{cases}$$

In order to show that (9) holds for the (geometric) error between flat and curved numerical fluxes that are based on the same one-dimensional flux functions, we recall from [8, Proof of Lemma 5] that for every compact $\mathscr{K} \subset \mathbb{R}$ there exists $c = c(\mathscr{K}) > 0$ such that

$$|c^n_{K,e}(u) - \bar{c}^n_{\bar{K},\bar{e}}(u)| =: |E^n_{K,e}(u)| \leq ch^2 \quad \forall u \in \mathscr{K}. \tag{13}$$

This estimate at hand it is an easy exercise to show that both, the Lax-Friedrichs and the Godunov scheme satisfy (9), as for any compact $\mathscr{K} \subset \mathbb{R}$

$$\|^{\mathrm{i}}G[\tilde{f}_1] - {}^{\mathrm{i}}G[\tilde{f}_2]\|_{L^\infty(\mathscr{K}^2)} \leq \|\tilde{f}_1 - \tilde{f}_2\|_{L^\infty(\mathscr{K})} \quad \mathrm{i} = \mathrm{LF}, \mathrm{GV}. \tag{14}$$

For the Engquist-Osher flux the situation is slightly different. However, it can be shown analogously to the derivation of (13) that

$$\left|\frac{\mathrm{d}\, c^n_{K,e}}{\mathrm{d}\, u}(u) - \frac{\mathrm{d}\, \bar{c}^n_{\bar{K},\bar{e}}}{\mathrm{d}\, u}(u)\right| \leq ch^2. \tag{15}$$

Moreover, the Engquist-Osher numerical flux operator

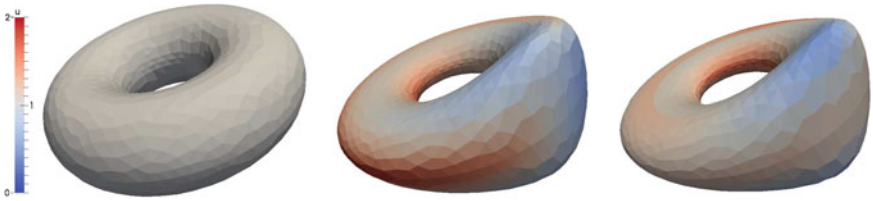
$$^{\mathrm{EO}}G[\tilde{f}](u,v) = \tilde{f}(0) + \int_0^u \max\{\tilde{f}'(s),0\}\,\mathrm{d}\, s + \int_0^v \min\{\tilde{f}'(s),0\}\,\mathrm{d}\, s$$

safisfies

$$\|^{\mathrm{EO}}G[\tilde{f}_1] - {}^{\mathrm{EO}}G[\tilde{f}_2]\|_{L^\infty(\mathscr{K}^2)} \leq C(\mathscr{K})\|\tilde{f}_1' - \tilde{f}_2'\|_{L^\infty(\mathscr{K})} \tag{16}$$

for any compact $\mathscr{K} \subset \mathbb{R}$. Thus, the Engquist-Osher flux satisfies (9).

Acknowledgments We gratefully acknowledge that the work of T.M. was supported by the German Research Foundation (DFG) via SFB TR 71 'Geometric Partial Differential Equations' and by the



(a) $t = 0$. (b) $t = 1.57$. (c) $t = 4.00$.

Fig. 1 Flat approximate solution for (1)–(2) at three different times with Godunov numerical flux functions for the following problem, which was originally studied in [8]: A deforming torus is considered as computational domain Γ and $T = 4$ as final time. Within the time interval $[0, 2]$ the *left half* of the torus undergoes compression whereas the *right half* is stretched, while $\Gamma(t)$ remains constant for $t \in [2, 4]$. A Burgers-type flux function $f = f(u, x) = \frac{1}{2}u^2(x_2, -x_1, 0)^T$ and constant initial values $u_0 \equiv 1$ are chosen. The time step size is chosen dynamically for each time step in order to guarantee stability. In spite of the constant initial values, a shock wave is induced due to the change of geometry (compression and rarefaction) and the nonlinearity of the flux function. Note that the actual computation was performed on a deforming polyhedron approximating the deforming torus. All simulations have been performed within the DUNE-FEM module using AluGrid as grid implementation. Confer [8] for implementation references and more detailed simulations

German National Academic Foundation. J.G. would like to thank the DFG for financial support of the project 'Modeling and sharp interface limits of local and non-local generalized Navier–Stokes–Korteweg Systems'.

We would like to express our gratitude to the two anonymous referees of [8] whose constructive criticism initiated this work.

References

1. Amorim, P., Ben-Artzi, M., LeFloch, P.G.: Hyperbolic conservation laws on manifolds: total variation estimates and the finite volume method. Methods Appl. Anal. **12**(3), 291–323 (2005)
2. Ben-Artzi, M., LeFloch, P.G.: Well-posedness theory for geometry-compatible hyperbolic conservation laws on manifolds. Ann. Inst. H. Poincaré Anal. Non Linéaire **24**(6), 989–1008 (2007)
3. Bothe, D., Prüss, J., Simonett, G.: Well-posedness of a two-phase flow with soluble surfactant. In: Nonlinear Elliptic and Parabolic Problems, Programming Nonlinear Differential Equations Applications, vol. 64, pp. 37–61. Birkhäuser, Basel (2005)
4. Calhoun, D.A., Helzel, C., LeVeque, R.J.: Logically rectangular grids and finite volume methods for PDEs in circular and spherical domains. SIAM Rev. **50**(4), 723–752 (2008)
5. Dziuk, G., Elliott, C.M.: Finite elements on evolving surfaces. IMA J. Numer. Anal. **27**(2), 262–292 (2007)
6. Dziuk, G., Kröner, D., Müller, T.: Scalar conservation laws on moving hypersurfaces. Interfaces Free Boundaries **15**(2), 203–236 (2013)
7. Giesselmann, J.: A convergence result for finite volume schemes on Riemannian manifolds. M2AN Math. Model. Numer. Anal. **43**(5), 929–955 (2009)
8. Giesselmann, J., Müller, T.: Geometric error of finite volume schemes for conservation laws on evolving surfaces. Numer. Math. (2014). doi:10.1007/s00211-014-0621-5

9. Giesselmann, J., Wiebe, M.: Finite volume schemes for balance laws on time-dependent surfaces. In: Numerical Methods for Hyperbolic Equations, pp. 251–258. CRC Press, London (2012)
10. Giraldo, F.X.: High-order triangle-based discontinuous galerkin methods for hyperbolic equations on a rotating sphere. J. Comput. Phys. **214**(2), 447–465 (2006)
11. LeFloch, P.G., Okutmustur, B., Neves, W.: Hyperbolic conservation laws on manifolds. An error estimate for finite volume schemes. Acta Math. Sin. (Engl. Ser) **25**(7), 1041–1066 (2009)
12. Lengeler, D., Müller, T.: Scalar conservation laws on constant and time-dependent riemannian manifolds. J. Differ. Equ. **254**(4), 1705–1727 (2013)
13. Lenz, M., Nemadjieu, S.F., Rumpf, M.: A convergent finite volume scheme for diffusion on evolving surfaces. SIAM J. Numer. Anal. **49**(1), 15–37 (2011)
14. Reister, E., Seifert, U.: Lateral diffusion of a protein on a fluctuating membrane. EPL (Europhysics Letters) **71**(5), 859 (2005)
15. Rossmanith, J.A.: A wave propagation algorithm for hyperbolic systems on the sphere. J. Comput. Phys. **213**(2), 629–658 (2006)
16. Williamson, D.L., Drake, J.B., Hack, J.J., Jakob, R., Swarztrauber, P.N.: A standard test set for numerical approximations to the shallow water equations in spherical geometry. J. Comput. Phys. **102**(1), 211–224 (1992)

Semi-implicit Alternating Discrete Duality Finite Volume Scheme for Curvature Driven Level Set Equation

Angela Handlovičová and Peter Frolkovič

Abstract Linear semi-implicit Alternating Discrete Duality Finite Volume (ADDFV) numerical scheme for the solution of regularized curvature driven level set equation is presented. The scheme requires in each time step to solve algebraic system with a half number of unknowns than necessary in standard DDFV scheme. The stability estimations are proved and comparisons for one numerical experiment are provided.

1 Introduction

The curvature driven level set equation [11]

$$u_t - |\nabla u| \nabla \cdot \left(\frac{\nabla u}{|\nabla u|} \right) = 0 \tag{1}$$

and its non-trivial generalizations are used in many applications, see the references in the quoted papers. In this paper we are mainly interested in numerical schemes for the regularized form of (1) as introduced in e.g. [4, 5]. Namely we study the following equation:

$$u_t - f(|\nabla u|) \nabla \cdot \left(\frac{\nabla u}{f(|\nabla u|)} \right) = 0, \tag{2}$$

A. Handlovičová (✉) · P. Frolkovič
Department of Mathematics, Slovak University of Technology,
Radlinského 11, 813 68 Bratislava, Slovakia
e-mail: angela.handlovicova@stuba.sk

P. Frolkovič
e-mail: peter.frolkovic@stuba.sk

J. Fuhrmann et al. (eds.), *Finite Volumes for Complex Applications VII - Methods and Theoretical Aspects,* Springer Proceedings in Mathematics & Statistics 77,
DOI: 10.1007/978-3-319-05684-5_32,

where the function f is of the form

$$f(z) = \min(\sqrt{z^2 + \varepsilon^2}, b). \tag{3}$$

The fixed regularization parameter $\varepsilon > 0$ in (3) is chosen as in [4] and another real fixed parameter b, $\varepsilon < b$, as in [5] that represents the upper bound of regularized function which is necessary from the numerical analysis point of view. The unknown function $u(t, x)$ in (2) is defined in $I \times \Omega$, where $\Omega \subset \Re^2$ is a rectangular domain, $I = [0, T]$, $T > 0$ is a time interval. We consider zero Dirichlet boundary conditions and prescribed initial condition,

$$u(t, x) = 0 \text{ on } I \times \partial\Omega, \quad u(0, x) = u^0(x), \ u^0 \in H_0^1(\Omega). \tag{4}$$

There are several approaches to numerical solutions of (1) and (2) that are based either on finite difference method [11], finite element method [2] or finite volume method [7, 9]. The Discrete Duality Finite Volumes (DDFV) scheme for elliptic problems is studied in [1, 3, 8]. Recently we have used DDFV for solving the proposed problem, see [6]. Here we study a new approach based on the DDFV method which we call Alternating Discrete Duality Finite Volumes (ADDFV) scheme. We prove the stability estimates for this scheme as it is done in [6] for DDFV scheme. We compare the ADDFV scheme with standard one using numerical example with known analytical solution.

2 Numerical Scheme

First we briefly recall the DDFV numerical scheme proposed in [6]. We choose a uniform discrete time step $\tau = \frac{T}{N_T}$ and replace the time derivative in (2) by the backward finite difference. The nonlinear coefficients in the scheme will be evaluated at previous time step, while the linear ones will be considered at the current time level. In this sense we will use a semi-implicit type of time discretization.

Our primal finite volume mesh $\mathcal{T}_h$ consists of squared cells $V_{ij} \in \mathcal{T}_h$ with the edge of length h, see the dashed lines grid in Fig. 1. The cells are associated with their barycenters $x_{ij} \in V_{ij}$ for $i = 1, \ldots, N_1$, $j = 1, \ldots, N_2$. The value of numerical solution associated with x_{ij} and $t_n = n\tau$ is denoted by u_{ij}^n, see Fig. 2.

The dual finite volume mesh $\overline{\mathcal{T}}_h$ is obtained by shifting the primal mesh in the north-east direction, see the solid lines rectangles in Fig. 1. The mesh consists of cells $\overline{V}_{ij} \in \overline{\mathcal{T}}_h$ associated with their barycenters $\overline{x}_{ij}$ for $i = 0, \ldots, N_1$, $j = 0, \ldots, N_2$ that are at the same time the corners of $V_{ij} \in \mathcal{T}_h$. Analogously to the primal mesh we denote $\overline{u}_{ij}^n \approx u(\overline{x}_{ij}, t^n)$, see Fig. 2. Note that the cells $\overline{V}_{ij}$ at the boundary $\partial\Omega$ are cut in a natural way so that $\overline{\Omega} = \bigcup_{\overline{V}_{ij} \in \overline{\mathcal{T}}_h} \overline{V}_{ij} = \bigcup_{V_{ij} \in \mathcal{T}_h} V_{ij}$, see Fig. 1.

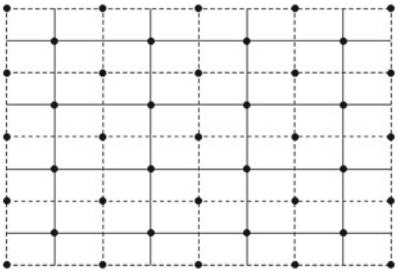


Fig. 1 Primal (*dashed lines rectangles*) and dual (*solid lines rectangles*) mesh

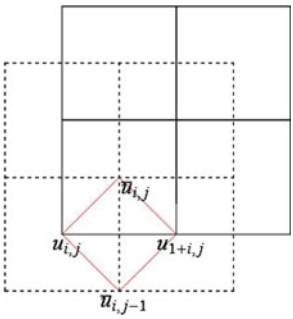


Fig. 2 Values of u in the primal mesh and the values of $\overline{u}$ in the dual mesh

Let $m(V_{ij}) = h^2$ denote the measure of V_{ij}. The edges of V_{ij} are denoted by e_{ij}^{pq} for $p, q = -1, 0, 1$ and $|p| + |q| = 1$ and $m(e_{ij}^{pq}) = h$. The line segments connecting the point x_{ij} with its neighbors $x_{i+p,j+q} \in D$ are denoted by σ_{ij}^{pq} and $m(\sigma_{ij}^{pq}) = h$. Analogous notations using a "bar" is used for the terms related to the cells $\overline{V}_{ij} \in \overline{\mathscr{T}}_h$.

For the approximation of gradients we use a diamond mesh which is the union of $\mathscr{D}_h$ and $\overline{\mathscr{D}}_h$ such that

$$\mathscr{D}_h = \bigcup_{(i,j)=(0,1)}^{(N_1,N_2)} D_{ij}, \quad \overline{\mathscr{D}}_h = \bigcup_{(i,j)=(1,0)}^{(N_1,N_2)} \overline{D}_{ij}\,.$$

The diamond cell D_{ij} has the vertices $\{x_{ij}, \overline{x}_{i,j-1}, x_{i+1,j}, \overline{x}_{ij},\}$ for $(i, j) = (1, 1), \ldots, (N_1 - 1, N_2)$, while for $i = 0$ and $i = N_1$ the diamonds are cut so that $D_{0j} \subset \Omega$ and $D_{N_1 j} \subset \Omega$. Analogously, the diamond cell $\overline{D}_{ij}$ has the vertices $\{x_{ij}, \overline{x}_{ij}, x_{i,j+1}, \overline{x}_{i-1,j}\}$ with the cut diamonds for $j = 0$ and $j = N_2$ such that $D_{i0} \subset \Omega$ and $D_{i\,N_2} \subset \Omega$ and that $\overline{\Omega} = \mathscr{D}_h \cup \overline{\mathscr{D}}_h$.

Having the diamond mesh one can define the approximative gradients ∇u_{ij}^n, resp. $\nabla \overline{u}_{ij}^n$ that are constant on D_{ij}, resp. on $\overline{D}_{ij}$ as follows:

$$\nabla u_{ij}^n = \frac{1}{h}\left(u_{i+1,j}^n - u_{ij}^n, \overline{u}_{ij}^n - \overline{u}_{i,j-1}^n\right), \quad \nabla \overline{u}_{ij}^n = \frac{1}{h}\left(\overline{u}_{ij}^n - \overline{u}_{i-1,j}^n, u_{i,j+1}^n - u_{ij}^n\right). \tag{5}$$

The approximations (5) require additional values that are defined using the zero Dirichlet boundary conditions and eventually the linear extrapolation,

$$u_{0j}^n = -u_{1j}^n, \quad u_{N_1+1\,j}^n = -u_{N_1-1\,j}^n, \quad u_{i0}^n = -u_{i1}^n, \quad u_{i\,N_2+1}^n = -u_{i\,N_2-1}^n, \tag{6}$$

$$\overline{u}_{i0}^n = \overline{u}_{i\,N_2}^n = \overline{u}_{0j}^n = \overline{u}_{N_1\,j}^n = 0. \tag{7}$$

Using (5) one can define the averaged gradients $\tilde{\nabla} u_{ij}^n$, resp. $\tilde{\nabla} \overline{u}_{ij}^n$ that are constant on V_{ij}, resp. $\overline{V}_{ij}$ computed as the arithmetic average of corresponding four diamond gradients. Finally we define the discrete initial values in the standard way:

$$u_{ij}^0 = \frac{1}{m(V_{ij})}\int_{V_{ij}} u^0(x)\,dx \;\; \forall V_{ij} \in \mathscr{T}_h, \quad \overline{u}_{ij}^0 = \frac{1}{m(\overline{V}_{ij})}\int_{\overline{V}_{ij}} u^0(x)\,dx \;\; \forall \overline{V}_{ij} \in \overline{\mathscr{T}_h}. \tag{8}$$

The DDFV scheme can be now obtained using the backward Euler scheme for time derivative and integrating the resulting equation over every V_{ij} and $\overline{V}_{ij}$. After using the divergence theorem and the central finite difference approximation for the directional derivatives we obtain the scheme of the form:

Definition 1 (Fully-discrete semi-implicit DDFV scheme).
Let $i = 1, \ldots, N_1$, $j = 1, \ldots, N_2$ and u_{ij}^0, $\overline{u}_{ij}^0$ be given discrete initial values as defined in (8). Then we search the unknown values u_{ij}^n, $\overline{u}_{ij}^n$, with $n = 1, \ldots, N_T$ satisfying

$$\frac{u_{ij}^n - u_{ij}^{n-1}}{f(|\tilde{\nabla} u_{ij}^{n-1}|)}\,\frac{h^2}{\tau} = \sum_{p=\pm 1}\left(\frac{u_{i+p,j}^n - u_{ij}^n}{f(|\nabla u_{i+\gamma,j}^{n-1}|)} + \frac{u_{i,j+p}^n - u_{ij}^n}{f(|\nabla \overline{u}_{ij+\gamma}^{n-1}|)}\right), \begin{cases} \gamma = 0 \text{ if } p = 1 \\ \gamma = -1 \text{ if } p = -1 \end{cases} \tag{9}$$

$$\frac{\overline{u}_{ij}^n - \overline{u}_{ij}^{n-1}}{f(|\tilde{\nabla} \overline{u}_{ij}^{n-1}|)}\,\frac{h^2}{\tau} = \sum_{p=\pm 1}\left(\frac{\overline{u}_{i+p,j}^n - \overline{u}_{ij}^n}{f(|\nabla \overline{u}_{i+\beta,j}^{n-1}|)} + \frac{\overline{u}_{i,j+p}^n - \overline{u}_{ij}^n}{f(|\nabla u_{ij+\beta}^{n-1}|)}\right), \begin{cases} \beta = 1 \text{ if } p = 1 \\ \beta = 0 \text{ if } p = -1 \end{cases} \tag{10}$$

using also the definitions (5)–(7).

This scheme must compute in every time step two coupled linear algebraic systems having each $N_1 \times N_2$ unknowns. To avoid this we propose the Alternating DDFV scheme with the basic idea to solve (9) only for $n = 2k + 1$ and (10) only for $n = 2k + 2$, where in what follows we use always $k = 1, \ldots, m = N_T/2 - 1$. Note that the schemes (9) and (10) are coupled only through the nonlinear coefficients

that are evaluated in an explicit way as usual for semi-implicit type of numerical schemes.

We study two variants of ADDFV. In both variants we compute the unknowns u_{ij}^1 from (9) with the time step τ and the unknowns $\overline{u}_{ij}^2$ from (10) with the time step 2τ so the nonlinear coefficients in (9) and (10) are always evaluated from u_{ij}^0 and $\overline{u}_{ij}^0$.

The first variant we call *Variant A*. It is obtained in a natural way by replacing the dependence of nonlinear coefficients as in (9) and (10) with the dependence on the approximations of ∇u at time level $n-2$. For that purpose one defines the intermediate missing values by linear interpolation in time, i.e. $u_{ij}^{2k} = (u_{ij}^{2k+1} + u_{ij}^{2k-1})/2$ and $\overline{u}_{ij}^{2k-1} = (\overline{u}_{ij}^{2k} + \overline{u}_{ij}^{2k-2})/2$.

The disadvantage of *Variant A* is that the nonlinear coefficients in the modified schemes (9) and (10) are evaluated in different time levels. Therefore we propose also the second variant of ADDFV that does not have this disadvantage, and, moreover, it allows to prove stability estimates following the approach of [6].

The *Variant B* will use instead of (5) the following approximations of ∇u:

$$\nabla u_{ij}^{2k-\frac{1}{2}} = \frac{1}{h}\left(u_{i+1,j}^{2k-1} - u_{ij}^{2k-1}, \overline{u}_{ij}^{2k} - \overline{u}_{i,j-1}^{2k}\right) \text{ on } D_{ij}, \tag{11}$$

$$\nabla \overline{u}_{ij}^{2k-\frac{1}{2}} = \frac{1}{h}\left(\overline{u}_{ij}^{2k} - \overline{u}_{i-1,j}^{2k}, u_{i,j+1}^{2k-1} - u_{ij}^{2k-1}\right) \text{ on } \overline{D}_{ij}. \tag{12}$$

Analogously we define $\tilde{\nabla} u_{ij}^{2k-\frac{1}{2}}$ and $\tilde{\nabla} \overline{u}_{ij}^{2k-\frac{1}{2}}$.

The *Variant B* of ADDFV can be defined for $k = 1, \ldots, m$ and $i = 1, \ldots, N_1,\ j = 1, \ldots, N_2$ as follows

$$\frac{u_{ij}^{2k+1} - u_{ij}^{2k-1}}{f(|\tilde{\nabla} u_{ij}^{2k-\frac{1}{2}}|)} \frac{h^2}{2\tau} = \sum_{p=\pm 1} \left(\frac{u_{i+p,j}^{2k+1} - u_{ij}^{2k+1}}{f(|\nabla u_{i+\gamma,j}^{2k-\frac{1}{2}}|)} + \frac{u_{i,j+p}^{2k+1} - u_{ij}^{2k+1}}{f(|\nabla \overline{u}_{i,j+\gamma}^{2k-\frac{1}{2}}|)} \right), \tag{13}$$

$$\frac{\overline{u}_{ij}^{2k+2} - \overline{u}_{ij}^{2k}}{f(|\tilde{\nabla} \overline{u}_{ij}^{2k-\frac{1}{2}}|)} \frac{h^2}{2\tau} = \sum_{p=\pm 1} \left(\frac{\overline{u}_{i,j+p}^{2k+2} - \overline{u}_{ij}^{2k+2}}{f(|\nabla u_{ij+\beta}^{2k-\frac{1}{2}}|)} + \frac{\overline{u}_{i+p,j}^{2k+2} - \overline{u}_{ij}^{2k+2}}{f(|\nabla \overline{u}_{i+\beta,j}^{2k-\frac{1}{2}}|)} \right). \tag{14}$$

Remark 1 L_∞ stability for the numerical solution of scheme (13), (14) can be shown in similar way as in [5] so we omit it here.

Theorem 1 *For the solution of the discrete scheme (13), (14) the following stability results holds:*

$$\sum_{k=1}^{m} \left(\sum_{V_{ij} \in \mathcal{T}_h} \frac{\left(u_{ij}^{2k+1} - u_{ij}^{2k-1}\right)^2}{f(|\tilde{\nabla} u_{ij}^{2k-\frac{1}{2}})} \frac{h^2}{2\tau} + \sum_{\overline{V}_{ij} \in \overline{\mathcal{T}}_h} \frac{\left(\overline{u}_{ij}^{2k+2} - \overline{u}_{ij}^{2k}\right)^2}{f(|\tilde{\nabla} \overline{u}_{ij}^{2k-\frac{1}{2}}|)} \frac{h^2}{2\tau} \right)$$

$$+\frac{1}{2b}\sum_{k=1}^{m}\left(\sum_{D_{ij}\in\mathscr{D}_h}\left(|\nabla u_{ij}^{2k+\frac{3}{2}}|-|\nabla u_{ij}^{2k-\frac{1}{2}}|\right)^2h^2+\sum_{\overline{D}_{ij}\in\overline{\mathscr{D}}_h}\left(|\nabla \overline{u}_{ij}^{2k+\frac{3}{2}}|-|\nabla \overline{u}_{ij}^{2k-\frac{1}{2}}|\right)^2h^2\right)$$
$$+\sum_{D_{ij}\in\mathscr{D}_h}|\nabla u_{ij}^{N_T-\frac{1}{2}}|h^2+\sum_{\overline{D}_{ij}\in\overline{\mathscr{D}}_h}|\nabla \overline{u}_{ij}^{N_T-\frac{1}{2}}|h^2\le C, \tag{15}$$

where C is a constant and depends only on data of the problem, not on h or τ.

Proof Note that the index $2k+\frac{3}{2}$ in (15) is used in (11) and (12) as $2(k+1)-\frac{1}{2}$. Analogously as in [5] and [6] we use a function F defined by

$$\forall s\in R_+,\ F(s)=\int_0^s\frac{z}{f(z)}\,dz,\quad F(s)\in\left[\frac{s^2}{2b},\frac{s^2}{2\varepsilon}\right]. \tag{16}$$

We use the following properties of F and f (see [5])

$$\forall c,d\in R_+,\int_c^d\frac{z\,dz}{f(z)}+\frac{(d-c)^2}{2f(c)}\le\frac{d}{f(c)}(d-c). \tag{17}$$

First we multiply (13) and (14) with the term $u_{ij}^{2k+1}-u_{ij}^{2k-1}$ and $\overline{u}_{ij}^{2k+2}-\overline{u}_{ij}^{2k}$, respectively, and sum them for $V_{ij}\in\mathscr{T}_h$ and $\overline{V}_{ij}\in\overline{\mathscr{T}}_h$. By rearranging terms in both equations and using some standard procedures in finite volume methods [5, 6] together with the fact that the gradients on D_{ij} and $\overline{D}_{ij}$ are constant, we obtain:

$$\sum_{V_{ij}\in\mathscr{T}_h}\frac{\left(u_{ij}^{2k+1}-u_{ij}^{2k-1}\right)^2}{f(|\tilde{\nabla}u_{ij}^{2k-\frac{1}{2}}|)}\frac{h^2}{2\tau}+\sum_{D_{ij}\in\mathscr{D}_h}\frac{(u_{ij}^{2k+1}-u_{i+1,j}^{2k+1})^2-(u_{ij}^{2k+1}-u_{i+1,j}^{2k+1})(u_{ij}^{2k-1}-u_{i+1,j}^{2k-1})}{f(|\nabla u_{ij}^{2k-\frac{1}{2}}|)}$$
$$+\sum_{\overline{D}_{ij}\in\overline{\mathscr{D}}_h}\frac{(u_{ij}^{2k+1}-u_{i,j+1}^{2k-1})^2-(u_{ij}^{2k+1}-u_{i+1,j}^{2k+1})(u_{ij}^{2k-1}-u_{i+1,j}^{2k-1})}{f(|\nabla \overline{u}_{ij}^{2k-\frac{1}{2}}|)}=0$$

and

$$\sum_{\overline{V}_{ij}\in\overline{\mathscr{T}}_h}\frac{\left(\overline{u}_{ij}^{2k+2}-\overline{u}_{ij}^{2k}\right)^2}{f(|\tilde{\nabla}\overline{u}_{ij}^{2k-\frac{1}{2}}|)}\frac{h^2}{2\tau}+\sum_{D_{ij}\in\mathscr{D}_h}\frac{(\overline{u}_{ij}^{2k+2}-\overline{u}_{i,j-1}^{2k+2})^2-(\overline{u}_{ij}^{2k+2}-\overline{u}_{i,j-1}^{2k+2})(\overline{u}_{ij}^{2k}-\overline{u}_{i,j-1}^{2k})}{f(|\nabla u_{ij}^{2k-\frac{1}{2}}|)}$$
$$+\sum_{\overline{D}_{ij}\in\overline{\mathscr{D}}_h}\frac{(\overline{u}_{ij}^{2k}-\overline{u}_{i-1,j}^{2k})^2-(\overline{u}_{ij}^{2k}-\overline{u}_{i-1,j}^{2k})(\overline{u}_{ij}^{2k-2}-\overline{u}_{i-1,j}^{2k-2})}{f(|\nabla \overline{u}_{ij}^{2k-\frac{1}{2}}|)}=0.$$

We sum the both equations and rewrite the result into the form

$$\sum_{V_{ij}\in\mathscr{T}_h}\frac{\left(u_{ij}^{2k+1}-u_{ij}^{2k-1}\right)^2}{f(|\tilde{\nabla}u_{ij}^{2k-\frac{1}{2}}|)}\frac{h^2}{2\tau}+\sum_{\overline{V}_{ij}\in\overline{\mathscr{T}}_h}\frac{\left(\overline{u}_{ij}^{2k+2}-\overline{u}_{ij}^{2k}\right)^2}{f(|\tilde{\nabla}\overline{u}_{ij}^{2k-\frac{1}{2}}|)}\frac{h^2}{2\tau}+$$

$$\sum_{D_{ij}\in\mathscr{D}_h}\frac{(\nabla u_{ij}^{2k+\frac{3}{2}})^2-\nabla u_{ij}^{2k+\frac{3}{2}}\cdot\nabla u_{ij}^{2k-\frac{1}{2}}}{f(|\tilde{\nabla}u_{ij}^{2k-\frac{1}{2}}|)}h^2+\sum_{\overline{D}_{ij}\in\overline{\mathscr{D}}_h}\frac{(\nabla\overline{u}_{ij}^{2k+\frac{3}{2}})^2-\nabla\overline{u}_{ij}^{2k+\frac{3}{2}}\cdot\nabla\overline{u}_{ij}^{2k-\frac{1}{2}}}{f(|\tilde{\nabla}\overline{u}_{ij}^{2k-\frac{1}{2}}|)}h^2=0.$$

From (3) and (16) we have (and similarly for the terms with "bar")

$$F(|\nabla u_{ij}^{2k+\frac{3}{2}}|)-F(|\nabla u_{ij}^{2k-\frac{1}{2}}|)=\int_{|\nabla u_{ij}^{2k-\frac{1}{2}}|}^{|\nabla u_{ij}^{2k+\frac{3}{2}}|}\frac{z\,dz}{f(z)}.$$

Using the property (17) we obtain

$$F(|\nabla u_{ij}^{2k+\frac{3}{2}}|)-F(|\nabla u_{ij}^{2k-\frac{1}{2}}|)+\frac{1}{2b}(|\nabla u_{ij}^{2k+\frac{3}{2}}|-|\nabla u_{ij}^{2k-\frac{1}{2}}|)^2$$

$$\leq\frac{|\nabla u_{ij}^{2k+\frac{3}{2}}|}{f(|\nabla u_{ij}^{2k-\frac{1}{2}}|)}(|\nabla u_{ij}^{2k+\frac{3}{2}}|-|\nabla u_{ij}^{2k-\frac{1}{2}}|)\leq\frac{(\nabla u_{ij}^{2k+\frac{3}{2}})^2-\nabla u_{ij}^{2k+\frac{3}{2}}\cdot\nabla u_{ij}^{2k-\frac{1}{2}}}{f(|\nabla u_{ij}^{2k-\frac{1}{2}}|)}$$

and analogously for $|\nabla\overline{u}_{ij}^{2k+\frac{3}{2}}|$ etc. Using the above estimates we have

$$\sum_{V_{ij}\in\mathscr{T}_h}\frac{\left(u_{ij}^{2k+1}-u_{ij}^{2k-1}\right)^2}{f(|\tilde{\nabla}u_{ij}^{2k-\frac{1}{2}}|)}\frac{h^2}{2\tau}+\sum_{\overline{V}_{ij}\in\overline{\mathscr{T}}_h}\frac{\left(\overline{u}_{ij}^{2k+2}-\overline{u}_{ij}^{2k}\right)^2}{f(|\tilde{\nabla}\overline{u}_{ij}^{2k-\frac{1}{2}}|)}\frac{h^2}{2\tau}+$$

$$\frac{1}{2b}\sum_{D_{ij}\in\mathscr{D}_h}\left(|\nabla u_{ij}^{2k+\frac{3}{2}}|-|\nabla u_{ij}^{2k-\frac{1}{2}}|\right)^2h^2+\frac{1}{2b}\sum_{\overline{D}_{ij}\in\overline{\mathscr{D}}_h}\left(|\nabla\overline{u}_{ij}^{2k+\frac{3}{2}}|-|\nabla\overline{u}_{ij}^{2k-\frac{1}{2}}|\right)^2h^2+$$

$$\sum_{D_{ij}\in\mathscr{D}_h}(F(|\nabla u_{ij}^{2k+\frac{3}{2}}|)-F(|\nabla u_{ij}^{2k-\frac{1}{2}}|))h^2+\sum_{\overline{D}_{ij}\in\overline{\mathscr{D}}_h}(F(|\nabla\overline{u}_{ij}^{2k+\frac{3}{2}}|)-F(|\nabla\overline{u}_{ij}^{2k-\frac{1}{2}}|))h^2\leq 0.$$

Summing the last inequalities over $k=1,\ldots,m$ we obtain

$$\sum_{k=1}^{m}\left(\sum_{V_{ij}\in\mathscr{T}_h}\frac{\left(u_{ij}^{2k+1}-u_{ij}^{2k-1}\right)^2}{f(|\tilde{\nabla}u_{ij}^{2k-\frac{1}{2}}|)}\frac{h^2}{2\tau}+\sum_{\overline{V}_{ij}\in\overline{\mathscr{T}}_h}\frac{\left(\overline{u}_{ij}^{2k+2}-\overline{u}_{ij}^{2k}\right)^2}{f(|\tilde{\nabla}\overline{u}_{ij}^{2k-\frac{1}{2}}|)}\frac{h^2}{2\tau}\right)+$$

$$\frac{1}{2b}\sum_{k=1}^{m}\left(\sum_{D_{ij}\in\mathscr{D}_h}\left(|\nabla u_{ij}^{2k+\frac{3}{2}}|-|\nabla u_{ij}^{2k-\frac{1}{2}}|\right)^2+\sum_{\overline{D}_{ij}\in\overline{\mathscr{D}}_h}\left(|\nabla\overline{u}_{ij}^{2k+\frac{3}{2}}|-|\nabla\overline{u}_{ij}^{2k-\frac{1}{2}}|\right)^2\right)h^2+$$

$$\sum_{D_{ij}\in\mathscr{D}_h}F(|\nabla u_{ij}^{N_T-\frac{1}{2}}|)h^2+\sum_{\overline{D}_{ij}\in\overline{\mathscr{D}}_h}F(|\nabla\overline{u}_{ij}^{N_T-\frac{1}{2}}|)h^2\le$$

$$\sum_{D_{ij}\in\mathscr{D}_h}F(|\nabla u_{ij}^{0}|)h^2+\sum_{\overline{D}_{ij}\in\overline{\mathscr{D}}_h}F(|\nabla\overline{u}_{ij}^{0}|)h^2.$$

Using (16) and the properties of $u^0(x)$ we obtain the desired estimate (15) . □

3 Numerical Experiments

To study the Experimental Order of Convergence (EOC) for two variants of ADDFV scheme and to compare it with standard DDFV we use the solution of (1) presented in [10] of the following form $u(x, y, t) = \min\{0.5(x^2 + y^2 - 1) + t, 0\}$. The problem is solved on $\Omega = [-1, 25, 1.25]^2$ and $[0, T] = [0, 0.3125]$. The regularization parameter is chosen $\varepsilon = h^2$ where $h = 2.5/n$, $n = 10, 20, 40, \ldots, 320$.

We denote $e_{ij}^n = h^2\left(u_{ij}^n - u(t_n, x_{ij})\right)^2$ and $\overline{e}_{ij}^n = h^2\left(\overline{u}_{ij}^n - u(t_n, \overline{x}_{ij})\right)^2$ to define two discrete L_2 errors

$$E_2 = \left(\tau\sum_{n=1}^{N_T}\frac{1}{2}\left(\sum_{V_{ij}}e_{ij}^n+\sum_{\overline{V}_{ij}}\overline{e}_{ij}^n\right)\right)^{\frac{1}{2}},\quad E_2^B = \left(\tau\sum_{k=1}^{m}\left(\sum_{V_{ij}}e_{ij}^{2k-1}+\sum_{\overline{V}_{ij}}\overline{e}_{ij}^{2k}\right)\right)^{\frac{1}{2}}.$$

The error E_2 is used for DDFV scheme and for variant A, the error E_2^B is used for the Variant B. The results for the fixed definition of time step $\tau = h^2$ are presented in Table 1. Note that the solution contains flat regions and a singular circular curve with a gradient jump, therefore we can not expect second order accuracy as it can be observed also for other FV schemes in [5] and [9]. However the numerical schemes converge also in this singular case and the EOCs are close to 1.

From Table 1 one can see that the ADDFV schemes performs faster than the DDFV scheme for particular n, but the pay off is lower accuracy. In the current version the implemented ADDFV scheme does not perform better than the DDFV

Table 1 The comparison of errors, EOCs, and CPU times for the DDFV and ADDFV schemes

n	E_2	EOC	CPU	E_2	EOC	E_2^B	EOC	CPU
10	5.52e-02	–	1.50e-02	6.08e-02	–	6.78e-02	–	1.50e-02
20	3.17e-02	0.79	4.70e-02	3.99e-02	0.61	4.63e-02	0.55	4.60e-02
40	1.68e-02	0.91	6.55e-01	2.23e-02	0.83	2.24e-02	0.97	5.15e-01
80	8.69e-03	0.95	9.27e+00	1.18e-02	0.92	1.20e-02	0.96	7.30e+00
160	4.42e-03	0.97	1.53e+02	6.12e-03	0.96	6.15e-03	0.97	1.12e+02
320	2.23e-03	0.98	2.05e+03	3.11e-03	0.98	3.11e-03	0.99	1.47e+02

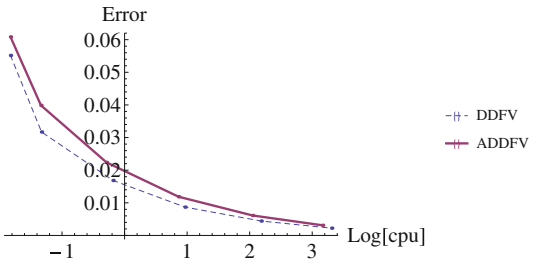


Fig. 3 The plot of E_2 over logarithm of CPU times for the DDFV and ADDFV scheme

scheme when comparing the errors with respect to the logarithm of CPU times as it can be seen in Fig. 3. This issue will be further investigated.

Acknowledgments This work was supported by grants APVV-0184-10 and VEGA 1/1137/12.

References

1. Andreianov, B., Boyer, F., Hubert, F.: Discrete duality finite volume schemes for Leray-Lions-type elliptic problems on general 2D meshes. Numer. Meth. Part. D. E. **23**(1), 145–195 (2007)
2. Deckelnick, K., Dziuk, G.: Error estimates for a semi-implicit fully discrete finite element scheme for the mean curvature flow of graphs. Interfaces Free Bound. **2**, 341–359 (2000)
3. Domelevo, K., Omnès, P.: A finite volume method for the laplace equation on almost arbitrary two-dimensional grids. M2AN. Math. Model. Numer. Anal. **39**(6), 1203–1249 (2005)
4. Evans, L.C., Spruck, J.: Motion of level sets by mean curvature I. J. Differ. Geom. **33**, 635–681 (1991)
5. Eymard, R., Handlovičová, A., Mikula, K.: Study of a finite volume scheme for the regularized mean curvature flow level set equation. IMA J. Numer. Anal. **31**(3), 813–846 (2011)
6. Handlovičová, A., Kotorová, D.: Convergence of a semi-implicit discrete duality finite volume scheme for the curvature driven level set equation in 2d. Kybernetika **49**(6), 829–854 (2013)
7. Handlovičová, A., Mikula, K., Sgallari, F.: Semi-implicit complementary volume scheme for solving level set like equations in image processing and curve evolution. Numer. Math. **93**, 675–695 (2003)
8. Hermeline, F.: A finite volume method for the approximation of diffusion operators on distorted meshes. J. Comput. Phys. **160**(2), 481–499 (2000)
9. Mikula, K., Sarti, A., Sgallari, F.: Co-volume method for Riemannian mean curvature flow in subjective surfaces multiscale segmentation. Comput. Visual. Sci. **9**(1), 23–31 (2006)

10. Oberman, A.: A convergent monotone difference scheme for motion of level sets by mean curvature. Numer. Math. **99**(2), 365–379 (2004)
11. Sethian, J.: Level set methods and fast marching methods: Evolving interfaces in computational geometry, fluid mechanics, computer vision, and material science. Cambridge University Press, New York (1999)

Convergence of the MAC Scheme for the Steady-State Incompressible Navier-Stokes Equations on Non-uniform Grids

R. Herbin, J.-C. Latché and K. Mallem

Abstract We prove in this paper the convergence of the Marker and cell (MAC) scheme for the discretization of the steady-state incompressible Navier-Stokes equations in primitive variables on non-uniform Cartesian grids, without any regularity assumption on the solution. *A priori* estimates on solutions to the scheme are proven; they yield the existence of discrete solutions and the compactness of sequences of solutions obtained with family of meshes the space step of which tends to zero. We then establish that the limit is a weak solution to the continuous problem.

1 Introduction

Let Ω be an open bounded domain of $\mathbb{R}^d$ with $d = 2$ or $d = 3$. We consider the steady-state incompressible Navier-Stokes equations, which read:

$$\mathrm{div}(\boldsymbol{u}) = 0 \quad \text{in } \Omega, \tag{1a}$$

$$-\Delta \boldsymbol{u} + (\boldsymbol{u} \cdot \nabla)\boldsymbol{u} + \nabla p = \boldsymbol{f} \quad \text{in } \Omega, \tag{1b}$$

$$\boldsymbol{u} = 0 \quad \text{on} \quad \partial\Omega, \tag{1c}$$

where $\boldsymbol{u}$ stands for the (vector-valued) velocity of the flow, p for the pressure and $\boldsymbol{f}$ is a given field of $L^2(\Omega)^d$. The weak formulation of the problem reads:

R. Herbin (✉) · K. Mallem
Aix-Marseille Université, Marseille, France
e-mail: raphaele.herbin@univ-amu.fr

K. Mallem
e-mail: khadidja.mallem@univ-amu.fr

J.-C. Latché
IRSN Cadarache, Saint-Paul-lès-Durance, France
e-mail: jean-claude.latche@irsn.fr

J. Fuhrmann et al. (eds.), *Finite Volumes for Complex Applications VII - Methods and Theoretical Aspects*, Springer Proceedings in Mathematics & Statistics 77,
DOI: 10.1007/978-3-319-05684-5_33,

$$\text{Find } (\boldsymbol{u}, p) \in H_0^1(\Omega)^d \times L_0^2(\Omega) \text{ such that, } \forall (\boldsymbol{v}, q) \in H_0^1(\Omega)^d \times L^2(\Omega),$$
$$\int_\Omega \nabla \boldsymbol{u} : \nabla \boldsymbol{v} \, \mathrm{d}\boldsymbol{x} + \int_\Omega ((\boldsymbol{u} \cdot \nabla)\boldsymbol{u}) \cdot \boldsymbol{v} \, \mathrm{d}\boldsymbol{x} - \int_\Omega p \, \mathrm{div}(\boldsymbol{v}) \, \mathrm{d}\boldsymbol{x} = \int_\Omega \boldsymbol{f} \cdot \boldsymbol{v} \, \mathrm{d}\boldsymbol{x}, \quad (2)$$
$$\int_\Omega q \, \mathrm{div}(\boldsymbol{u}) \, \mathrm{d}\boldsymbol{x} = 0,$$

where $L_0^2(\Omega)$ stands for the subspace of $L^2(\Omega)$ of zero mean-valued functions. The aim of this paper is to show, under minimal regularity assumptions on the solution, that sequences of approximate solutions obtained by the discretization of Problem (1) by the Marker-And-Cell (MAC) scheme converge to a solution of (2) as the mesh size tends to 0.

The Marker-And-Cell (MAC) scheme, introduced in the middle of the sixties [5], is one of the most popular methods [8, 9] for the approximation of the Navier-Stokes equations in the engineering framework, because of its simplicity, its efficiency and its remarkable mathematical properties. In the case of uniform meshes, finite difference techniques allow to obtain error estimates [7] for the vorticity-pressure formulation of (1) with some regularity conditions on the exact solution (H^2 regularity for the pressure) that are stronger than the natural conditions. Here, we give a convergence result with respect to the discretization of the Navier-Stokes equations in primitive variables on a non-uniform rectangular mesh, thanks to a finite volume expression of the scheme (see [1] for a first result in this direction in the case of the Stokes equations), and without regularity assumptions on the solutions. An essential feature of the studied scheme is that the (discrete) kinetic energy remains controlled. In particular, the velocity convection operator is approximated so as to be compatible with a discrete continuity equation on the dual cells; this discretization coincides with the usual discretization on uniform meshes [8], contrary to the scheme of [2]. Velocity and pressure estimates are thus obtained, which lead to the compactness of sequences of approximate solutions. We then show that the prospective limit is a weak solution of the Navier-Stokes equations. For short, we focus here on the analysis of the stability and consistency of the velocity convection operator, which is non-standard, while the study of (linear) diffusion, gradient and divergence operators is more classical. We can then conclude that the approximate solutions obtained with the MAC scheme converge (up to a subsequence since no uniqueness result is known for the continuous problem) to a weak solution of the Navier-Stokes equations. The present work is the first step in a project which consists in the mathematical study of the MAC scheme on non-uniform grids for the steady and time-dependent Navier-Stokes equations in primitive variables, and its extension to the variable density Navier-Stokes equations. Because of space restrictions, we only sketch the proofs here and refer to [6] for full details.

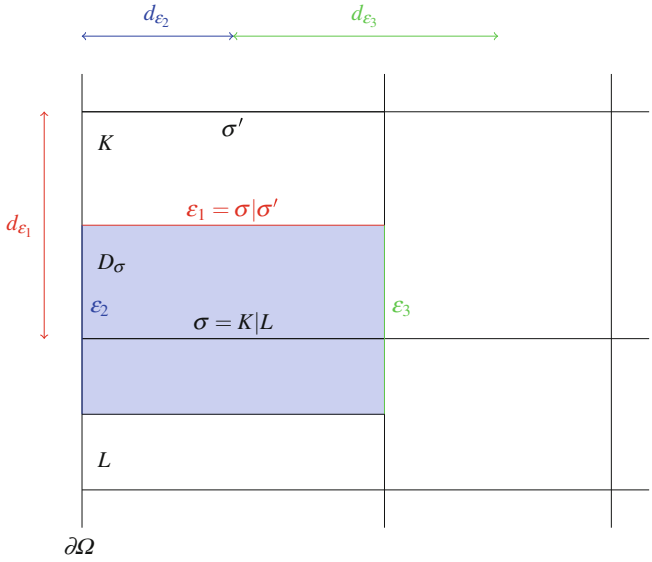


Fig. 1 Notations for control volumes and dual cells (for the second component of the velocity)

2 The MAC Scheme

We assume that the domain Ω is a union of rectangles ($d = 2$) or orthogonal parallelepipeds ($d = 3$). Let us introduce the MAC grids, namely

- the pressure (or primal) grid denoted by $\mathcal{M}$, which consists of a union of possibly non uniform rectangles; a generic cell of this grid will be denoted by K, and a generic face of such a cell by $\sigma \in \mathcal{E}$, where $\mathcal{E}$ denotes the set of all faces of the mesh. The set of faces that are orthogonal to the ith unit vector $\boldsymbol{e}_i$ of the canonical basis of $\mathbb{R}^d$ is denoted by $\mathcal{E}^{(i)}$, for $i = 1, \ldots, d$.
- the velocity (or dual) grids denoted by $\mathcal{M}^{(i)}$: one grid for each component $u^{(i)}$, $i = 1, \ldots, d$. A cell D_σ of the mesh $\mathcal{M}^{(i)}$ is associated to a face $\sigma \in \mathcal{E}^{(i)}$. For an internal face $\sigma = K|L$, it is defined as the union of two rectangles $D_{K,\sigma}$ and $D_{L,\sigma}$, where $D_{K,\sigma}$ (resp. $D_{L,\sigma}$) is the half-part of K (resp. L) adjacent to σ (see Fig. 1 for the 2D case); for an external face σ adjacent to the cell K, $D_\sigma = D_{K,\sigma}$. A dual face separating two dual cells D_σ and $D_{\sigma'}$ is denoted by $\varepsilon = \sigma|\sigma'$. To any dual face ε, we associate a distance d_ε as sketched on Fig. 1.

Hereafter, $|\cdot|$ stands indifferently for the d- or $(d-1)$-dimensional measure of a subset of $\mathbb{R}^d$ or $\mathbb{R}^{d-1}$ respectively.

We define the discrete pressure space $L_\mathcal{M}$ which stands for piecewise constant functions over each of the grid cells K of $\mathcal{M}$, and $H^{(i)}_\mathcal{M}$ which stands for the piecewise constant functions over each of the grid cells D_σ of $\mathcal{M}^{(i)}$, with $\sigma \in \mathcal{E}^{(i)}$. As in the continuous case, the Dirichlet boundary conditions are (partly) incorporated in

the definition of the velocity spaces, and, to this purpose, we introduce $H^{(i)}_{\mathcal{M},0} \subset H^{(i)}_{\mathcal{M}}$, $i = 1, \ldots, d$, defined as follows:

$$H^{(i)}_{\mathcal{M},0} = \left\{ u \in H^{(i)}_{\mathcal{M}},\ u(x) = 0\ \forall x \in D_\sigma,\ \sigma \in \mathcal{E}^{(i)}_{\rm ext} \right\},\ i = 1, \ldots, d.$$

where $\mathcal{E}^{(i)}_{\rm ext}$ is the subset of $\mathcal{E}^{(i)}$ of faces that lie on the boundary $\partial\Omega$ of the domain. We then set $\mathbf{H}_{\mathcal{M},0} = \prod_{i=1}^{d} H^{(i)}_{\mathcal{M},0}$.

Discrete divergence operator—We are now in position to define the discrete divergence operator $\mathrm{div}_{\mathcal{M}}$ from $\mathbf{H}_{\mathcal{M},0}$ to $L_{\mathcal{M}}$ by

$$\mathrm{div}_{\mathcal{M}} \boldsymbol{u}(\boldsymbol{x}) = \frac{1}{|K|} \sum_{\sigma \in \mathcal{E}(K)} F_{K,\sigma}(\boldsymbol{u}),\ \forall \boldsymbol{x} \in K,$$

where $F_{K,\sigma}(\boldsymbol{u})$ is the mass flux through a face σ of the set $\mathcal{E}(K)$ of faces of K, which reads, for $\sigma \in \mathcal{E}^{(i)}$:

$$F_{K,\sigma}(\boldsymbol{u}) = |\sigma|\, u_\sigma\, \boldsymbol{e}_i \cdot \boldsymbol{n}_{K,\sigma},$$

where $\boldsymbol{n}_{K,\sigma}$ denotes the unit normal vector to σ outward K. Note that we have the usual finite volume property of local conservativity of the flux through an interface $\sigma = K|L$ between the cells $K, L \in \mathcal{M}$, *i.e.* $F_{K,\sigma}(\boldsymbol{u}) = -F_{L,\sigma}(\boldsymbol{u})$.

Discrete gradient—The gradient in the discrete momentum balance equation is built as the dual operator of the discrete divergence, and reads:

$$\nabla_{\mathcal{M}} : \left| \begin{array}{l} L_{\mathcal{M}} \longrightarrow \mathbf{H}_{\mathcal{M},0} \\ p \longmapsto \nabla_{\mathcal{M}} p(\boldsymbol{x}) = (\eth_1 p, \ldots, \eth_d p)^t, \end{array} \right.$$

where $\eth_i p \in H^{(i)}_{\mathcal{M},0}$ is the discrete derivative of p in the i-th direction, defined by:

$$\eth_i p(\boldsymbol{x}) = \frac{|\sigma|}{|D_\sigma|} (p_L - p_K)\, \boldsymbol{n}_{K,\sigma} \cdot \boldsymbol{e}_i,\ \forall \boldsymbol{x} \in D_\sigma,\ \text{for } \sigma = K|L \in \mathcal{E}^{(i)}_{\rm int},\ i = 1, \ldots, d.$$

Note that the definition of the operator is complete, since the functions of $\mathbf{H}_{\mathcal{M},0}$ vanish on the dual cells associated to external faces.

Discrete Laplace operator—For $i = 1 \ldots, d$, we classically define the ith component $-\Delta^{(i)}_{\mathcal{M}}$ of the discrete Laplace operator from $H^{(i)}_{\mathcal{M},0}$ to $H^{(i)}_{\mathcal{M},0}$ by:

$$-\Delta_{\mathcal{M}} u^{(i)}(\boldsymbol{x}) = \frac{1}{|D_\sigma|} \sum_{\varepsilon \in \mathcal{E}(D_\sigma)} \frac{|\varepsilon|}{d_\varepsilon} [u^{(i)}]_{\sigma,\varepsilon},\ \forall \boldsymbol{x} \in D_\sigma,\ \ \text{for } \sigma \in \mathcal{E}^{(i)}.$$

where the sum in the right-hand-side is over the set $\mathcal{E}(D_\sigma)$ of (dual) faces of the dual cell D_σ, and where $[u^{(i)}]_{\sigma,\varepsilon} = u^{(i)}_\sigma - u^{(i)}_{\sigma'}$ if ε is an internal dual face

separating D_σ and $D_{\sigma'}$ and $[u^{(i)}]_{\sigma,\varepsilon} = u_\sigma^{(i)}$ if ε is included in the boundary $\partial\Omega$. Then the discrete Laplace operator $-\Delta_\mathcal{M}$, $\mathbf{H}_{\mathcal{M},0} \longrightarrow \mathbf{H}_{\mathcal{M},0}$, is given by $-\Delta_\mathcal{M}\boldsymbol{u} = (-\Delta_\mathcal{M}^{(1)}u^{(1)}, \ldots, -\Delta_\mathcal{M}^{(d)}u^{(d)})^t$.

Discrete convection operator—Let us consider the momentum equation (1b) for the ith component of the velocity, based on its associated dual mesh $\mathcal{M}^{(i)}$ and integrate it on a cell D_σ, $\sigma \in \mathcal{E}^{(i)}$. By the Stokes formula we then need to discretize $\sum_{\varepsilon\subset\partial D_\sigma}\int_\varepsilon u^{(i)}\boldsymbol{u}\cdot\boldsymbol{n}_{\varepsilon,\sigma}$, where $\boldsymbol{n}_{\varepsilon,\sigma}$ denotes the unit normal vector to ε outward D_σ. For $\varepsilon = \sigma|\sigma'$, the convection flux $\int_\varepsilon u^{(i)}\boldsymbol{u}\cdot\boldsymbol{n}_{\varepsilon,\sigma}$ is approximated by $|\varepsilon|\ F_{\sigma,\varepsilon}(\boldsymbol{u})u_\varepsilon^{(i)}$, $u_\varepsilon^{(i)} = (u_\sigma^{(i)} + u_{\sigma'}^{(i)})/2$, where $F_{\sigma,\varepsilon}(\boldsymbol{u})$ is the numerical mass flux through ε outward D_σ which we now define. We distinguish two cases:

- First case—The vector $\boldsymbol{e}_i$ is normal to ε , and ε is included in a primal cell K. Then the mass flux through $\varepsilon = D_\sigma|D_{\widetilde{\sigma}}$ is given by:

$$F_{\sigma,\varepsilon}(\boldsymbol{u}) = \frac{1}{2}\,(-F_{K,\sigma}(\boldsymbol{u}) + F_{K,\widetilde{\sigma}}(\boldsymbol{u})).$$

- Second case—The vector $\boldsymbol{e}_i$ is tangent to ε, and ε is the union of the halves of two primal faces σ' and σ'' such that $\sigma = K|L$ with $\sigma' \in \mathcal{E}(K)$ and $\sigma'' \in \mathcal{E}(L)$. Then we write $\varepsilon = \frac{\sigma'\sigma''}{K|L}$. The mass flux through ε is then given by:

$$F_{\sigma,\varepsilon}(\boldsymbol{u}) = \frac{1}{2}\,(F_{K,\sigma'}(\boldsymbol{u}) + F_{L,\sigma''}(\boldsymbol{u})).$$

  Note that, with this definition, $F_{\sigma,\varepsilon}(\boldsymbol{u}) = 0$ if $\varepsilon \subset \partial\Omega$, which is consistent with the boundary conditions (1c).

We can now define the operator $C_\mathcal{M}^{(i)}$ from $H_{\mathcal{M},0}^{(i)}$ to $H_{\mathcal{M},0}^{(i)}$ by

$$C_\mathcal{M}^{(i)}u^{(i)}(\boldsymbol{x}) = \frac{1}{|D_\sigma|}\sum_{\substack{\varepsilon\in\mathcal{E}(D_\sigma)\\ \varepsilon=\sigma|\sigma'}} |\varepsilon|\ F_{\sigma,\varepsilon}(\boldsymbol{u})\frac{u_\sigma^{(i)} + u_{\sigma'}^{(i)}}{2},\quad \forall \boldsymbol{x}\in D_\sigma,\ \text{for}\ \sigma\in\mathcal{E}_{\text{int}}^{(i)}.$$

Then the discrete convection operator $C_\mathcal{M}$ from $\mathbf{H}_{\mathcal{M},0}$ to $\mathbf{H}_{\mathcal{M},0}$ is defined by $C_\mathcal{M}\boldsymbol{u} = (C_\mathcal{M}^{(1)}u^{(1)}, \ldots, C_\mathcal{M}^{(d)}u^{(d)})^t$.

The scheme—With these notations, the discrete scheme reads:

$$\boldsymbol{u}\in\mathbf{H}_{\mathcal{M},0},\ p\in L_\mathcal{M},\ \int_\Omega p\,\mathrm{d}\boldsymbol{x} = 0, \tag{3a}$$

$$-\Delta_\mathcal{M}\boldsymbol{u} + C_\mathcal{M}\boldsymbol{u} + \nabla_\mathcal{M}p = \boldsymbol{f}_\mathcal{M}, \tag{3b}$$

$$\mathrm{div}_\mathcal{M}\boldsymbol{u} = 0, \tag{3c}$$

where $\boldsymbol{f}_\mathcal{M}$ stands for the projection of $\boldsymbol{f}$ onto $\mathbf{H}_{\mathcal{M},0}$ obtained by taking the mean value over each cell.

3 Variational form of the Scheme and Stability

We first recall the definition of the discrete H^1 inner product [3]. To this purpose, we multiply the discrete Laplace operator by $\boldsymbol{u}$ and integrate over the computational domain. A simple reordering of the sums, which may be seen as a discrete integration by parts, yields:

$$\forall (\boldsymbol{u}, \boldsymbol{v}) \in \mathbf{H}_{\mathcal{M},0}{}^2, \qquad \int_\Omega -\Delta_{\mathcal{M}} \boldsymbol{u} \cdot \boldsymbol{v} \, \mathrm{d}\boldsymbol{x} = [\boldsymbol{u}, \boldsymbol{v}]_{1,\mathcal{M},0} = \sum_{i=1}^{d} [u^{(i)}, v^{(i)}]_{1,\mathcal{M}^{(i)},0},$$

with:

$$[u^{(i)}, v^{(i)}]_{1,\mathcal{M}^{(i)},0} = \sum_{\varepsilon = \sigma | \sigma' \in \mathcal{E}^{(i)}} \frac{|\varepsilon|}{d_\varepsilon} \, [u^{(i)}]_{\sigma,\varepsilon} \, [v^{(i)}]_{\sigma,\varepsilon},$$

where $\mathcal{E}^{(i)}$ denotes the set of faces of the dual mesh $\mathcal{M}^{(i)}$, $i = 1, \ldots, d$. We may also define the discrete H^1-norms:

$$\|u^{(i)}\|^2_{1,\mathcal{M},0} = [u^{(i)}, u^{(i)}]_{1,\mathcal{M}^{(i)},0}, \text{ for } i = 1, \ldots, d, \text{ and } \|\boldsymbol{u}\|^2_{1,\mathcal{M},0} = [\boldsymbol{u}, \boldsymbol{u}]_{1,\mathcal{M},0}.$$

Let us now define the weak form of the nonlinear convection operator, which we denote by $\mathcal{C}_{\mathcal{M}}$:

$$\forall (\boldsymbol{u}, \boldsymbol{v}) \in \mathbf{H}_{\mathcal{M},0}{}^2, \quad \mathcal{C}_{\mathcal{M}}(\boldsymbol{u}, \boldsymbol{v}) = \sum_{i=1}^{d} \mathcal{C}^{(i)}_{\mathcal{M}}(u^{(i)}, v^{(i)}) = \sum_{i=1}^{d} \int_\Omega C^{(i)}_{\mathcal{M}}(u^{(i)}) \, v^{(i)} \, \mathrm{d}\boldsymbol{x}.$$

Then the weak formulation of the scheme reads:

$$\text{Find } (\boldsymbol{u}, p) \in \mathbf{H}_{\mathcal{M},0} \times L_{\mathcal{M}} \text{ such that } \int_\Omega p \, \mathrm{d}\boldsymbol{x} = 0 \text{ and, } \forall (\boldsymbol{v}, q) \in \mathbf{H}_{\mathcal{M},0} \times L_{\mathcal{M}},$$

$$[\boldsymbol{u}, \boldsymbol{v}]_{1,\mathcal{M},0} + \mathcal{C}_{\mathcal{M}}(\boldsymbol{u}, \boldsymbol{v}) - \int_\Omega p \, \mathrm{div}_{\mathcal{M}}(\boldsymbol{v}) \, \mathrm{d}\boldsymbol{x} = \int_\Omega \boldsymbol{f}_{\mathcal{M}} \cdot \boldsymbol{v} \, \mathrm{d}\boldsymbol{x}, \tag{4a}$$

$$\int_\Omega \mathrm{div}_{\mathcal{M}}(\boldsymbol{u}) \, q \, \mathrm{d}\boldsymbol{x} = 0. \tag{4b}$$

In order to prove the convergence of the scheme, we introduce an alternate convection form $\widetilde{\mathcal{C}}$, defined on the pressure grid and easier to manipulate in the proofs; it is also defined component by component, *i.e.* it may be written as

$$\widetilde{\mathcal{C}}_{\mathcal{M}}(\boldsymbol{u}, \boldsymbol{v}) = \sum_{i=1}^{d} \widetilde{\mathcal{C}}^{(i)}_{\mathcal{M}}(u^{(i)}, v^{(i)}), \forall (\boldsymbol{u}, \boldsymbol{v}) \in \mathbf{H}_{\mathcal{M},0}{}^2.$$

The following lemma defines $\widetilde{\mathcal{C}}^{(i)}$ and provides a bound for $\widetilde{\mathcal{C}}^{(i)} - \mathcal{C}^{(i)}$.

Lemma 1 (Alternate convection term) *Let $\boldsymbol{u} \in \mathbf{H}_{\mathcal{M},0}$, let $i \in \{1, \ldots, d\}$ and let $u,\ v \in H^{(i)}_{\mathcal{M},0}$. For $\sigma \in \mathcal{E}$, let us define $\widehat{u}_\sigma$ by:*

$$\widehat{u}_\sigma = u_\sigma \text{ if } \sigma \in \mathcal{E}^{(i)}, \quad \widehat{u}_\sigma = \frac{1}{\mathrm{card}(\mathcal{N}_\sigma)} \sum_{\sigma' \in \mathcal{N}_\sigma} u_{\sigma'} \text{ otherwise},$$

where, for any $\sigma \in \mathcal{E} \setminus \mathcal{E}^{(i)}$, $\mathcal{N}_\sigma = \{\sigma' \in \mathcal{E}^{(i)}, \overline{D_\sigma} \cap \sigma' \neq \emptyset\}$. For $K \in \mathcal{M}$, we denote by v_K the quantity $v_K = \frac{1}{2} \sum_{\sigma \in \mathcal{E}^{(i)}(K)} v_\sigma$. Let $\widetilde{\mathcal{C}}^{(i)}_{\mathcal{M}}$ be the nonlinear form defined by:

$$\widetilde{\mathcal{C}}^{(i)}_{\mathcal{M}}(u, v) = \sum_{K \in \mathcal{M}} v_K \sum_{\sigma \in \mathcal{E}(K)} F_{K,\sigma}(\boldsymbol{u}) \widehat{u}_\sigma,$$

and let $\mathcal{R}^{(i)}(u, v) = \mathcal{C}^{(i)}_{\mathcal{M}}(u, v) - \widetilde{\mathcal{C}}^{(i)}_{\mathcal{M}}(u, v)$. Then there exists $C \geq 0$ depending only on the regularity $\eta_{\mathcal{M}}$ of the mesh defined by

$$\eta_{\mathcal{M}} = \max \left\{ \frac{|K|}{|L|}, (K, L) \in \mathcal{M}^2; K|L \in \mathcal{E}_{\mathrm{int}} \right\}, \tag{5}$$

such that :

$$|\mathcal{R}^{(i)}(u, v)| \leq C\ h^\alpha\ \|\boldsymbol{u}\|_{1,\mathcal{M},0}\ \|u\|_{1,\mathcal{M}^{(i)},0}\ \|v\|_{1,\mathcal{M}^{(i)},0},$$

with $\alpha < 1$ if $d = 2$ and $\alpha = 1/2$ if $d = 3$.

Lemma 2 (Estimate on the convection term) *There exists $C > 0$, depending only on the regularity $\eta_{\mathcal{M}}$ of the mesh (5), such that:*

$$\forall (\boldsymbol{u}, \boldsymbol{v}) \in \mathbf{H}_{\mathcal{M},0}{}^2, \qquad |\mathcal{C}_{\mathcal{M}}(\boldsymbol{u}, \boldsymbol{v})| \leq C\ \|\boldsymbol{u}\|^2_{1,\mathcal{M},0}\ \|\boldsymbol{v}\|_{1,\mathcal{M},0}. \tag{6}$$

In addition:

$$\forall \boldsymbol{u} \in \mathbf{H}_{\mathcal{M},0}, \qquad \mathcal{C}_{\mathcal{M}}(\boldsymbol{u}, \boldsymbol{u}) = 0. \tag{7}$$

Essential arguments of the proof—The estimate (6) follows from Lemma 1 and a bound on $\widetilde{\mathcal{C}}_{\mathcal{M}}$ obtained with some simple algebra. The relation (7) relies on the fact that

$$\sum_{\varepsilon \in \mathcal{E}(D_\sigma)} F_{\sigma,\varepsilon}(\boldsymbol{u}) = 0.$$

thanks to the definition of the dual mass fluxes $F_{\sigma,\varepsilon}(\boldsymbol{u})$. □

The following stability result is a consequence of the above lemma, together with the duality of the discrete gradient and divergence operators, and the fact that

the MAC discretization satisfies the so-called *inf-sup* condition (see e.g. [4]). The existence of a solution then follows by a topological degree argument.

Proposition 1 (Existence and estimates) *There exists a solution to* (4)*, and there exists* $C > 0$ *depending only on the regularity* $\eta_{\mathcal{M}}$ *of the mesh, such that any solution of* (4) *satisfies the following stability estimate:*

$$\|\boldsymbol{u}\|_{1,\mathcal{M},0} + \|p\|_{L^2(\Omega)} \leq C \, \|\boldsymbol{f}\|_{(L^2(\Omega))^d}. \tag{8}$$

4 Convergence Analysis

The convergence of the scheme is stated in the following theorem.

Theorem 1 (Convergence of the scheme) *Let* $(\mathcal{M}_n)_{n\in\mathbb{N}}$ *be a sequence of meshes such that* $\max_{K\in\mathcal{M}_n} diam(K) \to 0$ *as* $n \to +\infty$*; assume that there exists* $\eta > 0$ *such that* $\eta_{\mathcal{M}_n} \leq \eta$ *for any* $n \in \mathbb{N}$ *(with* $\eta_{\mathcal{M}_n}$ *defined by* (5)*). Let* $(\boldsymbol{u}_n, p_n)$ *be a solution to* (4) *for* $\mathcal{M} = \mathcal{M}_n$*. Then there exists* $\bar{\boldsymbol{u}} \in H^1_0(\Omega)^d$ *and* $\bar{p} \in L^2(\Omega)$ *such that, up to a subsequence:*

- *the sequence* $(u_n)_{n\in\mathbb{N}}$ *converges to* $\bar{\boldsymbol{u}}$ *in* $L^2(\Omega)^d$*,*
- *the sequence* $(p_n)_{n\in\mathbb{N}}$ *weakly converges to* $\bar{p}$ *in* $L^2(\Omega)$*,*
- $(\bar{\boldsymbol{u}}, \bar{p})$ *is a solution to the weak formulation* (2)*.*

Main steps of the proof—The existence of a limit for the velocity in $L^2(\Omega)^d$ and the pressure in $L^2(\Omega)$ and the convergence of the sequence of discrete solutions follow from a compactness argument thanks to the estimates of Proposition 1. We then obtain that $\bar{\boldsymbol{u}}$ belongs to $H^1_0(\Omega)^d$ thanks to the estimate (8).

Finally, it remains to pass to the limit in the scheme, *i.e.*, in other words, to prove its (weak) consistency. For the diffusion and divergence operators, the proof is rather standard [3]. The consistency of the discrete gradient is readily obtained thanks to the fact that this operator is dual with the divergence. For the convection operator, essential difficulties are solved by switching from the nonlinear form $\mathcal{C}$ to $\widetilde{\mathcal{C}}$, thanks to the control on the error between $\mathcal{C}$ and $\widetilde{\mathcal{C}}$ given in Lemma 1. □

References

1. Blanc, P.: Error estimate for a finite volume scheme on a MAC mesh for the Stokes problem. In: Finite Volumes for Complex Applications II, pp. 117–124. Hermes Science Publishing, Paris (1999)
2. Chénier, E., Eymard R. nd Gallouët, T., Herbin, R.: An extension of the MAC scheme to locally refined meshes: convergence analysis for the full tensor time-dependent Navier-Stokes equations. Calcolo, to appear (2014)

3. Eymard, R., Gallouët, T., Herbin, R.: Finite volume methods. In: Ciarlet, P.G., Lions, J.L. (eds.) Techniques of Scientific Computing, Part III, Handbook of Numerical Analysis, VII, pp. 713–1020. North-Holland, Amsterdam (2000)
4. Gallouët, T., Herbin, R., Latché, J.: $W^{1,q}$ stability of the Fortin operator for the MAC scheme. Calcolo **69**, 63–71 (2012). See also http://hal.archives-ouvertes.fr/
5. Harlow, F., Welch, J.: Numerical calculation of time-dependent viscous incompressible flow of fluid with a free surface. Phys. Fluids **8**, 2182–2189 (1965)
6. Herbin, R., Latché, J., Mallem, K.: Numerical analysis of the MAC scheme for the Navier-Stokes equations in primitive variables. (in preparation)
7. Nicolaïdes, R., Wu, X.: Analysis and convergence of the mac scheme ii, Navier-Stokes equations. Math. Comp. **65**, 29–44 (1996)
8. Patankar, S.: Numerical heat transfer and fluid flow. Series in Computational Methods in Mechanics and Thermal Sciences, vol. XIII. Hemisphere Publishing Corporation, Washington (1980)
9. Wesseling, P.: Principles of Computational Fluid Dynamics. Springer, Berlin (2001)

Stochastic Modeling for Heterogeneous Two-Phase Flow

M. Köppel, I. Kröker and C. Rohde

Abstract The simulation of multiphase flow problems in porous media often requires techniques for uncertainty quantification to represent parameter values that are not known exactly. The use of the stochastic Galerkin approach becomes very complex in view of the highly nonlinear flow equations. On the other hand collocation-like methods suffer from low convergence rates. To overcome these difficulties we present a hybrid stochastic Galerkin finite volume method (HSG-FV) that is in particular well-suited for parallel computations. The new approach is applied to specific two-phase flow problems including the example of a porous medium with a spatially random change in mobility. We emphasize in particular the issue of parallel scalability of the overall method.

1 Introduction

We consider the influence of stochastic effects on a two-phase flow model, that governs the infiltration of a wetting fluid into a porous medium which is initially filled by a nonwetting fluid. Let us assume that both fluids are immiscible and incompressible, and let us neglect gravitational forces. The fractional flow formulation of the capillarity-free case for some domain $D \subset \mathbb{R}^2$ and time $T > 0$ leads to the following problem [5]:

M. Köppel (✉) · I. Kröker · C. Rohde
IANS, Universität Stuttgart, Pfaffenwaldring 57, 70569 Stuttgart, Germany
e-mail: markus.koeppel@mathematik.uni-stuttgart.de

I. Kröker
e-mail: ilja.kroeker@mathematik.uni-stuttgart.de

C. Rohde
e-mail: christian.rohde@mathematik.uni-stuttgart.de

J. Fuhrmann et al. (eds.), *Finite Volumes for Complex Applications VII - Methods and Theoretical Aspects*, Springer Proceedings in Mathematics & Statistics 77,
DOI: 10.1007/978-3-319-05684-5_34,

$$\mathbf{v} = -\mathbf{K}\lambda(S) \bigtriangledown p \quad \text{and} \quad \operatorname{div}(\mathbf{v}) = q \qquad \text{in } D \times (0,T), \tag{1}$$

$$\phi S_t + \operatorname{div}(\mathbf{v} f(\mathbf{x}, S)) - q = 0 \qquad \text{in } D \times (0,T). \tag{2}$$

The unknowns are the saturation of the wetting fluid $S = S(\mathbf{x}, t) \in [0,1]$, the global pressure $p = p(\mathbf{x}, t) \in \mathbb{R}$, and the total velocity field $\mathbf{v} = \mathbf{v}(\mathbf{x}, t) \in \mathbb{R}^2$. The total mobility $\lambda = \lambda(S)$ and the fractional flow function $f = f(\mathbf{x}, S)$ are given nonlinear functions of the saturation and additionally of space for the flux. Furthermore $\mathbf{K} = \mathbf{K}(\mathbf{x})$ stands for the intrinsic permeability, $\phi = \phi(\mathbf{x})$ for the porosity, and $q = q(\mathbf{x}, t)$ for a source or sink. Appropriate initial and boundary conditions have to be added.

Uncertainty can effect solutions of (1), (2) through e.g. given parameter functions, initial and boundary data. In this case the unknowns depend also on corresponding random variables. Let us first assume that the velocity field $\mathbf{v}$ is given and it remains to solve the hyperbolic transport equation for the saturation. Under generic conditions a representation in the form of a polynomial chaos expansion (PCE) exists. Restriction of a (stochastically) weak formulation to a finite number of modes leads to the stochastic Galerkin method. Combined with a finite volume discretization in space the PCE approach yields a coupled deterministic system to be solved. The degree of coupling increases with the non-linearity of the considered equations and with the order of polynomial expansion. This fact increases the computational effort and significantly reduces the scalability in parallelisation. We have suggested a hybrid stochastic Galerkin finite volume method (HSG-FV) in [2], that extends the methods presented in [8, 11], for general transport equations and will develop it here for the two-phase problem. Together with a review on the stochastic setting the new method is formulated in Sect. 2. We stress that the HSG-FV relies on an adaptive combination of PCE with a multi-element decomposition of the stochastic domain. It leads to a deterministic system that is significantly weaker coupled than the pure PCE approach. Therefore, the HSG-FV method allows for more efficient parallelization. In Sect. 3 we apply the HSG-FV to the two-phase flow problem (1), (2) with a nonlinear continuous flux function, present the finite volume method and numerical examples. Moreover the computational effort of the HSG-FV method is discussed at the end of the section. At last we briefly present the application of the HSG-FV method to the two-phase flow problem in a heterogeneous porous medium with randomly disturbed discontinuous flux function in Sect. 4.

2 Hybrid Stochastic Galerkin Representation

Polynomial Chaos Let $\theta = \theta(\omega)$ be a random variable on the probability space $(\Omega, \mathscr{F}, \mathbb{P})$, which satisfies $\theta \in L^2(\Omega)$. We assume that the distribution of θ is known and the probability density function (PDF) ρ is given. In this case the expectation of the random variable θ can be computed by $\mathbb{E}[\theta] := \int_\Omega \theta(\omega)\, \mathrm{d}\mathbb{P}(\omega) = \int \theta\, \mathrm{d}\rho(\theta)$. Then there exists a family $\{\phi_p(\theta)\}_{p\in\mathbb{N}_0}$ of $L^2(\Omega)$-orthonormal polynomials with respect to the PDF ρ. This means that $\{\phi_p(\theta)\}_{p\in\mathbb{N}_0}$ satisfies

$$\langle \phi_p(\theta),\, \phi_q(\theta)\rangle_{L^2(\Omega)} := \int_{\mathsf{I}} \phi_p(\theta)\phi_q(\theta)\,\mathrm{d}\rho(\theta) = \delta_{pq} \quad \text{for } p, q \in \mathbb{N}_0.$$

Here δ_{pq} denotes the Kronecker delta and I is the support of ϕ_p, for $p \in \mathbb{N}_0$. The choice of the polynomial basis depends on the PDF ρ. For example the Hermite polynomials could be used for the stochastic discretization of the Gauss distributed random variables, and Legendre polynomials allow the discretization of uniformly distributed random variables. Let $w = w(\mathbf{x}, t, \theta(\omega))$, $(\mathbf{x}, t) \in D \times [0, T]$, $\omega \in \Omega$ be a second order random field. Then w can be represented by the infinite series

$$w(\mathbf{x}, t, \theta(\omega)) = \sum\nolimits_{p=0}^{\infty} w^p(\mathbf{x}, t)\phi_p(\theta(\omega)), \quad (\mathbf{x}, t, \omega) \in D \times [0, T] \times \Omega.$$

The coefficients $w^p = w^p(\mathbf{x}, t)$, $(\mathbf{x}, t) \in D \times [0, T]$ are defined by $w^p := \langle w,\, \phi_p\rangle_{L^2(\Omega)}$ for $p \in \mathbb{N}_0$. The expectation of the random field w is given by w^0, and the variance is given by the series $\sum_{p=1}^{\infty}(w^p)^2$. The truncation up to polynomial order $N_\mathrm{o} \in \mathbb{N}$ yields a finite sum

$$\Pi^{N_\mathrm{o}}[w](\mathbf{x}, t, \theta(\omega)) := \sum\nolimits_{p=0}^{N_\mathrm{o}} w^p(\mathbf{x}, t)\phi_p(\theta(\omega)), \quad (\mathbf{x}, t, \omega) \in D \times [0, T] \times \Omega. \tag{3}$$

The Cameron-Martin theorem [3] shows the convergence of (3). For more explanations we refer to [4, 12].

Extension to the Hybrid stochastic Galerkin discretization For the sake of brevity we assume that θ is uniformly distributed on the interval [0,1], ($\theta \sim \mathscr{U}(0, 1)$). The main idea of the presented method is the dyadical decomposition of the stochastic domain [0,1] and the appropriate rescaling of the polynomial basis $\{\phi_p\}_{p\in\mathbb{N}_0}$. Due to $\theta \sim \mathscr{U}(0, 1)$ we consider orthonormal Legendre polynomials. For $N_\mathrm{o} \in \mathbb{N}_0$ and $N_\mathrm{r} \in \mathbb{N}_0$ we define the *stochastic element* by $\mathsf{I}_l^{N_\mathrm{r}} := [2^{-N_\mathrm{r}}l, 2^{-N_\mathrm{r}}(l + 1)]$, for $l = 0, \ldots, 2^{N_\mathrm{r}} - 1$, and a space of the piecewise polynomials $S^{N_\mathrm{o},\, N_\mathrm{r}} := \left\{w : [0, 1] \to \mathbb{R} \mid w_{|\mathsf{I}_l^{N_\mathrm{r}}} \in \mathbb{Q}_{N_\mathrm{o}}[\theta], \forall l \in \{0, \ldots, 2^{N_\mathrm{r}} - 1\}\right\}$, where $\mathbb{Q}_{N_\mathrm{o}}[\theta]$ denotes the space of real polynomials with degree $\leq N_\mathrm{o}$. The basis of $S^{N_\mathrm{o},\, N_\mathrm{r}}$ is spanned by the polynomials $\phi_{p,l}^{N_\mathrm{r}}$ defined by

$$\phi_{i,l}^{N_\mathrm{r}}(\xi) = \begin{cases} 2^{N_\mathrm{r}/2}\phi_i(2^{N_\mathrm{r}}\xi - l), & \xi \in \mathsf{I}_l^{N_\mathrm{r}}, \\ 0, & \text{else}, \end{cases} \quad i = 0, \ldots, N_\mathrm{o}, \quad l = 0, \ldots, 2^{N_\mathrm{r}} - 1.$$

The polynomials $\phi_{0,0}^{N_\mathrm{r}}, \ldots, \phi_{N_\mathrm{o},2^{N_\mathrm{r}}-1}^{N_\mathrm{r}}$ satisfy the orthogonality relation

$$\left\langle \phi_{i,k}^{N_\mathrm{r}},\, \phi_{j,l}^{N_\mathrm{r}} \right\rangle_{L^2(\Omega)} = \delta_{ij}\delta_{kl}, \tag{4}$$

and their support is given by the appropriate stochastic element $\text{supp}(\phi_{i,l}^{N_\text{r}}) = \mathsf{I}_l^{N_\text{r}}$. The projection $\Pi^{N_\text{o},N_\text{r}} : L^2(\Omega) \to S^{N_\text{o},N_\text{r}}$ of a second order random field $w(\mathbf{x},t,\cdot) \in L^2(\Omega)$ is defined by $\Pi^{N_\text{o},N_\text{r}}[w](\mathbf{x},t,\theta) := \sum_{l=0}^{2^{N_\text{r}}-1}\sum_{i=0}^{N_\text{o}} w_{i,l}^{N_\text{r}}(\mathbf{x},t)\phi_{i,l}^{N_\text{r}}(\theta)$, where the coefficients $w_{i,l}^{N_\text{r}}$ are defined by $w_{i,l}^{N_\text{r}}(\mathbf{x},t) := \left\langle w(\mathbf{x},t,\cdot),\ \phi_{i,l}^{N_\text{r}}\right\rangle_{L^2(\Omega)}$, for $0 \le p \le N_\text{o}$ and $0 \le l \le 2^{N_\text{r}} - 1$. The convergence of $\Pi^{N_\text{o},N_\text{r}}[u]$ for $N_\text{r},\ N_\text{o} \to \infty$ is discussed in [1]. The expectation and variance of the projection $\Pi^{N_\text{o},N_\text{r}}[w]$ can be computed by the following formulae:

$$\mathbb{E}[\Pi^{N_\text{o},N_\text{r}}[w](\mathbf{x},t)] = \sum_{l=0}^{2^{N_\text{r}}-1}\sum_{p=0}^{N_\text{o}} w_{p,l}^{N_\text{r}}(\mathbf{x},t)\left\langle \phi_{p,l}^{N_\text{r}},\ \phi_{0,0}^{0}\right\rangle_{L^2(\Omega)}, \tag{5}$$

$$\begin{aligned}\text{Var}[\Pi^{N_\text{o},N_\text{r}}[w](\mathbf{x},t)] = &\sum_{l=0}^{2^{N_\text{r}}-1}\sum_{p=0}^{N_\text{o}}\sum_{q=0}^{N_\text{o}} w_{p,l}^{N_\text{r}}(\mathbf{x},t)w_{q,l}^{N_\text{r}}(\mathbf{x},t)\left\langle \phi_{p,l}^{N_\text{r}}\phi_{q,l}^{N_\text{r}},\ \phi_{0,0}^{0}\right\rangle_{L^2(\Omega)} \\ &- \left(\mathbb{E}[\Pi^{N_\text{o},N_\text{r}}[w](\mathbf{x},t)]\right)^2.\end{aligned} \tag{6}$$

Together with the orthogonality relation (4) of $\phi_{q,l}^{N_\text{r}}$ for $q = 0,\ldots,N_\text{o}$, $l = 0,\ldots,2^{N_\text{r}} - 1$ and the fact $\phi_{0,0}^{0} \equiv 1$ for $\mathscr{U}(0,1)$ we obtain

$$\text{Var}[\Pi^{N_\text{o},N_\text{r}}[w](\mathbf{x},t)] = \sum_{l=0}^{2^{N_\text{r}}-1}\sum_{p=0}^{N_\text{o}}\left(w_{p,l}^{N_\text{r}}(\mathbf{x},t)\right)^2 - \left(\mathbb{E}[\Pi^{N_\text{o},N_\text{r}}[w](\mathbf{x},t)]\right)^2.$$

3 Hybrid Stochastic Galerkin for the Two-Phase Flow Problem with Continuous Flux Function

In the deterministic case the continuous fractional flux function is defined as equivalent to $f(\mathbf{x},S) \equiv f_w(S)$. $f(\mathbf{x},S) \equiv f_w(S)$. The fractional flux of the wetting phase $f_w : [0,1] \to \mathbb{R}$ is given by $f_w(S) = f_w(S,S^e) := \frac{\lambda_w(S,S^e)}{\lambda_w(S,S^e)+\lambda_o(S,S^e)}$. Here the mean mobility λ is given by $\lambda(S,S^e) = \lambda_o(S,S^e) + \lambda_w(S,S^e)$, where λ_w denotes the total mobility of the wetting phase and λ_o the total mobility of the non-wetting phase. The effective saturation S^e is defined by $S^e(S) := (S - S_{wc})/(1 - S_{or} - S_{wc})$, with the connate saturation $S_{wc} \in [0,1]$ and the irreducible saturation $S_{or} \in [0,1]$. If the condition

$$\lambda(S,S^e) = const \tag{7}$$

is fulfilled, then the total velocity field $\mathbf{v}$ does not depend on the change of the saturation S. We use this property of $\mathbf{v}$ to stress the influence of the random disturbance.

In this section we consider the application of the HSG discretization to the two-phase flow problem with a given randomly disturbed velocity field. For this sake we replace $\mathbf{v} = (v^x, v^y)$ in (1) by $\mathbf{v}_s$ given by $\mathbf{v}_s = (v^x + c\theta, v^y)$, for $c \in \mathbb{R}$ and $\theta \sim \mathscr{U}(0, 1)$. Further we replace $\mathbf{v}_s$ and S in the Eq. (2) by their HSG representations $\Pi^{N_\mathrm{o},N_\mathrm{r}}[\mathbf{v}_s]$ and $\Pi^{N_\mathrm{o},N_\mathrm{r}}[S]$ and obtain

$$\mathbf{v} = -\mathbf{K}\lambda(S) \bigtriangledown p \quad \text{and} \quad \operatorname{div}(\mathbf{v}) = q, \tag{8}$$

$$\Pi^{N_\mathrm{o},N_\mathrm{r}}[S]_t + \operatorname{div}\left(\Pi^{N_\mathrm{o},N_\mathrm{r}}[\mathbf{v}_s]\, f\left(\Pi^{N_\mathrm{o},N_\mathrm{r}}[S]\right)\right) - q = 0, \tag{9}$$

$$S(\cdot, 0) = S_0. \tag{10}$$

We test the Eq. (9) with $\phi_{p,l}^{N_\mathrm{r}}$ for $p = 0, \ldots, N_\mathrm{o}$ and $l = 0, \ldots, 2^{N_\mathrm{r}} - 1$, that is

$$\int_\Omega \left(\Pi^{N_\mathrm{o},N_\mathrm{r}}[S]_t + \operatorname{div}\left(\Pi^{N_\mathrm{o},N_\mathrm{r}}[\mathbf{v}_s]\, f\left(\Pi^{N_\mathrm{o},N_\mathrm{r}}[S]\right)\right) - q\right)\phi_{p,l}^{N_\mathrm{r}}\, \mathrm{d}\mathbb{P}(\omega),$$

and obtain the system

$$\partial_t S_\alpha^{N_\mathrm{r}} + \operatorname{div}\left\langle \Pi^{N_\mathrm{o},N_\mathrm{r}}[\mathbf{v}_s]\, f\left(\Pi^{N_\mathrm{o},N_\mathrm{r}}[S]\right),\, \phi_\alpha^{N_\mathrm{r}}\right\rangle_{L^2(\Omega)} - \left\langle q,\, \phi_\alpha^{N_\mathrm{r}}\right\rangle_{L^2(\Omega)} = 0, \tag{11}$$

with initial values

$$S_\alpha^{N_\mathrm{r}}(\cdot, 0) = \left\langle S_0,\, \phi_\alpha^{N_\mathrm{r}}\right\rangle_{L^2(\Omega)} \tag{12}$$

for the multi-index $\alpha = (p, l)$, $p = 0, \ldots, N_\mathrm{o}$ and $l = 0, \ldots, 2^{N_\mathrm{r}} - 1$. The HSG system (11) is symmetric hyperbolic [2].

Finite volume method For the computation of the numerical solution of the hyperbolic system (11), (12) the semi-discrete central-upwind scheme introduced by Kurganov and Petrova in [7] is applied. This central-upwind method allows to work with larger systems with a minimum of requirements on the eigenvalues. Together with the HSG discretization we obtain the following numerical scheme on the triangulation $\mathscr{T} = \bigcup T_j$ of $D = (-1, 1) \times (-1, 1)$, consisting of triangular cells T_j

$$\begin{aligned}\frac{d}{dt}\bar{\mathbf{S}}_j := &- \frac{1}{|T_j|}\sum\nolimits_{k=1}^{3} h_{jk}\left(\frac{a_{jk}^{in}\mathbf{F}(\tilde{\mathbf{S}}_{jk}, M_j(k), t) + a_{jk}^{out}\mathbf{F}(\tilde{\mathbf{S}}_j, M_j(k), t)}{a_{jk}^{in} + a_{jk}^{out}}\right)\cdot \mathbf{n}_{jk} \\ &+ \frac{1}{|T_j|}\sum\nolimits_{k=1}^{3} h_{jk}\frac{a_{jk}^{in}a_{jk}^{out}}{a_{jk}^{in} + a_{jk}^{out}}\left[\tilde{\mathbf{S}}_{jk}(M_j(k)) - \tilde{\mathbf{S}}_j(M_j(k))\right] + \mathbf{q}_j\end{aligned}$$

for $\alpha = 0, \ldots, P = (N_\mathrm{o} + 1)2^{N_\mathrm{r}} - 1$. Here $\bar{\mathbf{S}}_j = \left(\bar{S}^0, \ldots, \bar{S}^P\right)$ is the cell average on the triangle $T_j \in \mathscr{T}$. The flux vector is given by $\mathbf{F} = (F^0, \ldots, F^P)^T$, where

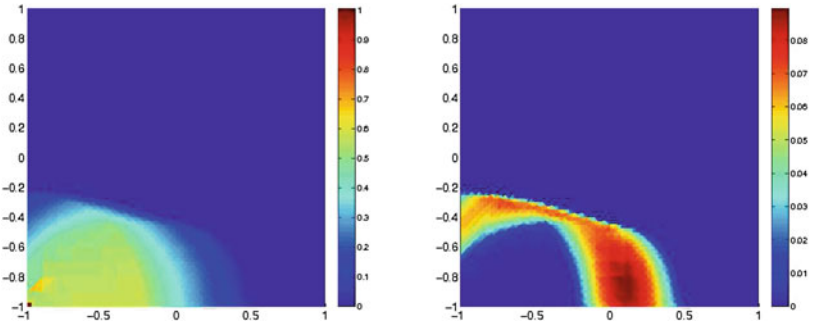


Fig. 1 **a** Expectation, **b** Variance for the problem (8)–(10) with a randomly perturbed velocity field and non-linear flux function. Computed with $N_\mathrm{r} = 5$, $N_\mathrm{o} = 4$, $T = 6$, spatial adaptivity with maximal refinement level 4

$$F^\alpha(\mathbf{S}, \mathbf{x}, t) := \left\langle f\left(\sum\nolimits_{\beta=0}^{P} S^\beta(\mathbf{x}, t)\phi_\beta^{N_\mathrm{r}}\right) \Pi^{N_\mathrm{o},N_\mathrm{r}}[\mathbf{v}_s](\mathbf{x}, t),\ \phi_\alpha^{N_\mathrm{r}}\right\rangle_{L^2(\Omega)} \quad \text{for } \alpha = 0, \ldots, P.$$

The initial values are given by $\bar{S}_j^{\alpha,0} := \frac{1}{T_j}\int_{T_j} S_0 \left\langle \phi_{0,0}^0,\ \phi_\alpha^{N_\mathrm{r}}\right\rangle_{L^2(\Omega)}$ for $\alpha = 0, \ldots, P$. For the triangle $T_j \in \mathscr{T}$, h_{jk} with $k = 1, 2, 3$ denotes the length of the k-th edge. The point $M_j(k)$ is the midpoint of the k-th edge and $\mathbf{n}_{jk}$ is the outer normal on the k-th edge, a_{jk}^{in} and a_{jk}^{out} are the so-called directional local speeds associated with the k-th edge. We use the Runge-Kutta method for the time discretization, the CFL-condition depends on the Jacobian of $\mathbf{F}$. For the computation of the reconstructions $\tilde{\mathbf{S}}_j$ and $\tilde{\mathbf{S}}_{jk}$ we refer to the work of Kurganov and Petrova [7].

Remark 1 The velocity field $\mathbf{v}$ is computed with the Taylor-Hood FEM approach, respective CG-solver, implemented in the FEM-toolbox *Alberta* [9]. The initial Delaunay triangulation is generated below the mesh generator *Triangle* [10]. We use an adaptive dynamic mesh refinement and coarsening, which uses discrete gradient heuristics and hierarchical refinement given by the bisection of the triangle on the longest edge. We perform our computation on the domain $D = (-1, 1) \times (-1, 1)$ with the initial edge-length 0.1 and max. refinement level 4.

Numerical experiments Let us apply the previously introduced numerical flux to (9), (10). We define mean mobility functions of the wetting λ_w and non-wetting λ_o phase by $\lambda_w(S, S^e) := \frac{(S^e(S))^2}{\mu_w(S^e(S)^2+(1-S^e(S))^2)}$ and $\lambda_o(S, S^e) := \frac{(1-S^e(S))^2}{\mu_o(S^e(S)^2+(1-S^e(S))^2)}$. Then the condition (7) is satisfied for $\mu_w = \mu_o = 0.3 \cdot 10^{-3}$. Therefore we can again use the velocity field $\mathbf{v} = (v^x, v^y)$ computed with the FEM framework *Alberta* at the first time-step during the computation. The randomly perturbed velocity field $\mathbf{v}_s$ is given by $\mathbf{v}_s = (v^x + c\theta, v^y)$, where $c = 0.1$ and $\theta \sim \mathscr{U}(0, 1)$. The irreducible and connate saturations are again given by $S_{or} = 0.3$ and $S_{wc} = 0.1$. Figure 1 shows the expectation and variance computed with (5), (6) at $T = 6$.

Table 1 (a) L^1-error for the problem (8)–(10) with a randomly perturbed velocity field and nonlinear flux function, at $T = 6$

		(a)					(b)		
N_o	$N_r = 2$	$N_r = 3$	$N_r = 4$	$N_r = 5$	N_o	$N_r = 2$	$N_r = 3$	$N_r = 4$	$N_r = 5$
2	2.74e-2	1.89e-2	1.10e-2	1.11e-3	2	18.5	21.2	17.8	19.6
3	2.54e-2	1.79e-2	1.0e-2	1.01e-3	3	48.4	56.2	46.2	49.7
4	2.34e-2	1.74e-2	1.10e-2	1.1e-3	4	109.9	129.3	107.3	114.3

(b) Computation time (in hours) for the problem (8)–(10) with a randomly perturbed velocity field and nonlinear flux function, at $T = 6$ computed on 2^{N_r} CPU's

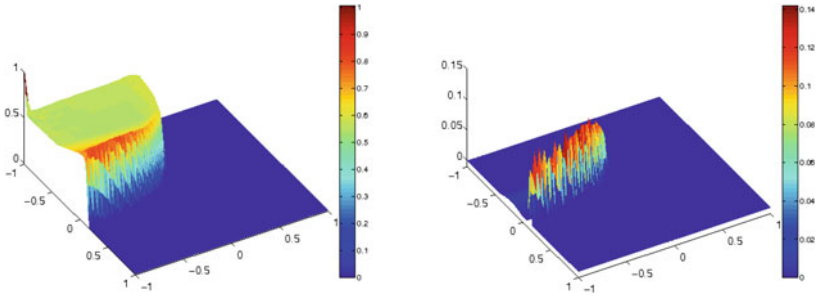
Up to our knowledge there is no analytical solution of the problem. A comparable simulation with the Monte Carlo finite volume (MC-FV) method in two space dimensions is not possible with nowadays computer power. Therefore we compare our numerical results with the most fine HSG-FV solution we could realize, that means $N_r = 6$ and $N_o = 4$. Due to the accuracy tests in one space dimension in comparison with a MC-solution, considered in [2], we can expect that this comparison represents the behaviour of the method correctly. Table 1 shows the L^1-error and computing times for $N_r = 1, \ldots, 5$ and $N_o = 1, \ldots, 4$. These results seem to indicate that the overall approach leads to convergence for increasing N_r and N_o. The computing times show, that the computational effort per node does not change significantly for increasing N_r and a constant number of stochastic elements $\mathrm{I}_l^{N_r}$ per node.

4 HSG for Two-Phase Flows in a Heterogeneous Porous Media

Now we focus on two-phase flow problems with a non-linear, spatially discontinuous flux function. A specific application is a heterogeneous porous medium characterised by two different materials. In the deterministic case the considered spatial domain D is decomposed such that $D = D^1 \cup D^2$. Within one subdomain D^i, $i = 1, 2$, the medium is supposed to be homogeneous. Hence, the descriptive parameters depend on the spatial position. By the introduction of a discontinuity function $\gamma : D \to [0, 1]$, in order to determine the location, and a uniformly distributed random variable θ, we define the randomly perturbed discontinuous fractional flux

$$f_w(\mathbf{x}, \gamma, S, \theta) := \gamma(x + c\theta, y) f^{1,w}(S) + (1 - \gamma(x + c\theta, y)) f^{2,w}(S), \quad \mathbf{x} \in D, \tag{13}$$

where $c \in \mathbb{R}$. The related HSG of the randomly perturbed problem (1), (2) is (non-strictly) hyperbolic (cf. [2, 6] for details). Figure 2 shows expectation and variance of the numerical solution of the problem (1), (2) with a randomly perturbed discontinuous flux (13) and deterministic velocity field $\mathbf{v}$ for $N_o = 3$ and $N_r = 3$ at $T = 15$. Computed with $S^1_{wc} = 0.1$, $S^1_{or} = 0.3$ and $S^2_{wc} = 0.4$, $S^2_{or} = 0.2$, $\mu_o = 3 \cdot 10^{-3}$, $\mu_w = 3 \cdot 10^{-3}$, $\theta \sim \mathscr{U}(0, 1)$ and coefficient $c = 0.4$.



(a) **(b)**

Fig. 2 **a** Expectation, **b** Variance for the quarter five-spot problem with random perturbed discontinuous flux. Computed with $N_r = 3$, $N_o = 3$ at $T = 15$

The numerical results show a realistic behaviour close to the discontinuity, in particular the variance shows the expected dependence of the uncertainty.

5 Outlook

In the future work we intend to develop an appropriate stochastic representation of the total velocity field **v** in the elliptic equation (1), and apply the developed numerical scheme to more general heterogeneous two-phase flow problems.

Acknowledgments The authors would like to thank the German Research Foundation (DFG) for financial support of the project within the Cluster of Excellence in Simulation Technology (EXC 310/1) at the University of Stuttgart.

References

1. Alpert, B.K.: A class of bases in L^2 for the sparse representation of integral operators. SIAM J. Math. Anal. **24**(1), 246–262 (1993). doi:10.1137/0524016
2. Bürger, R., Kröker, I., Rohde, C.: A hybrid stochastic Galerkin method for uncertainty quantification applied to a conservation law modelling a clarifier-thickener unit. ZAMM Z. Angew. Math. Mech. (2013). doi:10.1002/zamm.201200174
3. Cameron, R.H., Martin, W.T.: The orthogonal development of non-linear functionals in series of fourier-hermite functionals. Ann. Math. **2**(48), 385–392 (1947)
4. Ghanem, R.G., Spanos, P.D.: Stochastic Finite Elements: A Spectral Approach. Springer, New York (1991)
5. Helmig, R.: Multiphase Flow and Transport Processes in the Subsurface: A Contribution to the Modeling of Hydrosystems. Springer, Berlin (1997)
6. Kröker, I., Bürger, R., Rohde, C.: Uncertainty quantification for a clarifier-thickener model with random feed. In: Finite Volumes for Complex Applications VI, vol. 1, pp. 195–203. Springer, Berlin (2011)

7. Kurganov, A., Petrova, G.: Central-upwind schemes on triangular grids for hyperbolic systems of conservation laws. Numer. Meth. Partial Differ. Equ. **21**(3), 536–552 (2005). doi:10.1002/num.20049
8. Poëtte, G., Després, B., Lucor, D.: Uncertainty quantification for systems of conservation laws. J. Comput. Phys. **228**(7), 2443–2467 (2009). doi:10.1016/j.jcp.2008.12.018
9. Schmidt, A., Siebert, K.G.: Design of adaptive finite element software. In: The Finite Element Toolbox ALBERTA, with 1 CD-ROM (Unix/Linux). Lecture Notes in Computational Science and Engineering, vol. 42. Springer, Berlin (2005)
10. Shewchuk, J.R.: Triangle: engineering a 2D quality mesh generator and delaunay triangulator. In: Lin, M.C., Manocha, D. (eds.) Applied Computational Geometry: Towards Geometric Engineering. From the First ACM Workshop on Applied Computational Geometry. Lecture Notes in Computer Science, vol. 1148, pp. 203–222. Springer, Berlin (1996)
11. Tryoen, J., Le Maître, O., Ndjinga, M., Ern, A.: Intrusive Galerkin methods with upwinding for uncertain nonlinear hyperbolic systems. J. Comput. Phys. **229**(18), 6485–6511 (2010). doi:10.1016/j.jcp.2010.05.007
12. Xiu, D., Karniadakis, G.E.: Modeling uncertainty in flow simulations via generalized polynomial chaos. J. Comput. Phys. **187**(1), 137–167 (2003). doi:10.1016/S0021-9991(03)00092-5

A New Discretization Method for the Convective Terms in the Incompressible Navier-Stokes Equations

N. Kumar, J. H. M. ten Thije Boonkkamp and B. Koren

Abstract In this contribution we present the use of local one-dimensional boundary value problems (BVPs) to compute the interface velocities in the convective terms of the incompressible Navier-Stokes equations. This technique provides us with a better estimate for the interface velocities than linear interpolants.

1 Introduction

We present an accurate method to compute the interface velocities needed in the convective terms of the momentum equations by solving local one-dimensional BVPs. This method can be used as an improvement to a second-order accurate finite volume method on a staggered grid, with central difference discretization for the viscous and convective terms. Such a setup gives us an energy conserving discretization method for the incompressible Navier-Stokes equations [2]. The standard method for computing the interface velocities makes use of linear interpolation, i.e., by taking the average values, or alternatively, the upwind values. In this paper, we will solve a reduced form of the momentum equations, locally over a grid cell, in order to compute the interface velocities. The idea is inspired by the complete flux scheme for the advection-diffusion-reaction equation as presented in [3]. In this paper we consider the two-dimensional incompressible Navier-Stokes equations, the proposed method can be extended to the three-dimensional case.

N. Kumar (✉) · J. H. M. ten Thije Boonkkamp · B. Koren
Department of Mathematics and Computer Science, Eindhoven University of Technology, PO Box 513, 5600 MB Eindhoven, The Netherlands
e-mail: n.kumar@tue.nl

J. H. M. ten Thije Boonkkamp
e-mail: j.h.m.tenthijeboonkkamp@tue.nl

B. Koren
e-mail: b.koren@tue.nl

J. Fuhrmann et al. (eds.), *Finite Volumes for Complex Applications VII - Methods and Theoretical Aspects*, Springer Proceedings in Mathematics & Statistics 77,
DOI: 10.1007/978-3-319-05684-5_35,

In Sect. 2 of this paper, we outline the finite volume method for the incompressible Navier-Stokes equations. In Sect. 3 we describe the methods for solving the one-dimensional nonlinear BVPs. The computed interface velocities are then compared with highly accurate numerical solutions in Sect. 4. We conclude with results in Sect. 5.

2 Finite volume method

Consider the dimensionless incompressible Navier-Stokes equations, i.e.,

$$\nabla \cdot \mathbf{u} = 0, \tag{1a}$$

$$\frac{\partial \mathbf{u}}{\partial t} + \nabla \cdot (\mathbf{u}\mathbf{u}) = -\nabla p + \frac{1}{\mathrm{Re}} \nabla^2 \mathbf{u}, \tag{1b}$$

where $\mathbf{u} = (u, v)$ is the velocity of the fluid, p the pressure and Re the Reynolds number. We use the second-order finite volume method to discretize the above system of equations, as discussed in [1]. The spatial discretization is done using a staggered Cartesian grid, with the pressure and the velocity components defined at different locations, see Fig. 1. The semi-discrete form of Eq. (1a) and (1b) then reads:

$$\begin{aligned} Du(t) &= r_1(t), \\ |\Omega| u'(t) &= -C(u, v) + \frac{1}{\mathrm{Re}} Lu(t) - Gp(t) + r_2(t), \end{aligned}$$

where D, C, L and G represent the discrete divergence, convection, diffusion and gradient operators, respectively, and where $|\Omega|$ represents the measure of the control volumes [1]. The terms $r_1(t)$ and $r_2(t)$ give the boundary conditions for the system of equations. In two dimensions, $|\Omega|$ can be expressed as $|\Omega| = \mathrm{diag}(|\Omega^u_{i,j}|, |\Omega^v_{i,j}|)$, with $|\Omega^u_{i,j}| = |\Omega^v_{i,j}| = \Delta x \Delta y$. Let us consider the convective discretization for the u-component, i.e.,

$$\begin{aligned} \big(C^u(u, v)\big)_{i,j} = \Delta y \big(u^2_{i+1/2,j} - u^2_{i-1/2,j}\big) + \Delta x \big(v_{i+1/2,j}\, u_{i,j+1/2} \\ - v_{i+1/2,j-1}\, u_{i,j-1/2}\big). \end{aligned} \tag{2}$$

For computing $(C^u(u, v))_{i,j}$, we need methods to compute the interface velocities $u_{i+1/2,j}$, $v_{i+1/2,j}$ and $u_{i,j+1/2}$. In this paper we focus on the computation of $u_{i+1/2,j}$. The velocity $u_{i+1/2,j}$ can be simply taken as the average

$$u_{i+1/2,j} = \frac{u_{i,j} + u_{i+1,j}}{2}.$$

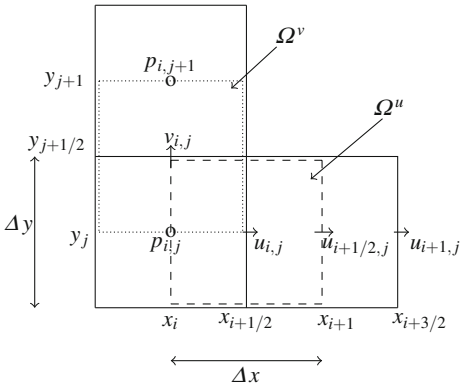


Fig. 1 Staggered grid structure for spatial discretization

In this paper, we aim to compute $u_{i+1/2,j}$ by solving a reduced form of the u-momentum equation locally. The u-momentum equation reads

$$u_t + (uu)_x + (uv)_y = -p_x + \frac{1}{\mathrm{Re}}\left(u_{xx} + u_{yy}\right).$$

Let us assume that the flow is locally steady and one-dimensional. Moreover, we ignore all terms involving y. Then the previous equation is reduced to

$$uu_x - \varepsilon u_{xx} = -p_x, \tag{3}$$

where $\varepsilon = 1/\mathrm{Re}$. Thus, we are left with a nonlinear differential equation. In the following we ignore the y-dependence of u and we simply write $u = u(x)$. We denote $u(x_i)$ as u_i. In order to get the interface velocity $u_{i+1/2}$ located at x_{i+1} we solve Eq. (3) for $x \in \left(x_{i+1/2}, x_{i+3/2}\right)$ subject to the boundary conditions

$$u(x_{i+1/2}) = u_i, \quad u(x_{i+3/2}) = u_{i+1}. \tag{4}$$

The following section details the computation of $u_{i+1/2,j}$ and briefly outlines the computation of $u_{i,j+1/2}$ and $v_{i+1/2,j}$.

3 Computing the Interface Velocities

The BVP (3)–(4) is difficult to solve due to the nonlinear term uu_x and the pressure gradient p_x. We simplify this by first solving a linearized problem without the pressure gradient and subsequently solving the linearized problem along with the pressure term.

Let U be in between u_i and u_{i+1}, or equal to either u_i or u_{i+1}. We linearize the nonlinear term of Eq. (3), to get

$$Uu_x - \varepsilon u_{xx} = -p_x.$$

In the derivation that follows, it is convenient to introduce the following notation, $\mathsf{a} = U/\varepsilon$ and $(.)' = \partial/\partial x$. We define the *local mesh Péclet number* (P) as

$$\mathsf{P} = \frac{U\Delta x}{\varepsilon} = \mathsf{a}\Delta x, \tag{5}$$

and the normalized coordinate by

$$\sigma(x) = \frac{x - x_{i+1/2}}{\Delta x} \qquad \text{for } x \in \left[x_{i+1/2}, x_{i+3/2}\right].$$

So the linearized equation can now be written as,

$$\varepsilon\left(u' - \mathsf{a}u\right)' = p'. \tag{6}$$

The problem given by Eq. (6) with boundary conditions (4), can now be split in two cases, the *homogeneous case*, having $p' = 0$, and the *inhomogeneous case*, in which we assume a piecewise linear pressure. We first consider the homogeneous case.

Homogeneous case. Using $u' - \mathsf{a}u = e^{\mathsf{a}x}\left(e^{-\mathsf{a}x}u\right)'$ and integrating Eq. (6) (with the assumption $p' = 0$), from $x_{i+1/2}$ to $x \in [x_{i+1/2}, x_{i+3/2}]$, and applying the boundary condition $u(x_{i+1/2}) = u_i$, gives

$$e^{-\mathsf{a}x}u(x) - e^{-\mathsf{a}x_{i+1/2}}u_i = \frac{C_1}{\mathsf{a}}\left(e^{-\mathsf{a}x_{i+1/2}} - e^{-\mathsf{a}x}\right), \quad (\mathsf{a} \neq 0).$$

Formulating in terms of σ and P, and imposing the other boundary condition $u(x_{i+3/2}) = u_{i+1}$, gives

$$u(x) = \frac{e^{-\mathsf{P}(1-\sigma(x))} - 1}{e^{-\mathsf{P}} - 1}u_i + \frac{e^{\mathsf{P}\sigma(x)} - 1}{e^{\mathsf{P}} - 1}u_{i+1}. \tag{7}$$

We assume that the grid is equidistant. Then putting $\sigma(x) = \frac{1}{2}$ in the above expression gives

$$u_{i+1/2} = A(-\mathsf{P}/2)u_i + A(\mathsf{P}/2)u_{i+1}, \tag{8}$$

where $A(z) = \left(e^z + 1\right)^{-1}$. Alternatively, $u_{i+1/2}$ can also be expressed as the sum of the average value and a correction term, as

$$u_{i+1/2} = \frac{(u_i + u_{i+1})}{2} + \left(A(\mathsf{P}/2) - \frac{1}{2}\right)(u_{i+1} - u_i). \tag{9}$$

From Eq. (8), we see that $u_{i+1/2}$ is a weighted average of u_i and u_{i+1}. It can be observed that in the limit $\mathsf{P} \to 0$, we recover the average value. For $\mathsf{P} = 0$, we

have $\mathsf{a} = 0$, implying $\varepsilon u'' = 0$, which gives $u(x_{i+1}) = (u_i + u_{i+1})/2$. In the limit $|\mathsf{P}| \to \infty$, we get $A(|\mathsf{P}/2|) = 0$, thereby giving $u_{i+1/2} = u_i$ or u_{i+1}, depending on the direction of the flow.

We compute the velocity $u_{i+1/2}$ iteratively, by initializing $U = (u_i + u_{i+1})/2$ and P as given in Eq. (5) and then compute $u_{i+1/2}$ using Eq. (8). For the next iteration, we take U to be the computed value of $u_{i+1/2}$ and update P accordingly, and then compute a new value of $u_{i+1/2}$ using Eq. (8). We continue this procedure until the values of $u_{i+1/2}$ computed after each iteration have converged, i.e., when the absolute difference between the values of $u_{i+1/2}$ computed at consecutive iterations has dropped below a fixed tolerance.

Inhomogeneous case In this case we solve the linearized boundary value problem given by Eq. (6), under the assumption that the pressure p is *piecewise linear*. We initially proceed as we did in the homogeneous case, so we have

$$e^{\mathsf{a}x}\left(e^{-\mathsf{a}x}u\right)' = \frac{1}{\varepsilon}\mathrm{I}(x) + C_1, \qquad \mathrm{I}(x) = \int_{x_{i+1}}^{x} p'(\xi)\mathrm{d}\xi, \tag{10}$$

and Eq. (10) then becomes

$$\left(e^{-\mathsf{a}x}u\right)' = \frac{1}{\varepsilon}e^{-\mathsf{a}x}\mathrm{I}(x) + C_1 e^{-\mathsf{a}x}. \tag{11}$$

The value $I(x)$ can be calculated as

$$I(x) = \begin{cases} \frac{1}{\Delta x}(p_{i+1} - p_i)(x - x_{i+1}) = (p_{i+1} - p_i)(\sigma(x) - \frac{1}{2}), & \text{for } 0 \le \sigma(x) \le \frac{1}{2}. \\ \frac{1}{\Delta x}(p_{i+2} - p_{i+1})(x - x_{i+1}) = (p_{i+2} - p_{i+1})(\sigma(x) - \frac{1}{2}), & \text{for } \frac{1}{2} \le \sigma(x) \le 1. \end{cases}$$

Integrating Eq. (11) from $x_{i+1/2}$ to $x \in \left[x_{i+1/2}, x_{i+3/2}\right]$ and using the boundary condition $u(x_{i+1/2}) = u_i$, we get

$$u(x) - e^{\mathsf{P}\sigma(x)}u_i = \frac{1}{\varepsilon}\int_{x_{i+1/2}}^{x} e^{\mathsf{a}(x-\xi)}\mathrm{I}(\xi)\mathrm{d}\xi + \frac{C_1}{\mathsf{a}}(e^{\mathsf{P}\sigma(x)} - 1).$$

We define

$$\mathrm{J}(x) \equiv \int_{x_{i+1/2}}^{x} e^{\mathsf{a}(x-\xi)}\mathrm{I}(\xi)\mathrm{d}\xi,$$

and use the boundary condition $u(x_{i+3/2}) = u_{i+1}$ to get the solution

$$u(x) = \frac{e^{-\mathsf{P}(1-\sigma(x))} - 1}{e^{-\mathsf{P}} - 1}u_i + \frac{e^{\mathsf{P}\sigma(x)} - 1}{e^{\mathsf{P}} - 1}u_{i+1} + \frac{1}{\varepsilon}\left(\mathrm{J}(x) - \frac{e^{\mathsf{P}\sigma(x)} - 1}{e^{\mathsf{P}} - 1}\mathrm{J}(x_{i+3/2})\right). \tag{12}$$

We now express the velocity $u(x)$ as the sum of a homogeneous part $u^h(x)$ and an inhomogeneous part $u^i(x)$ as

$$u(x) = u^h(x) + u^i(x).$$

The homogeneous part $u^h(x)$ of the velocity, as given by Eq. (7), depends on the convection-diffusion operator, whereas the inhomogeneous part, $u^i(x)$, depends on the pressure gradient. Computing the values of the integrals $\mathrm{J}(x_{i+1})$ and $\mathrm{J}(x_{i+3/2})$, and introducing

$$F(z) \equiv \frac{e^z - 1 - z}{z^2(e^z + 1)},$$

gives us

$$u(x_{i+1}) = u^h(x_{i+1}) + u^i(x_{i+1}), \tag{13a}$$

$$u^h(x_{i+1}) = \mathrm{A}(-\mathsf{P}/2)u_i + \mathrm{A}(\mathsf{P}/2)u_{i+1}, \tag{13b}$$

$$u^i(x_{i+1}) = -\frac{(\Delta x)^2}{4\varepsilon}\left[F(-\mathsf{P}/2)\frac{p_{i+1} - p_i}{\Delta x} + F(\mathsf{P}/2)\frac{p_{i+2} - p_{i+1}}{\Delta x}\right]. \tag{13c}$$

In this case also the computation of $u(x_{i+1})$ is iterative, where we begin by taking $U = u^h = (u_{i,j} + u_{i+1,j})/2$ and $u^i = 0$ and compute $u(x_{i+1})$ using the above equations. Now proceed as in case of the homogeneous case, until the values converge.

Till now we have discussed the computation of the interface velocity $u_{i+1/2,j}$ but for computing the convective term as given by Eq. (2), we also require $u_{i,j+1/2}$ and $v_{i+1/2,j}$. These velocities can also be computed using local BVPs. The interface velocity $u_{i,j+1/2}$ is computed from the BVP

$$Vu_y - \varepsilon u_{yy} = 0, \quad y_j < y < y_{j+1}, \tag{14a}$$

$$u(y_j) = u_j, \quad u(y_{j+1}) = u_{j+1}, \tag{14b}$$

and $v_{i+1/2,j}$, from

$$Uv_x - \varepsilon v_{xx} = 0, \quad x_i < x < x_{i+1}, \tag{15a}$$

$$v(x_i) = v_i, \quad v(x_{i+1}) = v_{i+1}. \tag{15b}$$

These velocities are also computed iteratively. We begin the iterations by defining $V = (v_{i,j} + v_{i+1,j})/2$ and $\mathsf{P}_v = V\Delta y/\varepsilon$ for BVP (14a) and (14b) and $U = (u_{i,j} + u_{i,j+1})/2$, $\mathsf{P}_u = U\Delta x/\varepsilon$ for BVP (15a) and (15b). We then compute $u_{i,j+1/2}$ and $v_{i+1/2,j}$ using an equation analogous to (8). For the next iteration, we assign V the value of $v_{i+1/2,j}$ computed in the previous iteration and U the value of $u_{i,j+1/2}$ computed in the previous iteration. With the new values of V and U we update P_v and P_u and recompute $u_{i,j+1/2}$ and $v_{i+1/2,j}$. We continue this procedure until the values converge. Computing the interface velocities in this manner results in the coupling between u and v interface velocities. The next section gives a validation of the method presented above.

Table 1 Validation for the homogeneous case, $u_{i+1/2,j}$ according to Eq. (8), ($\Delta x = 10^{-3}$)

Re	$u_{i+1/2,j}$	u_{num}	$\text{d}u_{\text{avg}}$	$\text{d}u_{up}$	# iterations
10	1.1749	1.1749	0.012	4.432	2
10^2	1.1735	1.1735	0.125	4.314	3
10^3	1.1609	1.1608	1.201	3.189	4
10^4	1.1254	1.1253	4.225	0.032	4
10^5	1.1250	1.1250	4.255	0	2
10^6	1.1250	1.1250	4.255	0	2

Table 2 Validation for inhomogeneous case, $u_{i+1/2,j}$ according to Eq. (13a)–(13c), ($\Delta x = 10^{-3}$)

Re	$u^h(x_{i+1})$	$u^i(x_{i+1})$	$u(x_{i+1})$	u_{num}	# iterations
10	1.1749	1.47×10^{-5}	1.1749	1.1749	2
10^2	1.1735	1.47×10^{-4}	1.1737	1.1738	3
10^3	1.1609	1.43×10^{-3}	1.1623	1.1637	4
10^4	1.1253	5.16×10^{-3}	1.1305	1.1356	4
10^5	1.1250	5.19×10^{-3}	1.1302	1.1354	4
10^6	1.1250	5.19×10^{-3}	1.1302	1.1354	4

4 Numerical Validation

In order to check the accuracy of the computed interface velocity $u_{i+1/2,j}$, we compare it with the value obtained by computing a very accurate numerical solution of the local boundary value problem.

For the validation of our computed interface velocities, we take $u_{i,j} = 1.125$ and $u_{i+1,j} = 1.225$. For this setup the value of $u_{i+1/2,j}$ obtained by central averaging (u_{avg}) is 1.175 and by taking the upwind value (u_{up}) it is 1.125. Let u_{num} denote the highly accurate numerically computed value of $u_{i+1/2,j}$. We next define the relative absolute differences $\text{d}u_{\text{avg}} = \frac{|u_{\text{avg}} - u_{i+1/2,j}|}{|u_{\text{avg}}|}$ and $\text{d}u_{\text{num}} = \frac{|u_{\text{num}} - u_{i+1/2,j}|}{|u_{\text{num}}|}$. The convergence criterion, i.e., the difference of the computed interface velocities from two consecutive iterations, is taken to be 10^{-7}. Table 1 gives the results obtained for the homogeneous case. It can be observed that the computed interface velocity is very accurate (when compared to u_{num}) and attains the upwind value for higher Reynolds numbers. Similarly, for the inhomogeneous case, let $u(x_{i+1})$ be the velocity computed using Eq. (13a)–(13c), as the sum of $u^h(x_{i+1})$ and $u^i(x_{i+1})$. Table 2 gives the obtained results. For this case a constant pressure gradient of -0.01175 has been assumed. The inhomogeneous part of the velocity increases with increasing Reynolds number. The effect of adding a pressure gradient can be seen in Fig. 2, where we have plotted the interface velocity with increasing Reynolds number. In absence of a pressure gradient the computed interface velocity is almost equal to the numerical solution see Fig. 2a. On adding the pressure gradient, the homogeneous

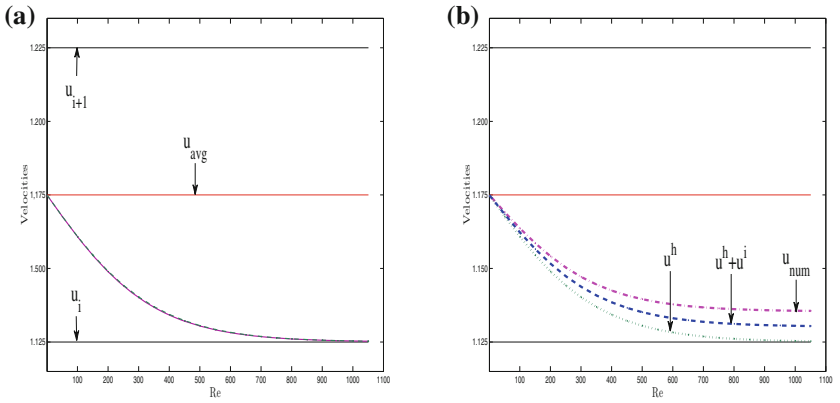


Fig. 2 Effect of adding a pressure gradient to the BVP, $u_i = 1.125$, $u_{i+1} = 1.225$, $\Delta x = 10^{-2}$ **a** Interface velocities in absence of a pressure gradient. **b** Interface velocities with a negative pressure gradient

part of the velocity remains the same, and the addition of the inhomogeneous part corrects the interface velocity $u_{i+1/2,j}$, in a physically proper way.

5 Conclusions

To compute interface velocities, we have proposed an iterative discretization method that depends on the local mesh Péclet number, P. For increasing P, the homogeneous part of the velocity attains the *upwind* value, and for decreasing values of P, it converges towards the average velocity. The pressure gradient plays an important role in the determination of $u_{i+1/2,j}$. For increasing pressure gradient, the difference between the approximations of $u_{i+1/2,j}$, u^h and $u^h + u^i$ and the numerical solution u_{num} increases see Fig. 2b. For a negative pressure gradient, the interface velocity increases, whereas for a positive pressure gradient it decreases. The increment/decrement grows with an increase in the absolute value of the pressure gradient.

We have applied the methods proposed in this paper to the two-dimensional flow in a lid-driven square cavity. It was observed that the difference in the velocities computed using the proposed iterative method and those computed using the average method is very small for small values of $\Delta x/\varepsilon$ but starts to increase as $\Delta x/\varepsilon$ increases. The gain of the present method is to be sought in the possibility to use much coarser grids, with the same accuracy as standard methods.

References

1. Sanderse, B.: Energy-conserving discretization methods for the incompressible Navier-Stokes equations: application to the simulation of wind-turbine wakes. Ph.D. thesis, Eindhoven University of Technology (2013)
2. Sanderse, B.: Energy-conserving Runge-Kutta methods for the incompressible Navier-Stokes equations. J. Comput. Phys. **233**, 100–131 (2013)
3. ten Thije Boonkkamp, J.H.M., Anthonissen, M.J.H.: The finite volume-complete flux scheme for advection-diffusion-reaction equations. J. Sci. Comput. **46**(1), 47–70 (2011)

Mimetic Finite Difference Schemes with Conditional Maximum Principle for Diffusion Problems

Konstantin Lipnikov

Abstract Numerical schemes that satisfy the maximum principle play important role in multiphysics codes. They reduce significantly various numerical artifacts. We describe a novel inexpensive practical algorithm for building mimetic finite difference schemes with conditional maximum principle on polygonal and polyhedral meshes for diffusion problems.

1 Introduction

Numerical schemes that preserve important properties of underlying PDEs lead in general to more robust computer simulations. These schemes reduce significantly or eliminate totally various numerical artifacts. An important property of a diffusion problem is the existence of the maximum principle (MP). In its simplest form, it states that in absence of external sources, the continuum solution has no internal extrema. This implies that physical quantities, such as temperature or chemical concentration, are always bounded by the boundary data.

It is well known that the second-order linear schemes for the diffusion equation satisfy the MP only under some conditions on the mesh and diffusion tensor. Analysis of sufficient conditions for the MP on unstructured simplicial meshes started in 70th, see e.g. [2]. For the mimetic finite difference (MFD) method, sufficient conditions were formulated in [5]; however, algorithms for their verification were developed for a limited class of meshes. Here, we propose a practical algorithm for verifying the sufficient conditions and building mimetic schemes with the MP for meshes with arbitrarily-shaped cells. The algorithm satisfies a few requirements of emerging computer architectures: large flops per memory ratio and data locality.

K. Lipnikov (✉)
Los Alamos National Laboratory, Los Alamos, NM 87544, USA
e-mail: lipnikov@lanl.gov

J. Fuhrmann et al. (eds.), *Finite Volumes for Complex Applications VII - Methods and Theoretical Aspects*, Springer Proceedings in Mathematics & Statistics 77,
DOI: 10.1007/978-3-319-05684-5_36,

A detailed description of the MFD method can be found in the recently published book [8] which focuses on numerical solution of elliptic PDEs. In general, the concept of mimetic (or compatible in general) schemes can be applied to a greater variety of PDEs (see [4] and references therein). The incomplete list of related compatible discretization methods includes discrete duality finite volume (FV), hybrid FV and mixed FV methods. For diffusion problems, the algebraic equivalence of the MFD, mixed FV, and hybrid FV methods has been shown in [3].

The paper outline is as follows. In Sect. 2, we describe briefly the MFD method. In Sect. 3, we formulate sufficient conditions for the MP and present a practical algorithm for verifying them and selecting an optimal scheme. In Sect. 4, we analyze numerically the complexity of the proposed algorithm.

2 A Family of Mimetic Finite Difference Schemes

Let $\Omega \subset \Re^d$, $d = 2$ or 3, be a bounded domain with a Lipschitz boundary Γ. We consider the following mixed formulation of the elliptic equation:

$$\mathbf{u} = -\mathbb{K}\nabla p \quad \text{and} \quad \operatorname{div} \mathbf{u} = b, \tag{1}$$

where p is pressure, $\mathbf{u}$ is velocity, $\mathbb{K}$ is symmetric diffusion tensor, and b is source term. To simplify the presentation, we assume that $p = 0$ on Γ.

Let Ω_h be a conformal partition of Ω into polyhedral ($d = 3$) or polygonal ($d = 2$) cells c. We denote by $|\mathrm{c}|$ the volume (area in 2D) of cell c. For face f of cell c, we denote by $|\mathrm{f}|$ its area (length in 2D) and by $\mathbf{n}_{\mathrm{f,c}}$ its exterior unit normal vector. We assume that the diffusion tensor has constant value $\mathbb{K}_\mathrm{c}$ in cell c.

The discrete pressure space Q_h consists of one degree of freedom per cell, p_c, and one degree of freedom per face, p_f, approximating the average pressure value in c and f, respectively. Thus, the dimension of Q_h equals to the number of mesh cells plus the number of mesh faces.

The discrete velocity space X_h consists of one degree of freedom, $u_{\mathrm{f,c}}$, per face f of cell c, which approximates the average flux $\mathbf{u} \cdot \mathbf{n}_{\mathrm{f,c}}$ across face f. Thus, the dimension of X_h equals to the number of boundary faces plus twice the number of interior faces. For each vector $\mathbf{u}_h \in X_h$, we denote by $\mathbf{u}_{h,\mathrm{c}}$ its restriction to cell c, i.e. $\mathbf{u}_{h,\mathrm{c}} = \{u_{\mathrm{f,c}}\}_{\mathrm{f}\in\partial\mathrm{c}}$. The mass conservation law implies the following condition:

$$u_{\mathrm{f,c}_1} = -u_{\mathrm{f,c}_2}, \tag{2}$$

for each face f shared by cells c_1 and c_2.

Integrating the second equation in (1) over cell c, we obtain:

$$\operatorname{div}^h \mathbf{u}_{h,\mathrm{c}} = b_c, \quad \operatorname{div}^h \mathbf{u}_{h,\mathrm{c}} = \frac{1}{|\mathrm{c}|} \sum_{\mathrm{f}\in\partial\mathrm{c}} |\mathrm{f}|\, u_{\mathrm{f,c}}, \quad b_c = \frac{1}{|\mathrm{c}|} \int_\mathrm{c} b \mathrm{d}V. \tag{3}$$

What is left is to discretize the first equation in (1).

2.1 Consistency and Stability Conditions

The constitutive equation is discretized using the consistency and stability conditions. The *consistency condition* is the exactness property and can be formulated in a variety of ways. Here we use the FV viewpoint that is more appropriate for the practical implementation of the MFD method; however, it hides its theoretical roots.

Let us consider a cell c with n_c faces f_i. We assume that there exists a linear relationship between the discrete unknowns,

$$\begin{pmatrix} u_{f_1,c} \\ u_{f_2,c} \\ \vdots \\ u_{f_{n_c},c} \end{pmatrix} = -\mathbb{W}_c \begin{pmatrix} |f_1|\,(p_{f_1} - p_c) \\ |f_2|\,(p_{f_2} - p_c) \\ \vdots \\ |f_{n_c}|\,(p_{f_{n_c}} - p_c) \end{pmatrix}, \tag{4}$$

with a symmetric and positive definite matrix $\mathbb{W}_c$. Then, the mimetic scheme is defined by collecting Eqs. (4), (2), (3) and imposing the homogeneous boundary conditions, i.e. $p_f = 0$ for all $f \in \Gamma$.

To define matrix $\mathbb{W}_c$, we require that (4) is exact for any linear function p and the corresponding constant vector function $\mathbf{u}$. It is sufficient to consider $d + 1$ linearly independent pressure functions: $p_0 = 1$, $p_1 = x$, $p_2 = y$ and $p_3 = z$ in 3D. Obviously, formula (4) is trivial for $p_0 = 1$ and $\mathbf{u}_0 = \mathbf{0}$. Taking pairs p_i and $\mathbf{u}_i = -\mathbb{K}_c \nabla p_i$, calculating vectors of degrees of freedom and inserting them in (4), we obtain d matrix equations:

$$\mathbb{N}_{c,i} = \mathbb{W}_c \mathbb{R}_{c,i}, \qquad 1 \le i \le d, \tag{5}$$

where $\mathbb{N}_i$ and $\mathbb{R}_i$ are $n_c \times 1$ vectors. These vectors can be calculated using only areas, centroids and normals to faces of c which results in relatively simple calculations for an arbitrary-shaped cell (see [8, Chap. 5] or [4] for more details).

Let us define $n_c \times d$ matrices $\mathbb{N}_c = [\mathbb{N}_{c,1}, \ldots, \mathbb{N}_{c,d}]$ and $\mathbb{R}_c = [\mathbb{R}_{c,1}, \ldots, \mathbb{R}_{c,d}]$. It has been proved in [1] that a particular solution to matrix Eq. (5) is

$$\mathbb{W}_c^{(0)} = \frac{1}{|c|} \mathbb{N}_c \, \mathbb{K}_c^{-1} \, \mathbb{N}_c^T.$$

The rank of this matrix is d which is strictly less than n_c. Thus, to build a positive definite $n_c \times n_c$ matrix $\mathbb{W}_c$, we have to add a stabilization term $\mathbb{W}_c^{(1)}$ such that $\mathbb{W}_c^{(1)} \mathbb{R}_c = 0$. The *stability condition* imposes lower and upper bounds on this term. More precisely, it requires the matrix $\mathbb{W}_c = \mathbb{W}_c^{(0)} + \mathbb{W}_c^{(1)}$ to be spectrally equivalent to a scalar matrix:

$$a_\star \frac{1}{|c|} \|\mathbf{u}_{h,c}\|^2 \le \mathbf{u}_{h,c}^T \mathbb{W}_c \mathbf{u}_{h,c} \le a^\star \frac{1}{|c|} \|\mathbf{u}_{h,c}\|^2 \qquad \forall \mathbf{u}_{h,c}, \tag{6}$$

where $a_\star$ and $a^\star$ are mesh independent positive constants, and $\|\cdot\|$ denotes the Eucleadian norm.

2.2 A Family of Stable Mimetic Schemes

Let us introduce a full rank matrix $\mathbb{D}_c$ such that its columns span the null space of matrix $\mathbb{R}_c^T$, i.e. $\mathbb{R}_c^T\ \mathbb{D}_c = \mathbb{O}$, where $\mathbb{O}$ denotes a generic zero matrix. We assume that the columns of $\mathbb{D}_c$ are orthonormal vectors. Then,

$$\mathbb{W}_c^{(1)} = \mathbb{D}_c\, \mathbb{P}_c\, \mathbb{D}_c^T,$$

where $\mathbb{P}_c$ is a symmetric positive definite matrix of parameters. The stability condition does not allow $\mathbb{P}_c$ to have arbitrarily small or large eigenvalues. In practice, a good choice for $\mathbb{P}_c$ is the scalar matrix $\alpha_c\mathbb{I}$ where $\alpha_c = \frac{1}{n_c}\text{trace}(\mathbb{W}_c^{(0)})$. In this case, the condition number of $\mathbb{W}_c$ depends only on the anisotropy of tensor $\mathbb{K}_c$ and the shape-regularity constants of cell c.

3 Mimetic Schemes with the Maximum Principle

For a polyhedral mesh, a family of admissible mimetic schemes is quite large. Indeed for each polyhedral cell with n_c faces, we have $(n_c - d + 1) \times (n_c - d)/2$ parameters forming the symmetric matrix $\mathbb{P}_c$. Ideally, these parameters have to be selected to enforce the MP.

3.1 Sufficient Conditions

We recall sufficient conditions for the MP proposed in [5]. Inserting (4) into (2) and (3), we obtain a system of algebraic equations for the pressure unknown $\mathbf{p}_h \in Q_h$:

$$\mathbb{A}\mathbf{p}_h = \mathbf{b}_h, \quad \mathbb{A} = \sum_{c\in\Omega_h} \mathscr{N}_c\, \mathbb{A}_c\, \mathscr{N}_c^T,$$

where $\mathscr{N}_c$ is an assembling matrix with 0 and 1 entries. The sufficient conditions for the MP are such that each cell matrix $\mathbb{A}_c$ is a singular M-matrix. If so, the global matrix $\mathbb{A}$ is a singular M-matrix. Eliminating equations corresponding to the Dirichlet boundary conditions, $p_f = 0$ for $f \in \Gamma$, we obtain an M-matrix [5]. Hence, solution $\mathbf{p}_h$ satisfies the MP.

Let us rewrite the mass balance equation (3) in the algebraic form $\mathbb{B}_{\mathrm{c}}^T \mathbf{u}_{h,\mathrm{c}} = |\mathrm{c}|\, b_{\mathrm{c}}$ where $\mathbb{B}_{\mathrm{c}}$ is the column matrix, $\mathbb{B}_{\mathrm{c}} = (|\mathrm{f}_1|, \ldots, |\mathrm{f}_{n_{\mathrm{c}}}|)^T$. We define a square diagonal matrix $\mathbb{C}_{\mathrm{c}}$ such that $\mathbb{C}_{\mathrm{c}}\mathbf{1} = \mathbb{B}_{\mathrm{c}}$, where $\mathbf{1}$ is a generic vector with all entries equal to 1. According to [5], the cell-based matrix has the following structure:

$$\mathbb{A}_{\mathrm{c}} = \begin{pmatrix} \mathbb{C}_{\mathrm{c}}^T \mathbb{W}_{\mathrm{c}} \mathbb{C}_{\mathrm{c}} & -\mathbb{C}_{\mathrm{c}}^T \mathbb{W}_{\mathrm{c}} \mathbb{B}_{\mathrm{c}} \\ -\mathbb{B}_{\mathrm{c}}^T \mathbb{W}_{\mathrm{c}} \mathbb{C}_{\mathrm{c}} & \mathbb{B}_{\mathrm{c}}^T \mathbb{W}_{\mathrm{c}} \mathbb{B}_{\mathrm{c}} \end{pmatrix}.$$

Lemma 1 ([5]). *The matrix $\mathbb{A}_{\mathrm{c}}$ is a singular M-matrix if $\mathbb{W}_{\mathrm{c}}$ is an M-matrix and the vector $\mathbb{W}_{\mathrm{c}}\, \mathbb{B}_{\mathrm{c}}$ has non-negative entries.*

3.2 Simplex Method for Matrix $\mathbb{W}_{\mathrm{c}}$

The simplex method is used twice in the construction of matrix $\mathbb{W}_{\mathrm{c}}$ that satisfies the conditions of Lemma 1. First, it answers the question of the existence of at least one such matrix. Second, it finds an optimal (in some sense) matrix when a few matrices satisfy Lemma 1. In this section, we drop out subscript 'c' from all matrices.

We illustrate this method for the quadrilateral cell, i.e. $n_{\mathrm{c}} = 4$, $d = 2$. Despite its simple shape, the direct construction of an M-matrix $\mathbb{W}$ was an open problem until now. The matrix of parameters is a 2×2 matrix characterized typically by three parameters. However, since the simplex method requires all parameters to be non-negative, we need four parameters to describe negative off-diagonal entries:

$$\mathbb{P} = \begin{pmatrix} a_1 & a_3 - a_4 \\ a_2 - a_4 & a_2 \end{pmatrix}, \quad a_i \geq 0.$$

Unless we enforce somehow the positive definiteness of matrix $\mathbb{P}$, we can only guarantee the symmetry of $\mathbb{W}$. Direct control of the properties of $\mathbb{P}$ is undesirable, since it leads to a nonlinear optimization problem. Fortunately, the properties of a M-matrix allow us to circumvent this problem. The first set of *linear* inequalities enforces the Z-matrix property for $\mathbb{W} = \mathbb{W}^{(0)} + \mathbb{W}^{(1)}$:

$$a_1 \mathbb{D}_{1i} \mathbb{D}_{1j} + a_2 \mathbb{D}_{2i} \mathbb{D}_{2j} + (a_3 - a_4)\big(\mathbb{D}_{1i} \mathbb{D}_{2j} + \mathbb{D}_{2i} \mathbb{D}_{1j}\big) \leq -\mathbb{W}_{ij}^{(0)} \quad \forall i < j.$$

Recall that a Z-matrix $\mathbb{W}$ is an M-matrix if there exists a vector $\mathbf{v}$ with non-negative entries such that $\mathbb{W}\mathbf{v} \geq \varepsilon > 0$, i.e. all entries of this matrix-vector product are strictly positive. We take $\mathbf{v} = \mathbb{B}$, so that the later property implies the second condition of Lemma 1. Since $\mathbb{W}^{(0)}\mathbf{v} = 0$, the resulting set of inequalities reads:

$$\sum_{j=1}^{n_{\mathrm{c}}} |\mathrm{f}_j| \Big(a_1 \mathbb{D}_{1i} \mathbb{D}_{1j} + a_2 \mathbb{D}_{2i} \mathbb{D}_{2j} + (a_3 - a_4)\big(\mathbb{D}_{1i} \mathbb{D}_{2j} + \mathbb{D}_{2i} \mathbb{D}_{1j}\big) \Big) \geq \varepsilon_i \quad \forall i,$$

where $\varepsilon_i > 0$. In practice, we take $\varepsilon_i = \lambda_{\min}(\mathbb{K}_c)/(10|c|)$. This choice leads to mesh-independent coefficients $a_\star$ and $a^\star$ in the stability condition (6).

The objective functional that the simplex method maximizes is the sum of all entries in $\mathbb{W}$ (this maximizes the diagonal dominance of $\mathbb{W}$):

$$\max_{a_i \geq 0} \Phi(\{a_i\}), \qquad \Phi(\{a_i\}) = \sum_{i,j=1}^{n_c} \mathbb{W}_{ij}. \tag{7}$$

Note that other linear objective functionals can be also admissible. The simplex method requires to convert the inequality constraints to equality constraints. We introduce the slack (or surplus, or logical) *non-negative* variables s_{ij} and s_i:

$$a_1 \mathbb{D}_{1i}\mathbb{D}_{1j} + a_2 \mathbb{D}_{2i}\mathbb{D}_{2j} + (a_3 - a_4)\big(\mathbb{D}_{1i}\mathbb{D}_{2j} + \mathbb{D}_{2i}\mathbb{D}_{1j}\big) + s_{ij} = -\mathbb{W}_{ij}^{(0)} \tag{8}$$

and

$$\sum_{j=1}^{n_c} |\mathrm{f}_j| \Big(a_1 \mathbb{D}_{1i}\mathbb{D}_{1j} + a_2 \mathbb{D}_{2i}\mathbb{D}_{2j} + (a_3 - a_4)\big(\mathbb{D}_{1i}\mathbb{D}_{2j} + \mathbb{D}_{2i}\mathbb{D}_{1j}\big)\Big) + s_i = \varepsilon_i. \tag{9}$$

The total number of slack variables is $n_c(n_c + 1)/2$. The slack variables are treated like the original parameters a_i until the last moment when they are just ignored. Each slack variable is the amount by which the original inequality is satisfied. The optimization problem is now to find the maximum of functional Φ subject to the equality constraints (8), (9) and the inequality constraints $a_i \geq 0$, $s_{ij} \geq 0$, $s_i \geq 0$.

To launch the simplex method, we need to prescribe valid (i.e. non-negative) initial values for the variables a_i, s_{ij} and s_i so that the above equalities are satisfied. In general, finding such a guess is equally as difficult as finding an optimal solution. Fortunately, computation of valid initial values can be done by the simplex itself.

Let us assume that the right-hand sides in (8) are non-negative which can be easily achieved by multiplying the corresponding equations by -1. Then, we introduce $n_c(n_c - 1)/2$ additional artificial (or logical) variables y_{ij} such that

$$a_1 \mathbb{D}_{1i}\mathbb{D}_{1j} + a_2 \mathbb{D}_{2i}\mathbb{D}_{2j} + (a_3 - a_4)\big(\mathbb{D}_{1i}\mathbb{D}_{2j} + \mathbb{D}_{2i}\mathbb{D}_{1j}\big) + s_{ij} + y_{ij} = -\mathbb{W}_{ij}^{(0)}. \tag{10}$$

This transformation gives equivalent equations only if $y_{ij} = 0$. To find such non-negative solution, we consider an auxiliary optimization problem:

$$\max_{a_i,\, s_{ij},\, s_i,\, y_{ij} \geq 0} \Psi, \quad \Psi = -\sum_{i<j} y_{ij}$$

subject to constraints (9) and (10). The maximum of this functional on a set of non-negative solutions is obviously zero. For this auxiliary functional it is easy to find a valid initial guess by setting $a_i = s_{ij} = 0$ and calculating y_{ij}, s_i from (9) and (10).

If the auxiliary problem does not have a solution, the original problem has no valid initial guess and an M-matrix $\mathbb{W}$ does not exist.

Remark 1 In a computer program, the artificial variables y_{ij} have to be introduced only when $\mathbb{W}_{ij}^{(0)} > 0$.

The simplex method performs linear operations on the set of linear constraints (9), (10) plus the linear objective functional (first Ψ, then Φ) using certain rules [6]. Each transformation does not decrease the value of the objective functional.

Various generalizations. The simplex method can be also applied to the nodal mimetic schemes [8, Chap. 6], In the case of parabolic equations, positive terms from a time discretization relax the positive definiteness conditions of type (9) which leads to a larger feasible set. Additional computational efficiency can be achieved by combining the simplex method with the primal-dual interior point method [6].

4 Numerical Analysis

In the numerical experiments we used the algorithm `simplx` from [7] with a few modifications that happened to be critical for meshes with flat cells typically used in porous media applications. Specifically, we changed the pivot rule and enforced stability with respect to round-off errors. We verified that the proposed algorithm returns diagonal mass matrix $\mathbb{W}_c$ for a Voronoi cell and a scalar diffusion tensor.

For time-dependent simulations that require to generate matrices $\mathbb{W}_c$ on each time step, complexity of the numerical scheme cannot be ignored. In our experiments with large physics codes, the simplex method has been reaching an optimal solution in something between $n_c^2/2$ and n_c^2 pivot steps. We have not yet met the worst-case scenario shown in Fig. 1.

We illustrate complexity of the simplex method with two experiments. In the first experiment, we take the unit square and the shape-regular pentagon with diameter 2.51 shown in Fig. 2 and change randomly positions of their vertices. This simulates a mechanical deformation of porous media, e.g. due to a land subsidence. The perturbation changes each vertex coordinate by 0.2ξ where $-1 \le \xi \le 1$ is a random function. In the second experiment, we fix the shape of cells shown in Fig. 2, plus the unit square, and rotate gradually the anisotropic diffusion tensor $\mathbb{K}_c = \mathrm{diag}\{1, 3\}$ in 2D and $\mathbb{K}_c = \mathrm{diag}\{1, 2, 3\}$ in 3D about the z-axis. This simulates a change of dispersion tensor, e.g. due to pumping in or out of a subsurface reservoir.

In both experiments, the CPU times are averaged over 1000 different realizations. The results presented in Table 1 show that the calculation of an M-matrix $\mathbb{W}_c$ is 3–6 times more expensive than the calculation based on the original formula $\mathbb{W}_c^{(1)} = \alpha_c \mathbb{D}_c \mathbb{D}_c^T$. On the other hand, the optimal M-matrix contains on average 40 % zero entries which has a few interesting implications for multigrid solvers. The performance of the multigrid solvers is near-optimal for M-matrices and our

Table 1 Complexity of a single matrix generation in microseconds for random perturbation (columns 2 and 3) and tensor rotation (columns 4 and 5)

Cell type	Monotone MFD	Original MFD	Monotone MFD	Original MFD
Quad	15.3	5.05	14.7	4.91
Pentagon	28.0	6.62	29.3	6.64
Hexahedron	–	–	48.7	8.92

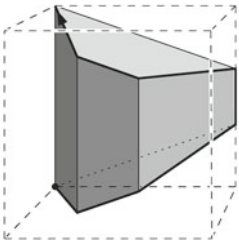


Fig. 1 The worst-case scenario. The set of feasible solutions forms a Klee Minty cube. The Dantzig's simplex method initialized at a vertex of this cube passes through all its vertices making exponentially many pivot steps

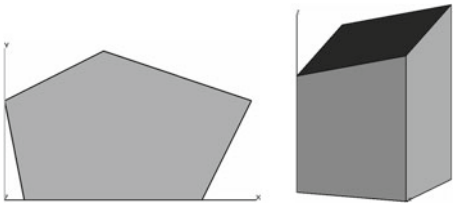


Fig. 2 The cells used in the numerical experiments: pentagon and hexahedron with planar faces

preliminary experiments indicate reduction of the cost of one V-cycle which offsets a bit the higher complexity of the proposed mimetic scheme.

Due to page limitation, experiments showing advantage of the simplex method in modeling dispersive transport in porous media will be presented at the conference.

Acknowledgments This work was performed under the auspices of the National Nuclear Security Administration of the US Department of Energy at Los Alamos National Laboratory under Contract No. DE-AC52-06NA25396. The author gratefully acknowledges the partial support of the US Department of Energy Office of Science Advanced Scientific Computing Research (ASCR) Program in Applied Mathematics Research and Office of Environmental Management Advanced Simulation Capability for Environmental Management (ASCEM) Program.

References

1. Brezzi, F., Lipnikov, K., Simoncini, V.: A family of mimetic finite difference methods on polygonal and polyhedral meshes. Math. Models Methods Appl. Sci. **15**(10), 1533–1551 (2005)
2. Ciarlet, P., Raviart, P.A.: Maximum principle and uniform convergence for the finite element method. Comput. Meth. Appl. Mech. Eng. **2**(1), 17–31 (1973)

3. Droniou, J., Eymard, R., Gallouet, T., Herbin, R.: A unified approach to mimetic finite difference, hybrid finite volume and mixed finite volume method. Math. Model. Methods Appl. Sci. **20**(2), 1–31 (2010)
4. Lipnikov, K., Manzini, G., Shashkov, M.: Mimetic finite difference method. J. Comput. Phys. **257**(Part-B), 1163–1228 (2014)
5. Lipnikov, K., Manzini, G., Svyatskiy, D.: Analysis of the monotonicity conditions in the mimetic finite difference method for elliptic problems. J. Comput. Phys. **230**, 2620–2642 (2011)
6. Matoušek, J., Gärtner, B.: Understanding and using linear programming, p. 228. Springer, Berlin (2007)
7. Press, W., Teukolsky, S., Vetterling, W., Flannery, B.: Numerical recipes in C, p. 949. Cambridge University Press, Cambridge New York Port Chester Melbourne Sydney (1992)
8. Beirao da Veiga, L., Lipnikov, K., Manzini, G.: Mimetic finite difference method for elliptic problems. Modeling, Simulation & Applications, vol. 11, p. 408. Springer, Berlin (2014)

Discrete Relative Entropy for the Compressible Stokes System

Thierry Gallouët, David Maltese and Antonín Novotný

Abstract In this paper, we propose a discretization for the nonsteady compressible Stokes Problem. This scheme is based on Crouzeix-Raviart approximation spaces. The discretization of the momentum balance is obtained by the usual finite element technique. The discrete mass balance is obtained by a finite volume scheme, with an upwinding of the density. The time discretization will be implicit in time. We prove the existence of a discrete solution. We prove that our scheme satisfies a discrete version of the relative entropy. As a consequence, we obtain an error estimate for this system. This preliminary work will be used in order to obtain a error estimate for the compressible Navier-Stokes system and has to the author's knowledge not been studied previously.

1 Introduction

Let Ω an open bounded domain with lipschitz boundary subset of $\mathbb{R}^d$, $d = 2, 3$. We consider the following system

$$\partial_t \varrho + \operatorname{div}(\varrho \boldsymbol{u}) = 0,\ t \in (0, T),\ x \in \Omega \tag{1}$$

$$\partial_t \boldsymbol{u} - \mu \Delta \boldsymbol{u} - (\mu + \lambda) \nabla \operatorname{div} \boldsymbol{u} + \nabla_x p(\varrho) = \mathbf{0},\ t \in (0, T),\ x \in \Omega \tag{2}$$

T. Gallouët (✉)
L.A.T.P., UMR 6632, Universite de Provence, Marseille cedex 13,
13453 Aix-en-provence, France
e-mail: thierry.gallouet@univ-amu.fr

D. Maltese · A. Novotný
IMATH, EA 2134, Universite du Sud Toulon-Var, BP 20132, 83957 La Garde, France
e-mail: david.maltese@univ-tln.fr

A. Novotný
e-mail: novotny.an@gmail.com

J. Fuhrmann et al. (eds.), *Finite Volumes for Complex Applications VII - Methods and Theoretical Aspects*, Springer Proceedings in Mathematics & Statistics 77,
DOI: 10.1007/978-3-319-05684-5_37,

supplemented with the following initial conditions and boundary condition

$$\varrho(0,x)=\rho_0(x),\ \boldsymbol{u}(0,x)=\boldsymbol{u}_0,\ \boldsymbol{u}_{|\partial\Omega}=0. \tag{3}$$

We suppose that the pressure satisfies $p\in C(\mathbb{R}_+)\cap C^2(\mathbb{R}_+^*)$, $p(0)=0$ and $\lim_{+\infty}\frac{p'(\rho)}{\rho^{\gamma-1}}=p_\infty>0$ for $\gamma\geq 2$. Moreover if $\gamma\in[\frac{6}{5},2[$ we suppose also that $\liminf_0\frac{p'(\rho)}{\rho^{\alpha-1}}=p_0>0$, with $\alpha\leq 0$.

2 Weak Solutions, Relative Entropies

In this part, we give the definition of (finite energy) weak solutions for our system. We give the definition of the relative entropy. In the following we denote $\mathcal{H}(\varrho)=\rho\int_1^\rho\frac{p(t)}{t^2}\,\mathrm{dt}$. Let us denote $C_c^\infty([0,T]\times\Omega,\mathbb{R}^3)$ the space of all smooth functions on $[0,T]\times\Omega$ compactly supported in $[0,T]\times\Omega$.

Definition 1 Let $(\varrho_0,\boldsymbol{u}_0)\in L^\gamma(\Omega)\times H_0^1(\Omega)$ such that $\varrho_0\geq 0$ a.e in Ω. We shall say that $(\varrho,\boldsymbol{u})$ is a finite energy weak solution to the problem (1)–(3) emanating from the initial data $(\varrho_0,\boldsymbol{u}_0)$ if

$$\varrho\in L^\infty(0,T;L^\gamma(\Omega))\cap C_w([0,T],L^\gamma(\Omega)),\ \rho\geq 0\ p.p \text{ in } (0,T)\times\Omega,$$

$$\boldsymbol{u}\in L^2(0,T;H_0^1(\Omega))\cap C_w([0,T],L^2(\Omega))$$

and :

– The continuity equation (1) is satisfied in the following weak sense

$$\int_\Omega\varrho(\tau,\cdot)\varphi(\tau,\cdot)\,\mathrm{dx}-\int_\Omega\varrho_0\varphi(0,\cdot)=\int_0^\tau\int_\Omega\varrho(t,x)\partial_t\varphi(t,x)\,\mathrm{dx}\,\mathrm{dt}+\int_0^\tau\int_\Omega\varrho\boldsymbol{u}\cdot\nabla_x\varphi\,\mathrm{dx}\,\mathrm{dt}, \tag{4}$$

$\forall\tau\in[0,T],\ \forall\varphi\in C^\infty([0,T]\times\overline{\Omega})$.

– The momentum equation (2) is satisfied in the following weak sense

$$\int_\Omega\boldsymbol{u}\cdot\psi(\tau,x)\,\mathrm{dx}-\int_\Omega\mathbf{u}_0\cdot\psi(0,\cdot)\,\mathrm{dx}=\int_0^\tau\int_\Omega\boldsymbol{u}\cdot\partial_t\psi+p(\varrho)\,\mathrm{div}_x\,\psi-\mu\nabla_x\boldsymbol{u}:\nabla_x\psi-(\mu+\lambda)\,\mathrm{div}_x\,\boldsymbol{u}\,\mathrm{div}_x\,\psi\,\mathrm{dx}\,\mathrm{dt}, \tag{5}$$

$\forall\tau\in[0,T],\ \forall\psi\in C_c^\infty([0,T]\times\Omega,\mathbb{R}^3)$.

– The following energy inequality is satisfied

$$\int_\Omega \frac{1}{2}|\boldsymbol{u}|^2+\mathcal{H}(\varrho)\,\mathrm{dx}+\int_0^\tau\int_\Omega \mu||\nabla_x\boldsymbol{u}||^2+(\mu+\lambda)(\mathrm{div}_x\,\boldsymbol{u})^2\,\mathrm{dx}\,\mathrm{dt}$$
$$\leq \int_\Omega \frac{1}{2}|\boldsymbol{u}_0|^2+\mathcal{H}(\varrho_0)\,\mathrm{dx}, \tag{6}$$

a.e $\tau\in[0,T]$.

2.1 Relative Entropy Inequality, Weak-Strong Uniqueness

The method of relative entropy has been successfully applied to partial differential equations of different types. Relative entropies are non-negative quantities that provide a kind of distance between two solutions of the same problem, one of which typically enjoys some extra regularity properties (see [2] for more details)

Definition 2 We define the relative entropy of $(\rho,\boldsymbol{u})$ with respect to $(r,\boldsymbol{U})$ by

$$\mathcal{E}([\varrho,\boldsymbol{u}],[r,\boldsymbol{U}])=\int_\Omega \frac{1}{2}|\boldsymbol{u}-\boldsymbol{U}|^2+E(\varrho,r)\,\mathrm{dx} \tag{7}$$

where $E(\rho,r)=\mathcal{H}(\rho)-\mathcal{H}'(r)(\rho-r)-\mathcal{H}(r)$. We also define a remainder, denoted by $\mathcal{R}$, as

$$\mathcal{R}=\int_\Omega \nabla_x\boldsymbol{U}:\nabla_x(\boldsymbol{U}-\boldsymbol{u})\,\mathrm{dx}+\int_\Omega(r-\varrho)\partial_t\mathcal{H}'(r)+\nabla_x\mathcal{H}'(r)\cdot(r\boldsymbol{U}-\varrho\boldsymbol{u})\,\mathrm{dx}$$
$$-\int_\Omega \mathrm{div}_x\,\boldsymbol{U}(p(\varrho)-p(r))\,\mathrm{dx}+\int_\Omega\partial_t\boldsymbol{U}\cdot(\boldsymbol{U}-\boldsymbol{u})\,\mathrm{dx}\,. \tag{8}$$

Theorem 1 *Let* $(\rho,\boldsymbol{u})$ *be a weak solution of (1)– (3) in the sense of the definition 1 emanating from the initial condition* $(\rho_0,\boldsymbol{u}_0)$. *Then* $(\rho,\boldsymbol{u})$ *satisfy the relative energy inequality:*

$$\mathcal{E}([\varrho,\boldsymbol{u}],[r,\boldsymbol{U}])(\tau)+\int_0^\tau\int_\Omega \mu||\nabla_x(\boldsymbol{u}-\boldsymbol{U})||^2+(\mu+\lambda)(\mathrm{div}_x(\boldsymbol{u}-\boldsymbol{U}))^2\,\mathrm{dx}\,\mathrm{dt}$$
$$\leq \mathcal{E}([\varrho_0,\boldsymbol{u}_0],[r(0),\boldsymbol{U}(0)])+\int_0^\tau \mathcal{R}([\varrho,\boldsymbol{u}],[r,\boldsymbol{U}])(t)\,\mathrm{dt} \tag{9}$$

a.e $\tau\in[0,T]$, *where* $r\in C^\infty([0,T]\times\overline{\Omega},\mathbb{R}^*_+)$ *and* $\boldsymbol{U}\in C^\infty([0,T]\times\Omega,\mathbb{R}^3)$.

Proof See [2].

Remark 1 For the choice of $r = \overline{\rho}$ and $\boldsymbol{U} = 0$, the relative energy inequality (9) reduces to the standard energy inequality.
Moreover, the relative energy inequality can be used to show that suitable weak solutions comply with the weak-strong uniqueness principle, meaning, a weak and strong solution emanating from the same initial data coincide as long as the latter exists. This can be seen by taking the strong solution as the test functions r, U in the relative entropy inequality (see [2]).

3 The Numerical Scheme

Now suppose that Ω is a bounded open set of $\mathbb{R}^d$, polygonal if $d = 2$ and polyhedral if $d = 3$. Let $\mathcal{T}$ be a decomposition of the domain Ω in simplices, which we call hereafter a triangulation of Ω, regardless of the space dimension. By $\mathcal{E}(K)$, we denote the set of the edges (d $= 2$) or faces (d $= 3$) σ of the elements $K \in \mathcal{T}$; for short, each edge or face will be called an edge hereafter. The set of all edges of the mesh is denoted by $\mathcal{E}$; the set of edges included in the boundary of Ω is denoted by $\mathcal{E}_{\text{ext}}$ and the set of internal edges (i.e $\mathcal{E} \setminus \mathcal{E}_{\text{ext}}$) is denoted by $\mathcal{E}_{\text{int}}$. The decomposition $\mathcal{T}$ is assumed to be regular in the usual sense of the finite element literature, and, in particular, $\mathcal{T}$ satisfies the following properties: $\overline{\Omega} = \cup_{K\in\mathcal{T}} \overline{K}$; if $K, L \in \mathcal{T}$, then $\overline{K} \cap \overline{L} = \emptyset$, $\overline{K} \cap \overline{L}$ is a vertex or $\overline{K} \cap \overline{L}$ is a common edge of K and L, which is denoted by $K|L$. For $K \in \mathcal{T}$ and $\sigma \in \mathcal{E}(K)$, we define $D_{K,\sigma}$ as the cone with basis σ and with vertex the mass center of K. For each internal edge of the mesh $\sigma = K|L$, $\boldsymbol{n}_{KL}$ stands for the unit normal vector of σ, oriented form K to L (so that $\boldsymbol{n}_{KL} = -\boldsymbol{n}_{LK}$). By $|K|$ and $|\sigma|$ we denote the (d and $d-1$ dimensional) measure, respectively, of an element K and of an edge σ, and h_K and h_σ stand for the diameter of K and σ, respectively. We measure the regularity of the mesh through the parameter θ defined by:

$$\theta = \inf\{\frac{\xi_K}{h_K}, K \in \mathcal{T}\} \tag{10}$$

where ξ_K stands for the diameter of the largest ball included in K. The space discretization relies on the Crouzeix-Raviart element. The reference element is the unit d-simplex and the discrete functional space is the space P_1 of affine polynomials. The degrees of freedom are determined by the following set of edge functionals:

$$\{F_\sigma, \sigma \in \mathcal{E}(K)\},\ F_\sigma(v) = \frac{1}{|\sigma|}\int_\sigma v \,\mathrm{d}\gamma\,.$$

The mapping from the reference element to the actual one is the standard affine mapping. Finally, the continuity of the average value of a discrete function v across each edge of the mesh, $F_\sigma(v)$, is required, thus the discrete space V_h is defined as follows:

$$V_h = \{v \in L^2(\Omega),\ \forall K \in \mathcal{T},\ v_{|K} \in \mathbb{P}_1(K) \text{ and } \forall \sigma \in \mathcal{E}_{\text{int}}, \sigma = K|L,\ F_\sigma(v_{|K}) = F_\sigma(v_{|L}), \forall \sigma \in \mathcal{E}_{\text{ext}},\ F_\sigma(v) = 0\}.$$

The space of approximation for the velocity is the space $\boldsymbol{W}_h$ of vector-valued functions each component of which belongs to V_h : $\boldsymbol{W}_h = (V_h)^d$. The pressure and the density are approximated by the space L_h of piecewise constant functions:

$$L_h = \{q \in L^2(\Omega), q_{|K} = \text{constant}, \forall K \in \mathcal{T}\}.$$

We will also denote $L_h^+ = \{q \in L_h,\ q_K \geq 0,\ \forall K \in \mathcal{T}\}$ and $L_h^{++} = \{q \in L_h,\ q_K > 0,\ \forall K \in \mathcal{T}\}$.

It is well-know that this discretization is nonconforming in $H^1(\Omega)^d$. We then define, for $1 \leq i \leq d$ and $u \in V_h$, $\partial_{h,i} u$ as the function of $L^2(\Omega)$ which is equal to the derivative of u with respect to the ith space variable almost everywhere. This notation allows us to define the discrete gradient, denoted by ∇_h for both scalar and vector-valued discrete functions and the discrete divergence of vector-valued discrete functions, denoted by div_h. We denote $||\cdot||_{1,b}$ the broken Sobolev H^1 semi-norm, which is defined for scalar as well as for vector-valued functions by

$$||v||_{1,b}^2 = \sum_{K \in \mathcal{T}} \int_K |\nabla v|^2 \, \mathrm{dx} = \int_\Omega |\nabla_h v|^2 \, \mathrm{dx}\,.$$

We denote by $\{u_{i,\sigma}, \sigma \in \mathcal{E}_{\text{int}}, 1 \leq i \leq d\}$ the set of velocity degrees of freedom We denote by φ_σ the usual Crouzeix-Raviart shape function associated to $\sigma \in \mathcal{E}_{\text{int}}$, i.e. the scalar function of V_h such that $F_\sigma(\varphi_\sigma) = 1$ and $F_{\sigma'}(\varphi_\sigma) = 0, \forall \sigma' \neq \sigma$.

Similarly, each degree of freedom for the density is associated to a cell K, and the set of density degrees of freedom is denoted by $\{\rho_K, K \in \mathcal{T}\}$. We define by r_h the following interpolation operator $r_h : H_0^1(\Omega) \to V_h$ by

$$r_h(v) = \sum_{\sigma \in \mathcal{E}_{\text{int}}} F_\sigma(v)\varphi_\sigma.$$

This operator naturally extends to vector-valued functions and we keep the same notation r_h for both the scalar and vector case.

Let us consider a partition $0 = t^0 < t^1 < \ldots < t^N = T$ of the time interval $[0, T]$, which, for the sake of simplicity, we suppose uniform. Let Δt be the constant time step $\Delta t = t^n - t^{n-1}$ for $n = 1, \ldots, N$. Let $(\rho^0, \boldsymbol{u}^0) \in L_h \times \boldsymbol{W}_h$.

Following [6] we consider the following numerical scheme :

Find $(\varrho^n)_{1 \leq n \leq N} \subset L_h$, $(\boldsymbol{u}^n)_{1 \leq n \leq N} \subset \mathbf{W}_h$ *such that* $\forall n = 1, \ldots, N$

$$|K| \frac{\varrho_K^n - \varrho_K^{n-1}}{\Delta t} + \sum_{\sigma \in \mathcal{E}(K), \sigma = K|L} |\sigma| \Big(\boldsymbol{u}_\sigma^n \cdot \boldsymbol{n}_{KL}\Big)^+ \rho_K^n - |\sigma| \Big(\boldsymbol{u}_\sigma^n \cdot \boldsymbol{n}_{KL}\Big)^-$$
$$\rho_L^n = 0, \forall K \in \mathcal{T} \tag{11}$$

$$\frac{|D_\sigma|}{\Delta t}(u_{i,\sigma}^n - u_{i,\sigma}^{n-1}) + \mu \sum_{K\in\mathcal{T}} \int_K \nabla u_i^n \cdot \nabla \varphi_\sigma \,\mathrm{dx} + (\mu+\lambda) \sum_{K\in\mathcal{T}} \int_K \mathrm{div}(\boldsymbol{u}^n)\,\mathrm{div}(\varphi_\sigma \boldsymbol{e}_i)\,\mathrm{dx}$$
$$- \sum_{K\in\mathcal{T}} \int_K p_K^n \,\mathrm{div}(\varphi_\sigma \boldsymbol{e}_i)\,\mathrm{dx} = 0, \forall \sigma \in \mathcal{E}_{int}, 1 \le i \le d \qquad (12)$$

with $p_K^n = p(\rho_K^n)$, $a^+ = \max(a, 0)$, $a^- = -\min(a, 0)$.

As usual, to the discrete unknowns, we associate piecewise constant functions on time intervals and on primal or dual meshes, so the density $\rho_{\Delta t,h}$, the pressure $p_{\Delta t,h}$ and the velocity $\boldsymbol{u}_{\Delta t,h}$ are defined almost everywhere on $(0, T) \times \Omega$ by

$$\varrho_{\Delta t,h}(t,x) = \sum_{n=1}^{N} \sum_{K\in\mathcal{T}} \varrho_K^n \mathbf{1}_{(t^{n-1},t^n)} \mathbf{1}_K, \quad \rho_{\Delta t,h}(t,x) = \sum_{n=1}^{N} \sum_{K\in\mathcal{T}} \rho_K^n \mathbf{1}_{(t^{n-1},t^n)} \mathbf{1}_K,$$

$$\boldsymbol{u}_{\Delta t,h}(t,x) = \sum_{n=1}^{N} \sum_{K\in\mathcal{T}} \boldsymbol{u}_\sigma^n \mathbf{1}_{(t^{n-1},t^n)} \mathbf{1}_{D_\sigma}.$$

3.1 Existence, Positivity and Stabilities Properties

Theorem 2 (Existence and positivity) *Let* $(\rho^0, \boldsymbol{u}^0) \in L_h^{++} \times \boldsymbol{W}_h$. *Then the problem (11), (12) admits at least a solution* $(\varrho^n)_{1\le n\le N} \subset L_h^{++}$, $(\boldsymbol{u}^n)_{1\le n\le N} \subset \boldsymbol{W}_h$.

Proof See [5].

Theorem 3 (Energy estimate) *Let* $(\varrho_0, \boldsymbol{u}_0) \in L^\gamma(\Omega) \times H_0^1(\Omega, \mathbb{R}^3)$, *such that* $\varrho_0 > 0$ *a.e* $x \in \Omega$.

Let $\varrho_K^0 = \frac{1}{|K|}\int_K \varrho_0 \,\mathrm{dx}$ *and* $\boldsymbol{u}^0 = r_h(\boldsymbol{u}_0)$.

Let $(\varrho^n, \boldsymbol{u}^n) \in L_h^{++} \times \boldsymbol{W}_h, n = 1, ..., N$ *be a solution of (11), (12) emanating from the initial data* $(\varrho^0, \boldsymbol{u}^0)$. *Then we have the following balance discrete energy*

$$\max_{n=0,\dots,N} \sum_{K\in\mathcal{T}} |K|\,\mathcal{H}(\varrho_K^n) + \max_{n=0,\dots,N} \sum_{i,\sigma\in\mathcal{E}_{\mathrm{int}}} \frac{1}{2}|D_\sigma|(u_{i,\sigma}^n)^2 + \mu\Delta t \sum_{n=0}^{N} ||\boldsymbol{u}^n||_{1,b}^2$$
$$+ (\mu+\lambda)\Delta t \sum_{k=0}^{N} ||\,\mathrm{div}_h\,\boldsymbol{u}^n||_{L^2(\Omega)}^2 \le c(d, \theta_0, \varrho_0, \boldsymbol{u}_0),$$

Proof See [5].

3.1.1 Discrete Relative Entropy Inequality

The following result is crucial for the rest of the article. It can be seen as a discrete balance version of (9).

Theorem 4 *Let* $(\varrho_0, \boldsymbol{u}_0) \in L^\gamma(\Omega) \times H_0^1(\Omega, \mathbb{R}^3)$, *such that* $\varrho_0(x) > 0$ *a.e* $x \in \Omega$ *and* $\mathcal{H}(\varrho_0) \in L^1(\Omega)$.

Let $\varrho_K^0 = \frac{1}{|K|}\int_K \varrho_0 \,\mathrm{dx}$ *and* $\boldsymbol{u}^0 = r_h(\boldsymbol{u}_0)$.

Let $(\varrho^n, \boldsymbol{u}^n) \in L_h \times \boldsymbol{W}_h, n = 1, ..., N$ *be a solution of (11), (12) emanating from the initial data* $(\rho^0, \boldsymbol{u}^0)$. *Let* $(r, \boldsymbol{U}) \in C^1([0,T] \times \overline{\Omega}) \cap C^2([0,T] \times \overline{\Omega}, \mathbb{R}^3)$ *such that* $r(t,x) > 0, \forall (t,x) \in [0,T] \times \overline{\Omega}$ *and* $\boldsymbol{U}(t)_{|\partial\Omega} = \boldsymbol{0}$. *Let* $\boldsymbol{U}_h^n = r_h(\boldsymbol{U}(t^n))$, $r_K^n = \frac{1}{|K|}\int_K r(t^n, x)\,\mathrm{dx}$ *Then we have the following inequality*

$$
\begin{aligned}
&\sum_{i,\sigma\in\mathcal{E}_{\mathrm{int}}} \frac{1}{2}\frac{|D_\sigma|}{\Delta t}\Big((u_{i,\sigma}^n - U_{i,\sigma}^n)^2 - (u_{i,\sigma}^{n-1} - U_{i,\sigma}^{n-1})^2\Big) \\
&+ \sum_{K\in\mathcal{T}} \frac{|K|}{\Delta t}\Big(E(\varrho_K^n | r_K^n) - E(\varrho_K^{n-1} | r_K^{n-1})\Big) \\
&+ \mu ||\boldsymbol{u}^n - \boldsymbol{U}^n||_{1,b}^2 + (\mu+\lambda)||\operatorname{div}_h(\boldsymbol{u}^n - \boldsymbol{U}_h^n)||_{L^2(\Omega)}^2 \\
&\le \sum_{i,\sigma\in\mathcal{E}_{\mathrm{int}}} \frac{|D_\sigma|}{\Delta t}(U_{i,\sigma}^n - u_{i,\sigma}^n)(U_{i,\sigma}^n - U_{i,\sigma}^{n-1}) + \mu \sum_{K\in\mathcal{T}} \int_K \nabla \boldsymbol{U}_h^n : \nabla(\boldsymbol{U}_h^n - \boldsymbol{u}^n)\,\mathrm{dx} \\
&+ (\mu+\lambda)\int_\Omega \operatorname{div}_h \boldsymbol{U}_h^n \operatorname{div}_h(\boldsymbol{U}_h^n - \boldsymbol{u}^n)\,\mathrm{dx} + \sum_{K\in\mathcal{T}} \operatorname{div}_K^{\mathrm{up}}(\varrho^n \boldsymbol{u}^n)\mathcal{H}'(r_K^n) \\
&+ \sum_{K\in\mathcal{T}} \frac{|K|}{\Delta t}(r_K^n - \rho_K^n)(\mathcal{H}'(r_K^n) - \mathcal{H}'(r_K^{n-1})) - \int_\Omega p^n \operatorname{div} \boldsymbol{U}_h^n \,\mathrm{dx} + \mathcal{R}^{n,h} \qquad (13)
\end{aligned}
$$

where $\Delta t \sum_{n=1}^N |\mathcal{R}^{n,h}| \le c(\varrho_0, \boldsymbol{u}_0, r, \boldsymbol{U})\Delta t$.

The following result is the main result of our article and it is a consequence of the previous. We give an error estimate for our system.

Theorem 5 *Let* $(\varrho_0, \boldsymbol{u}_0) \in L^\gamma(\Omega) \times H_0^1(\Omega, \mathbb{R}^3)$, *such that* $\varrho_0(x) > 0$ *a.e* $x \in \Omega$ *and* $\mathcal{H}(\varrho_0) \in L^1(\Omega)$.

Let $\varrho_K^0 = \frac{1}{|K|}\int_K \varrho_0 \,\mathrm{dx}$ *and* $\boldsymbol{u}^0 = r_h(\boldsymbol{u}_0)$.

Let $(\varrho^n, \boldsymbol{u}^n) \in L_h \times \boldsymbol{W}_h, n = 1, ..., N$ *be a solution of (11), (12) emanating from the initial data* $(\varrho^0, \boldsymbol{u}^0)$. *Let* $(r, \boldsymbol{U}) \in C^1([0,T] \times \overline{\Omega}) \cap C^2([0,T] \times \overline{\Omega}, \mathbb{R}^3)$ *be a strong solution of (1)–(3) such that* $\forall (t,x) \in [0,T] \times \overline{\Omega}, r(t,x) > 0$. *Let* $\boldsymbol{U}_h^n = r_h(\boldsymbol{U}(t^n))$, $r_K^n = \frac{1}{|K|}\int_K r(t^n, x)\,\mathrm{dx}$. *Then we have the following inequality*

$$
\begin{aligned}
&\sum_{i,\sigma\in\mathcal{E}_{\rm int}} \frac{1}{2}\frac{|D_\sigma|}{\Delta t}\Big((u_{i,\sigma}^n - U_{i,\sigma}^n)^2 - (u_{i,\sigma}^{n-1} - U_{i,\sigma}^{n-1})^2\Big)\\
&\quad + \sum_{K\in\mathcal{T}} \frac{|K|}{\Delta t}\Big(E(\varrho_K^n|r_K^n) - E(\varrho_K^{n-1}|r_K^{n-1})\Big)\\
&\quad + \mu||\boldsymbol{u}^n - \boldsymbol{U}^n||_{1,b}^2 + (\mu+\lambda)||\operatorname{div}_h(\boldsymbol{u}^n - \boldsymbol{U}_h^n)||_{L^2(\Omega)}^2\\
&\quad \le \sum_{K\in\mathcal{T}} (r_K^n - \varrho_K^n)\int_K \frac{\nabla p(r^n)}{r^n}\cdot(\boldsymbol{u}^n - \boldsymbol{U}_h^n)\,\mathrm{dx}\\
&\quad - \sum_{K\in\mathcal{T}}\int_K \Big(p^n - p'(r_K^n)(\varrho_K^n - r_K^n) - p(r_K^n)\Big)\operatorname{div}\boldsymbol{U}_h^n\,\mathrm{dx}\\
&\quad + \mathcal{R}^{n,h}
\end{aligned}
\tag{14}
$$

where $\Delta t\sum_{n=1}^N |\mathcal{R}^{n,h}| \le C(\theta_0,\varrho_0,\boldsymbol{u}_0)(h^{\epsilon(\gamma)}+\Delta t)$ *with* $\epsilon(\gamma)=\frac{1}{2}$ *for* $\gamma\ge\frac{3}{2}$ *and* $\epsilon(\gamma)=\frac{5}{2}-\frac{3}{\gamma}$ *for* $\gamma\in[\frac{6}{5},\frac{3}{2}]$, *and we obtain the following estimation* $||\boldsymbol{u}_{\delta t,h}-\boldsymbol{U}||_{L^\infty(0,T;L^2(\Omega))}^2 + ||\varrho_{\delta t,h}-r||_{L^\infty(0,T,L^\gamma(\Omega))}^\gamma \le C(\theta_0,\varrho_0,\boldsymbol{u}_0)(h^{\epsilon(\gamma)}+\Delta t)$.

Proof We begin with a algebraic inequality whose straightforward proof is left to the reader

Lemma 1 *Let* $0<a<b<\infty$. *Then there exists* $c=c(a,b)>0$ *such that for all* $\rho\in[0,\infty[$ *and* $r\in[a,b]$ *there holds*

$$
E(\rho|r)\ge c(a,b)\Big(1_{[\frac{a}{2},2b]} + \rho^\gamma 1_{\mathbb{R}_+\setminus[\frac{a}{2},2b]} + (\rho-r)^2 1_{\mathbb{R}_+\setminus[\frac{a}{2},2b]}\Big). \tag{15}
$$

We return to (14). We set $a=\min_{[0,T]\times\overline{\Omega}} r$ and $b=\max_{[0,T]\times\overline{\Omega}} r$. We write

$$
\begin{aligned}
&\sum_{K\in\mathcal{T}}\int_K \Big(p^n - p'(r_K^n)(\rho_K^n - r_K^n) - p(r_K^n)\Big)\operatorname{div}\boldsymbol{U}_h^n\,\mathrm{dx}\\
&\quad = \sum_{K,\rho_K^n\in[a/2,2b]}\int_K \Big(p^n - p'(r_K^n)(\rho_K^n - r_K^n) - p(r_K^n)\Big)\operatorname{div}\boldsymbol{U}_h^n\,\mathrm{dx}\\
&\qquad + \sum_{K,\rho_K^n\in\mathbb{R}_+\setminus[a/2,2b]}\int_K \Big(p^n - p'(r_K^n)(\rho_K^n - r_K^n) - p(r_K^n)\Big)\operatorname{div}\boldsymbol{U}_h^n\,\mathrm{dx}
\end{aligned}
$$

Now using the behavior of p as ρ goes to infinity and (15) we obtain

$$
|\sum_{K\in\mathcal{T}}\int_K \Big(p^n - p'(r_K^n)(\rho_K^n - r_K^n) - p(r_K^n)\Big)\operatorname{div}\boldsymbol{U}_h^n\,\mathrm{dx}| \le c(r,\boldsymbol{U})\sum_{K\in\mathcal{T}}|K|E(\rho_K^n|r_K^n)
$$

We write

$$\begin{aligned}
&\sum_{K\in\mathcal{T}}(r_K^n-\rho_K^n)\int_K\frac{\nabla p(r^n)}{r^n}\cdot(\boldsymbol{u}^n-\boldsymbol{U}_h^n)\,\mathrm{dx}\\
&\quad=\sum_{\rho_K^n<\frac{a}{2}}(r_K^n-\rho_K^n)\int_K\frac{\nabla p(r^n)}{r^n}\cdot(\boldsymbol{u}^n-\boldsymbol{U}_h^n)\,\mathrm{dx}\\
&\quad+\sum_{\rho_K^n\in[\frac{a}{2},2b]}(r_K^n-\rho_K^n)\int_K\frac{\nabla p(r^n)}{r^n}\cdot(\boldsymbol{u}^n-\boldsymbol{U}_h^n)\,\mathrm{dx}\\
&\quad+\sum_{\rho_K^n>2b}(r_K^n-\rho_K^n)\int_K\frac{\nabla p(r^n)}{r^n}\cdot(\boldsymbol{u}^n-\boldsymbol{U}_h^n)\,\mathrm{dx}\,.
\end{aligned}$$

Using (15) and Poincare's inequality we obtain $\forall\delta>0$,

$$\begin{aligned}
&|\sum_{\rho_K^n<\frac{a}{2}}(r_K^n-\rho_K^n)\int_K\frac{\nabla p(r^n)}{r^n}\cdot(\boldsymbol{u}^n-\boldsymbol{U}_h^n)\,\mathrm{dx}\,|\\
&\quad\le c(r,\delta)\sum_{K\in\mathcal{T}}|K|E(\rho_K^n|r_K^n)+\delta||\boldsymbol{u}^n-\boldsymbol{U}_h^n||_{1,b}^2,\\
&|\sum_{\rho_K^n\in[\frac{a}{2},2b]}(r_K^n-\rho_K^n)\int_K\frac{\nabla p(r^n)}{r^n}\cdot(\boldsymbol{u}^n-\boldsymbol{U}_h^n)\,\mathrm{dx}\,|\\
&\quad\le c(r,\delta)\sum_{K\in\mathcal{T}}|K|E(\rho_K^n|r_K^n)+\delta||\boldsymbol{u}^n-\boldsymbol{U}_h^n||_{1,b}^2.
\end{aligned}$$

Now we have

$$\sum_{\rho_K^n>2b}|K|(\rho_K^n)^\gamma\le c\sum_{K\in\mathcal{T}}|K|E(\rho_K^n|r_K^n),\qquad\sum_{\rho_K^n>2b}|K|(\rho_K^n)^{\gamma/2}\le c\sum_{K\in\mathcal{T}}|K|E(\rho_K^n|r_K^n)$$

Then,

$$\begin{aligned}
&|\sum_{\rho_K^n>2b}(r_K^n-\rho_K^n)\int_K\frac{\nabla p(r^n)}{r^n}\cdot(\boldsymbol{u}^n-\boldsymbol{U}_h^n)\,\mathrm{dx}\,|\\
&\quad\le c(r)\sum_{\rho_K^n>2b}\max(\rho_K^n,(\rho_K^n)^{\gamma/2})\int_K||\boldsymbol{u}^n-\boldsymbol{U}_h^n||\,\mathrm{dx}\\
&\quad c(r)\sum_{\rho_K^n>2b}\sqrt{|K|}(\rho_K^n)^{\gamma/2}||\boldsymbol{u}^n-\boldsymbol{U}_h^n||_{L^2(K)}\\
&\quad+c(r)\sum_{\rho_K^n>2b}|K|^{1/\gamma}\rho_K^n||\boldsymbol{u}^n-\boldsymbol{U}_h^n||_{L^{\gamma'}(K)}
\end{aligned}$$

$$
\begin{aligned}
&C(r,\delta)\sum_{K\in\mathcal{T}}|K|E(\rho_K^n|r_K^n)+\delta||\boldsymbol{u}^n-\boldsymbol{U}_h^n||_{1,b}^2\\
&+c(r,\delta)\sum_{K\in\mathcal{T}}|K|E(\rho_K^n|r_K^n)+\delta||\boldsymbol{u}^n-\boldsymbol{U}_h^n||_{L^{\gamma'}(\Omega)}^{\gamma'}\\
&\le C(r,\delta)\sum_{K\in\mathcal{T}}|K|E(\rho_K^n|r_K^n)+\delta||\boldsymbol{u}^n-\boldsymbol{U}_h^n||_{1,b}^2\\
&+c(r,\delta)\sum_{K\in\mathcal{T}}|K|E(\rho_K^n|r_K^n)+\delta||\boldsymbol{u}^n-\boldsymbol{U}_h^n||_{L^6(\Omega)}^6\\
&\le C(r,\delta)\sum_{K\in\mathcal{T}}|K|E(\rho_K^n|r_K^n)+\delta||\boldsymbol{u}^n-\boldsymbol{U}_h^n||_{1,b}^2
\end{aligned}
$$

since $\gamma\ge\frac{6}{5}$. We obtain finally

$$
\begin{aligned}
&\sum_{i,\sigma\in\mathcal{E}_{\rm int}}\frac{1}{2}\frac{|D_\sigma|}{\Delta t}\Big((u_{i,\sigma}^n-U_{i,\sigma}^n)^2-(u_{i,\sigma}^{n-1}-U_{i,\sigma}^{n-1})^2\Big)\\
&+\sum_{K\in\mathcal{T}}\frac{|K|}{\Delta t}\Big(E(\rho_K^n|r_K^n)-E(\rho_K^{n-1}|r_K^{n-1})\Big)\\
&\le c(r,\boldsymbol{U},\mu)\Big(\sum_{i,\sigma\in\mathcal{E}_{\rm int}}|D_\sigma|(u_{i,\sigma}^n-U_{i,\sigma}^n)^2+\sum_{K\in\mathcal{T}}|K|E(\rho_K^n|r_K^n)\Big)+\mathcal{R}^{n,h}.
\end{aligned}
$$

References

1. Eymard, R., Gallouët, T., Herbin, R.: Finite volume methods. Handb. Numer. Anal **7**, 713–1018 (2000)
2. Feireisl, E., Jin, B.J., Novotny, A.: Relative entropies, suitable weak solutions, and weak strong uniqueness for the compressible Navier-Stokes system. http://arxiv.org/abs/1111.3082 (2011)
3. Feireisl, E., Novotny, A.: Weak-strong uniqueness property for the full Navier-Stokes-Fourier system. http://arxiv.org/abs/1111.4256 (2011)
4. Fettah, A., Gallouët, T.: Numerical approximation of the general compressible Stokes problem. IMA J. Numer. Anal. (2012)
5. Gallouët, T., Gastaldo, L., Herbin, R., Latché, J.C.: An unconditionally stable pressure correction scheme for the compressible barotropic Navier-Stokes equations. ESAIM. Math. Model. Numer. Anal. **42**(02), 303–331 (2008)
6. Gallouët, T., Herbin, R., Latché, J.C.: A convergent finite element-finite volume scheme for the compressible Stokes problem part I : the isothermal case. http://arxiv.org/abs/0712.2798 (2007)
7. Herbin, R., Latché, J.C., Nguyen, T.T., et al.: Consistent explicit staggered schemes for compressible flows part I: the barotropic Euler equations. In: ESAIM Proceedings, Juillet 2013, vol. 40, pp. 83–102 (2013)

A Mixed Explicit Implicit Time Stepping Scheme for Cartesian Embedded Boundary Meshes

Sandra May and Marsha Berger

Abstract We present a mixed explicit implicit time stepping scheme for solving the linear advection equationMay, Sandra on a Cartesian embedded boundary mesh. The scheme represents a new approach for overcoming the small cell problem—that standard finite volume schemes are not stable on the arbitrarily small *cut cells*. It uses implicit time stepping on cut cells for stability. On standard Cartesian cells, explicit time stepping is employed. This keeps the cost small and makes it possible to extend existing schemes from Cartesian meshes to Cartesian embedded boundary meshes. The coupling is done by *flux bounding*, for which we can prove a TVD result. We present numerical results in one and two dimensions showing second-order convergence in the L^1 norm and between first- and second-order convergence in the L^∞ norm.

1 Cut Cells and the Small Cell Problem

Cartesian embedded boundary methods, also referred to as *cut cell methods*, have been used increasingly in recent years to simulate flow around objects with complicated geometry. They are an alternative to unstructured or body-fitted grids. Cut cell methods cut the object out of a Cartesian background grid, resulting in irregular cells around the object, the so-called cut cells. Most cells are Cartesian cells for which standard methods can be used. Special methods must be developed for cut cells.

S. May (✉)
ETH Zurich, Rämistrasse 101, 8092 Zurich, Switzerland
e-mail: sandra.may@sam.math.ethz.ch

M. Berger
Courant Institute of Mathematical Sciences, 251 Mercer Street,
New York, NY 10012, USA
e-mail: berger@cims.nyu.edu

J. Fuhrmann et al. (eds.), *Finite Volumes for Complex Applications VII - Methods and Theoretical Aspects*, Springer Proceedings in Mathematics & Statistics 77,
DOI: 10.1007/978-3-319-05684-5_38,

Our goal is to extend a well-established, second-order projection method [2–4] for simulating the incompressible Euler equations from Cartesian meshes to embedded boundary meshes, to enable us to simulate flow around more complicated objects than currently possible. This will require extensions to both steps of the projection algorithm to account for the cut cells: *Step 1*, the update of the velocity field $\mathbf{u}^n$ to $\mathbf{u}^{n+1}$ without enforcing the incompressibility condition, and *Step 2*, the projection. In Step 1, a MUSCL scheme [3, 9] is employed. Due to the specifics of this predictor-corrector scheme, the equation that is solved is closer to the linear advection equation

$$s_t + \nabla \cdot (\mathbf{u}s) = 0, \quad \nabla \cdot \mathbf{u} = 0, \tag{1}$$

than to the nonlinear, incompressible Euler equations. Therefore, an important intermediate step consists in finding a solution to the *small cell problem* for the linear advection equation, which is the focus of this contribution.

The small cell problem refers to this: for standard finite volume schemes, the time step Δt depends on the size of the cell. One would like to choose the time step based on the size of the regular Cartesian cells; cut cells, however, can be arbitrarily small. Two methods for solving the small cell problem are the h-box method [6, 11] and the flux redistribution method [8, 10, 14]. The first one is second-order at the boundary, but has only been implemented in two dimensions due to its complexity. The latter is only first-order at the boundary, but has been used successfully in two and three dimensions.

Our main idea is to use an implicit time stepping scheme on cut cells for stability. On standard Cartesian cells, we use the MUSCL scheme. This approach avoids excessive cost and is compatible with our goal of extending an existing method to cut cells. We have developed a suitable method of switching between explicit and implicit time stepping, which we refer to as *flux bounding*. We will first focus on one dimension and discuss theoretical results for flux bounding as well as present numerical results. Then, we will briefly describe the extension of the scheme to two dimensions and present numerical results for advection along a ramp and in the interior of a circle.

Remark 1 The scheme presented in the following is *not* an IMEX-scheme. We use implicit time stepping on cut cells, and explicit time stepping on most other cells. This is different from the classical IMEX approach.

2 The Mixed Scheme in One Dimension

We consider the linear advection equation in one dimension given by

$$s_t + us_x = 0, \quad u > 0 \quad \text{constant}, \tag{2}$$

and we define the CFL number $\lambda = \frac{u\Delta t}{\Delta x}$. We will first describe how we switch between the schemes and then which implicit scheme we use.

As a model of the behavior of a cut cell grid in one dimension, we consider a grid that contains equidistant cells of length Δx and one small cell of length $\alpha\Delta x$, $\alpha \in (0, 1]$, in the middle. We use the index "0" for the small cell (see Fig. 1). For the explicit scheme on the regular cells of the grid we use the MUSCL scheme

$$s_i^{n+1} = s_i^n - \lambda\left(s_{i+1/2}^{n+1/2,L} - s_{i-1/2}^{n+1/2,L}\right), \quad \text{with} \quad s_{i+1/2}^{n+1/2,L} = s_i^n + (1-\lambda)s_{x,i}^n \frac{\Delta x}{2} \tag{3}$$

and $s_{x,i}^n \approx \partial_x s(x_i, t^n)$.

On the small cell, we want to use a fully implicit scheme for stability. Also, we want to keep the region with implicit fluxes as small as possible. We have investigated various ways of switching between the explicit and implicit scheme [13] that achieve these goals and have found *flux bounding* to be the most suitable approach. The idea is illustrated in Fig. 1: First, all cells with index less or equal to –2 or greater or equal to 2 are updated using fluxes computed by the explicit MUSCL scheme (those marked with E). This automatically prescribes the fluxes between cells –2 and –1 and between cells 1 and 2, $F_{-3/2}$ and $F_{3/2}$. In a second step, the fluxes between cells –1 and 0 and cells 0 and 1 are determined. To guarantee stability on the small cell, we use implicit fluxes (I) for $F_{-1/2}$ and $F_{1/2}$. As a consequence, cells –1 and 1 are transition cells, updated using both explicit and implicit fluxes. Note that this approach is automatically conservative.

Example 1 Consider explicit Euler time stepping with piecewise constant data for the explicit scheme and implicit Euler time stepping with piecewise constant data as the implicit scheme. Using flux bounding, the update for cells –1, 0, and 1 is

$$s_{-1}^{n+1} = s_{-1}^n - \lambda\left(s_{-1}^{n+1} - s_{-2}^n\right), \tag{4a}$$

$$s_0^{n+1} = s_0^n - \frac{\lambda}{\alpha}\left(s_0^{n+1} - s_{-1}^{n+1}\right), \tag{4b}$$

$$s_1^{n+1} = s_1^n - \lambda\left(s_1^n - s_0^{n+1}\right). \tag{4c}$$

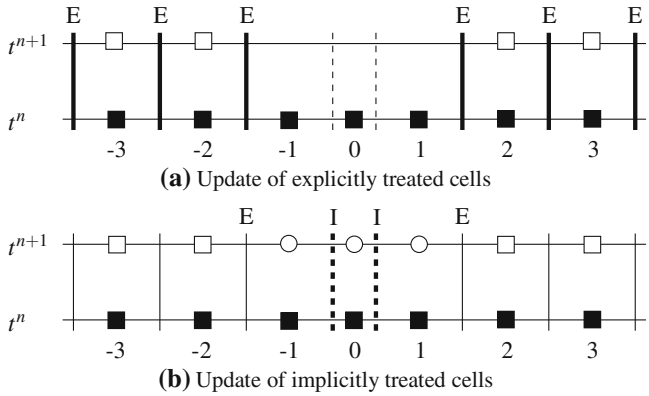
Then we have the following monotonicity/stability results for flux bounding:

Lemma 1 *The scheme described in Example 1 is monotone for* $0 \le \lambda \le 1$.

Proof 1 We focus on cells –1, 0, and 1. We rewrite Eqs. (4a)–(4c) as

$$\begin{aligned}
s_{-1}^{n+1} &= \frac{1}{1+\lambda} s_{-1}^n + \frac{\lambda}{1+\lambda} s_{-2}^n, \\
s_0^{n+1} &= \frac{1}{1+\frac{\lambda}{\alpha}} s_0^n + \frac{\frac{\lambda}{\alpha}}{1+\frac{\lambda}{\alpha}} s_{-1}^{n+1} = \frac{1}{1+\frac{\lambda}{\alpha}} s_0^n + \frac{\frac{\lambda}{\alpha}}{1+\frac{\lambda}{\alpha}} \left[\frac{1}{1+\lambda} s_{-1}^n + \frac{\lambda}{1+\lambda} s_{-2}^n\right], \\
s_1^{n+1} &= (1-\lambda)\, s_1^n + \lambda s_0^{n+1},
\end{aligned}$$

i.e., as convex combination of values at t^n. This implies the claim.



(a) Update of explicitly treated cells

(b) Update of implicitly treated cells

Fig. 1 With flux bounding, first update all explicit cells (*empty squares*), then use the explicitly computed fluxes $F_{-3/2}$ and $F_{3/2}$ and the implicitly computed fluxes $F_{-1/2}$ and $F_{1/2}$ to update the values in cells –1, 0, and 1 (*empty circles*)

If instead we use the explicit MUSCL scheme, with implicit Euler time stepping and piecewise constant data for the small cell, we have the following result [13]:

Lemma 2 *Consider the MUSCL scheme with minmod slope limiter as explicit scheme and implicit Euler time stepping with piecewise constant data as implicit scheme. Also use piecewise constant data in the cells two away from the small cell. If the mixed scheme uses flux bounding as described above, then the resulting scheme is TVD for* $0 \leq \lambda \leq 1$.

Based on these results, we use flux bounding to switch between the explicit and implicit scheme. We note that flux bounding generalizes to two and three dimensions in a straightforward way.

Besides the question of how to switch between the schemes, we need to find a suitable implicit scheme to be used with the MUSCL scheme. Implicit Euler time stepping is unconditionally SSP (strong stability preserving), but only first-order accurate. Our goal is to find a second-order implicit time stepping scheme that possesses very good stability properties, and that is compatible with the MUSCL scheme. Unfortunately, there is no second-order method that is unconditionally SSP [12]. Most second-order implicit schemes only allow for a time step that corresponds to a CFL of 2 or 4. Cut cells, however, can be arbitrarily small. We currently use the trapezoidal rule with slope reconstruction for the implicit method, since this is linearity preserving in combination with MUSCL. However the trapezoidal rule is not zero stable, and it is possible to construct examples for which the mixed method using the trapezoidal rule produces overshoots. It is not clear yet whether this will be a problem in our intended applications. We are working on a FCT approach [7, 15] for combining the trapezoidal rule with implicit Euler to preserve monotonicity; for compressible flow applications this will definitely be necessary.

Table 1 Numerical results in one dimension for the mixed scheme using implicit Euler and trapezoidal rule with unlimited slope reconstruction

Implicit method	N	L^1 error	L^1 order	L^∞ error	L^∞ order
Implicit Euler	80	1.22e–03	–	1.38e–02	–
	160	3.16e–04	1.95	6.66e–03	1.05
	320	8.21e–05	1.95	3.30e–03	1.01
Trapezoidal	80	2.10e–04	–	1.58e–03	–
	160	5.68e–05	1.89	3.51e–04	2.17
	320	1.48e–05	1.94	7.41e–05	2.24

Numerical Results: We consider the mixed scheme consisting of

- MUSCL with unlimited slope reconstruction,
- either implicit Euler or trapezoidal rule, both with unlimited slope reconstruction,
- flux bounding to connect the explicit and implicit method.

The switch in schemes in the transition cells leads to a loss of cancellation of the leading terms in the truncation error. Using implicit Euler in the small cell the one step error is then first order, and using the trapezoidal rule the one step error is second order, instead of the respective second and third order results when using the same scheme everywhere. We want to test numerically whether this error accumulates. We consider again the cut cell mesh shown in Fig. 1 on $[0, 1+\alpha\Delta x]$ with $\alpha = 10^{-4}$. We use initial data $\sin\left(\frac{2\pi x}{1+\alpha\Delta x}\right)$, periodic boundary conditions, and solve $s_t + s_x = 0$ with CFL number $\lambda = 0.8$. The final time is chosen such that the sinusoidal curve has returned to its original position. The results are shown in Table 1. We observe second-order convergence for both the L^1 and the L^∞ error for combining the MUSCL scheme with the trapezoidal rule. For using implicit Euler time stepping, the method converges with second order in L^1 and with first order in L^∞. Therefore, we do not observe accumulation of the transition error for this test. The observation that errors do not accumulate in the standard way on cut cell meshes has been noted before.

3 The Mixed Scheme in Two Dimensions

We again use flux bounding to connect the implicit and explicit scheme. Cartesian cells next to cut cells act as transition cells: for these cells, both explicit and implicit fluxes are used to update the cell values. This is sketched in Fig. 2a.

Since we are using an implicit scheme, we need to solve an implicit system in each iteration. The number of unknowns corresponds to the number of cut cells plus the number of transition cells. For using piecewise constant data or unlimited slope reconstruction on the cut cells and transition cells, the system is a linear (sparse) system. Inflow boundary conditions are prescribed at the appropriate locations in space and time. Outflow boundary conditions are computed using the respective (explicit

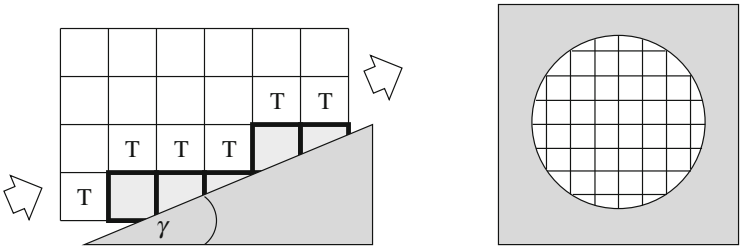


Fig. 2 **a** Cut cell grid for ramp test: Cut cells are shaded in *grey*, transition cells are marked by a "T". Thick *black* lines indicate implicit flux. **b** Sketch of cut cell grid for advection in interior of *circle*

or implicit) fluxes. For now we use pseudo-timestepping to solve the equations to make it easier to test various alternatives.

Our software in two (and three dimensions) is based on the open-access library *BoxLib* [5] for massively parallel, block-structured AMR algorithms. We combine this with *patchCubes*, part of the software package *Cart3D* [1], which generates the necessary cut cell information (cut cell and cut face centroids and volumes).

As in one dimension, we test the accuracy of the method and examine whether the transition error accumulates. For the following tests we focus on the mixed scheme using trapezoidal rule time stepping and unlimited slope reconstruction. We first consider the test sketched in Fig. 2a: Linear advection along a ramp with angle γ to the Cartesian background grid. This results in minimum volume fractions (area of a cut cell divided by area of a Cartesian cell) between 10^{-3} and 10^{-5}. We advect a smooth test function parallel to the ramp. The test function is a one-dimensional quadratic with respect to the line perpendicular to the ramp.

Table 2 shows the error measured in the L^1 and L^∞ norm for four ramp angles. We observe second-order convergence in the L^1 norm for all angles. This is expected. Since MUSCL alone is exact for quadratics, the maximum error occurs along the boundary, and the L^∞ norm is a suitable measure for the accuracy of the mixed scheme. The convergence rates lie between 1.3 and 1.8. It is common for cut cell methods to have *zig-zag* convergence since the error is not smooth at the cut cells. This indicates that in higher dimensions the error can accumulate, since the boundary is characteristic. We are currently examining approaches to improve this.

Next we consider advection in the interior of a circle with radius 1 (see Fig. 2b) using the spatially varying velocity field $(-y, x)^T$, with initial data $s(x, y, 0) = 1 + \exp(-60 * ((x - 0.85)^2 + y^2))$, so the peak is located close to the boundary, but not exactly at it. We compute for a time corresponding to one full rotation, and run 16 simulations with $\Delta x = \Delta y$ varying between 0.003 and 0.010. The L^1 error converges with second order. Figure 3 shows the L^∞ errors, measured over all cells (L^∞_{all}), transition cells only (L^∞_{T}), and cut cells only (L^∞_{C}). The L^∞_{all} error converges with second order, and is larger than the L^∞_{T} and L^∞_{C} error, since for these cell sizes the maximum error is at the peak. The L^∞_{T} and L^∞_{C} errors show the typical lack of

Table 2 *Advection along a non-coordinate-aligned ramp:* L^1 and L^∞ error for four ramp angles

γ	N^2	L^1 error	L^1 order	L^∞ error	L^∞ order
5^o	32^2	1.24e–06	–	2.42e–04	–
	64^2	2.60e–07	2.26	7.87e–05	1.62
	128^2	6.03e–08	2.11	2.73e–05	1.53
	256^2	1.62e–08	1.89	7.81e–06	1.80
20^o	32^2	3.32e–06	–	3.96e–04	–
	64^2	9.10e–07	1.87	1.39e–04	1.51
	128^2	2.37e–07	1.94	4.70e–05	1.57
	256^2	6.30e–08	1.91	1.88e–05	1.32
30^o	32^2	7.18e–06	–	4.67e–04	–
	64^2	1.83e–06	1.97	1.69e–04	1.47
	128^2	4.75e–07	1.95	6.19e–05	1.45
	256^2	1.25e–07	1.93	2.39e–05	1.37
40^o	32^2	1.08e–05	–	4.24e–04	–
	64^2	2.56e–06	2.07	1.43e–04	1.56
	128^2	6.24e–07	2.04	5.05e–05	1.51
	256^2	1.55e–07	2.01	1.85e–05	1.44

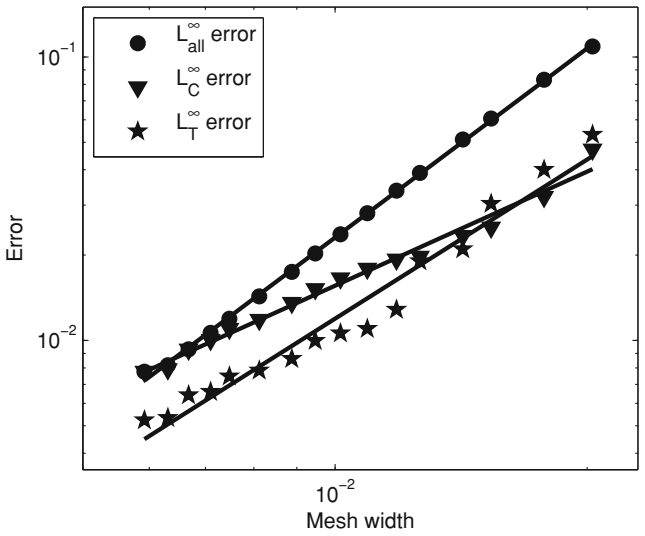


Fig. 3 *Advection in the interior of a circle:* Convergence plots for the L^∞ error measured over all cells, transition cells only, and cut cells only (L^∞_{all}, L^∞_{T}, and L^∞_{C}). The straight lines correspond to the least-squares fits with slopes 2.2, 1.9, and 1.3, respectively

smoothness in the convergence and some decay in accuracy. A least-squares fit shows convergence rates of 1.9 and 1.3, respectively. We are investigating some ideas on how to increase this to full second-order accuracy.

4 Summary

We have presented a new approach for solving the linear advection equation on a Cartesian embedded boundary grid based on using implicit time stepping on cut cells. The scheme is automatically conservative, and stable for arbitrarily small cells, though not yet fully second-order. Extensions to three dimensions and to include a projection algorithm on cut cells are under way.

Acknowledgments The authors would like to thank Ann Almgren, John Bell, and Andy Nonaka from Lawrence Berkeley National Laboratory for providing and helping the authors with the software packages BoxLib and VarDen. This work was supported in part by the DOE office of Advanced Scientific Computing under grant DE-FG02-88ER25053 and by AFOSR grant FA9550-13-1-0052. S. M. was also supported by ERC STG. N 306279, SPARCLE.

References

1. Aftosmis, M.J., Berger, M.J., Melton, J.E.: Robust and efficient Cartesian mesh generation for component-based geometry. AIAA J. **36**(6), 952–960 (1998)
2. Almgren, A.S., Bell, J.B., Colella, P., Howell, L.H., Welcome, M.L.: A conservative adaptive projection method for the variable density incompressible Navier-Stokes equations. J. Comput. Phys. **142**, 1–46 (1998)
3. Almgren, A.S., Bell, J.B., Szymczak, W.G.: A numerical method for the incompressible Navier-Stokes equations based on an approximate projection. SIAM J. Sci. Comput. **17**(2), 358–369 (1996)
4. Bell, J.B., Colella, P., Glaz, H.M.: A second-order projection method for the incompressible Navier-Stokes equations. J. Comput. Phys. **85**(2), 257–283 (1989)
5. Bell, J.B., et al.: BoxLib user's guide. Technical Report, CCSE, Lawrence Berkeley National Laboratory. https://ccse.lbl.gov/BoxLib/BoxLibUsersGuide.pdf (2012)
6. Berger, M.J., Helzel, C., LeVeque, R.: H-box methods for the approximation of hyperbolic conservation laws on irregular grids. SIAM J. Numer. Anal. **41**(3), 893–918 (2003)
7. Boris, J.P., Book, D.L.: Flux corrected transport. I. SHASTA, a fluid transport algorithm that works. J. Comput. Phys. **11**, 38–69 (1973)
8. Chern, I.L., Colella, P.: A conservative front tracking method for hyperbolic conservation laws. Technical Report, Lawrence Livermore National Laboratory, Livermore. Preprint UCRL-97200. (1987)
9. Colella, P.: Multidimensional upwind methods for hyperbolic conservation laws. J. Comput. Phys. **87**, 171–200 (1990)
10. Colella, P., Graves, D.T., Keen, B.J., Modiano, D.: A Cartesian grid embedded boundary method for hyperbolic conservation laws. J. Comput. Phys. **211**, 347–366 (2006)
11. Helzel, C., Berger, M.J., LeVeque, R.: A high-resolution rotated grid method for conservation laws with embedded geometries. SIAM J. Sci. Comput. **26**, 785–809 (2005)
12. Ketcheson, D.I., Macdonald, C.B., Gottlieb, S.: Optimal implicit strong stability preserving Runge-Kutta methods. Appl. Numer. Math. **59**, 373–392 (2009)
13. May, S.: Embedded boundary methods for flow in complex geometries. Ph.D. Thesis, Courant Institute of Mathematical Sciences, New York University (2013)
14. Pember, R., Bell, J.B., Colella, P., Crutchfield, W., Welcome, M.L.: An adaptive Cartesian grid method for unsteady compressible flow in irregular regions. J. Comput. Phys. **120**, 278–304 (1995)
15. Zalesak, S.T.: Fully multidimsional flux-corrected transport algorithms for fluids. J. Comput. Phys. **31**, 335–362 (1979)

Finite-Volume Analysis for the Cahn-Hilliard Equation with Dynamic Boundary Conditions

Flore Nabet

Abstract This work is devoted to the convergence analysis of a finite-volume approximation of the 2D Cahn-Hilliard equation with dynamic boundary conditions. The method that we propose couples a 2d-finite-volume method in a bounded, smooth domain $\Omega \subset \mathbb{R}^2$ and a 1d-finite-volume method on $\partial\Omega$. We prove convergence of the sequence of approximate solutions. One of the main ingredient is a suitable space translation estimate that gives a limit in $L^\infty\left(0, T, H^1(\Omega)\right)$ whose trace is in $L^\infty\left(0, T, H^1(\partial\Omega)\right)$.

1 Introduction

We consider a smooth, connected and bounded domain $\Omega \subset \mathbb{R}^2$ and $\Gamma = \partial\Omega$ its boundary. Let $T > 0$ be given.

We are interested here in the following phase separation model in material science (refered to as the Cahn-Hilliard equation with dynamic boundary conditions):

Find the concentration of one of the two phases $c : (0, T) \times \Omega \to \mathbb{R}$ satisfying:

$$\begin{cases} \partial_t c &= \Delta\mu, & \text{in } (0, T) \times \Omega; \\ \mu &= -\Delta c + f_b'(c), & \text{in } (0, T) \times \Omega; \\ \partial_t c_{|\Gamma} &= \Delta_\parallel c_{|\Gamma} - f_s'(c_{|\Gamma}) - \partial_n c, & \text{on } (0, T) \times \Gamma; \\ \partial_n \mu &= 0, & \text{on } (0, T) \times \Gamma; \\ c(0, .) &= c_0, & \text{in } \Omega; \end{cases} \tag{1}$$

where we have introduced an intermediate unknown: the chemical potential μ.

F. Nabet (✉)
Aix Marseille Université, CNRS, Centrale Marseille, I2M,
UMR 7373,13453 Marseille, France
e-mail: flore.nabet@univ-amu.fr

J. Fuhrmann et al. (eds.), *Finite Volumes for Complex Applications VII - Methods and Theoretical Aspects*, Springer Proceedings in Mathematics & Statistics 77,
DOI: 10.1007/978-3-319-05684-5_39,

The trace of c on Γ is noted $c_{|\Gamma}$, $\Delta_{\|}$ is the Laplace-Beltrami operator on Γ and ∂_n is the normal derivative at the boundary. The Cahn-Hilliard potentials f_b and f_s are nonlinear and they correspond respectively to the bulk and the surface free energy densities. In fact, several physical parameters should appear in the Cahn-Hilliard equation to account for physical properties of the studied system. However, these constants affect the readability of the problem. Thus, we have chosen to write the Problem (1) without these parameters.

We impose the homogeneous Neumann boundary condition for the chemical potential since no mass exchange can occur through the boundary. For many years, different authors studied the Cahn-Hilliard equation associated with the Neumann boundary condition for the order parameter c. In some cases, however, this condition is too restrictive to account for the interaction of the mixture with the walls. For this reason, physicists [4, 7] have recently introduced the Cahn-Hilliard system with dynamic boundary conditions (1). The associated free energy is the sum of a bulk free energy $\mathscr{F}_b$ and a surface free energy $\mathscr{F}_s$:

$$\mathscr{F}(c) = \underbrace{\int_\Omega \left(\frac{1}{2}|\nabla c|^2 + f_b(c)\right)}_{:=\mathscr{F}_b(c)} + \underbrace{\int_\Gamma \left(\frac{1}{2}\left|\nabla_{\|} c_{|\Gamma}\right|^2 + f_s(c_{|\Gamma})\right)}_{:=\mathscr{F}_s(c)}. \tag{2}$$

The dynamic boundary condition on c is chosen in such a way that the total free energy decreases with respect to time:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathscr{F}(c(t,.)) = -\int_\Omega |\nabla \mu(t,.)|^2 - \int_\Gamma \left|\partial_t c_{|\Gamma}(t,.)\right|^2, \quad t \in [0, T[.$$

The potentials are supposed to satisfy standard assumptions:
Assumptions 1. :

- *Dissipativity:* $\liminf_{|c|\to\infty} f_b''(c) > 0$ and $\liminf_{|c|\to\infty} f_s''(c) > 0$.
- *Polynomial growth for* f_b: there exist $C_b > 0$ and a real $p \geq 2$ such that:

$$\left|f_b^{(m)}(c)\right| \leq C_b\left(1 + |c|^{p-m}\right), \quad m \in \{0, 1, 2\}.$$

A typical choice for f_b is the double-well function $f_b(c) = c^2(1-c)^2$.

From a theoretical point of view, this system has already been studied (see for example [6] and the references therein). From a numerical point of view, we have several results. In [4, 7], authors propose a finite-difference framework but without proof of convergence. A convergence result is proved in [2] with a finite element space semi-discretization, but in a slab with periodic boundary conditions in lateral directions. In this paper, we propose a convergence analysis of a finite-volume scheme for the space discretization. This method is well adapted to the coupling between the dynamics in the domain and those on the boundary by the flux term $\partial_n c$. Moreover, this kind of scheme preserves the mass and accounts naturally for the non-flat geometry of the boundary and the associated Laplace-Beltrami operator.

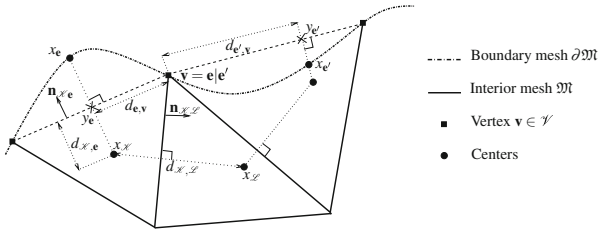


Fig. 1 Finite-volume meshes

2 The Discrete Setting

2.1 The Finite-Volume Meshes and Notation

We recall that the domain Ω is not polygonal and that we have to solve an equation on Γ. Thus, our notation will be slightly different than the usual finite-volume notation (see for example [3]). Let $\mathfrak{M}$ be a decomposition of Ω into polygonal subsets (called control volumes and noted $\mathscr{K} \in \mathfrak{M}$) except perhaps for those on the boundary which can have a curved edge. For each control volume $\mathscr{K} \in \mathfrak{M}$, we associate a point $x_{\mathscr{K}}$ which satisfies the orthogonality condition (see [3]). The main differences with the usual finite-volume notation are those on the boundary mesh $\partial\mathfrak{M}$. This mesh is constituted of the set of curved edges σ on the boundary Γ. With respect to the interior mesh, we keep the usual notation (Fig. 1) except for control volumes $\mathscr{K} \in \mathfrak{M}$ with one edge σ, at least, belonging to the boundary. In this case, $\mathscr{K}$ is not polygonal (σ is curved), we note $\widetilde{\mathscr{K}}$ the polygon formed by the vertices of $\mathscr{K}$ and by $m_{\widetilde{\mathscr{K}}}$ its Lebesgue measure. Note that $\widetilde{\mathscr{K}}$ may not be included in Ω. We will use two different notations for an element of $\partial\mathfrak{M}$: we note $\mathbf{e}$ when we consider it as a control volume belonging to $\partial\mathfrak{M}$ and we note σ when we consider it as the edge of an interior control volume $\mathscr{K} \in \mathfrak{M}$.

Let $\mathbf{e} \in \partial\mathfrak{M}$ be a boundary control volume and $\widetilde{\mathbf{e}}$ the corresponding chord. Their length are respectively noted $m_{\mathbf{e}}$ and $m_{\widetilde{\mathbf{e}}}$. If $\mathscr{K} \in \mathfrak{M}$ is the control volume such that $\mathbf{e} \subset \partial\mathscr{K}$, we set $x_{\mathbf{e}}$ as the intersection between Γ and the straight line passing through $x_{\mathscr{K}}$ and orthogonal to $\widetilde{\mathbf{e}}$. Let $y_{\mathbf{e}}$ be the intersection between the line $(x_{\mathscr{K}} x_{\mathbf{e}})$ and the chord $\widetilde{\mathbf{e}}$. We define $d_{\mathscr{K},\mathbf{e}}$ as the distance between the centers $x_{\mathscr{K}}$ and $y_{\mathbf{e}}$. Let $\mathscr{V}$ be the set of the vertices included in Γ and $d_{\mathbf{e},\mathbf{v}}$ be the distance between the center $y_{\mathbf{e}}$ and the vertex $\mathbf{v} \in \mathscr{V}$. For a vertex $\mathbf{v} = \mathbf{e}|\mathbf{e}' \in \Gamma$ which separates the control volumes $\mathbf{e}$ and $\mathbf{e}'$, we note $d_{\mathbf{e},\mathbf{e}'}$ the sum of $d_{\mathbf{e},\mathbf{v}}$ and $d_{\mathbf{e}',\mathbf{v}}$.

We can notice that all these quantities are computed by just knowing the coordinates of the vertices of the mesh in Γ. Thus, we do not need to know the equation of the boundary Γ.

We define the mesh size by: $h_{\mathscr{T}} = \sup\{\mathrm{diam}(\mathscr{K}), \mathscr{K} \in \mathfrak{M}\}$. In the results below, all the constants depend on a certain measure of regularity of the mesh. This is

classical and for the sake of simplicity, we do not give here its explicit value. In short, if this quantity is bounded when the mesh size tends to 0, this amounts to assume that the control volumes do not become flat when the mesh is refined.

2.2 Discrete Unknowns

With respect to the time discretization, we introduce a positive integer N. Then, we uniformly partition the temporal interval $[0, T]$ with the time step: $\Delta t = T/N$. Thus, for $n \in \{0, \cdots, N\}$, we define $t^n = n\Delta t$.

For each time step t^n, we denote the concentration unknowns by $c^n_{\mathscr{T}} = (c^n_{\mathfrak{M}}, c^n_{\partial\mathfrak{M}}) \in \mathbb{R}^{\mathscr{T}}$ and the chemical potential unknowns by $\mu^n_{\mathscr{T}} = (\mu^n_{\mathfrak{M}}, \mu^n_{\partial\mathfrak{M}}) \in \mathbb{R}^{\mathscr{T}}$. Regarding the chemical potential, we have the homogeneous Neumann boundary condition; thus we can define the boundary unknown $\mu^n_{\partial\mathfrak{M}} \in \mathbb{R}^{\partial\mathfrak{M}}$ as follows:

$$\mu^n_{\mathbf{e}} = \mu^n_{\mathscr{K}}, \quad \forall \mathbf{e} \in \partial\mathfrak{M} \text{ such that } \mathbf{e} = \sigma \in \mathscr{E}_{\mathscr{K}}.$$

Finally, let $u^{\Delta t}_{\mathfrak{M}}$ (respectively $u^{\Delta t}_{\partial\mathfrak{M}}$) be the piecewise constant function in $]0, T[\times\Omega$ (respectively $]0, T[\times\Gamma$) such that for all $t \in [t^n, t^{n+1}[$:

$$u^{\Delta t}_{\mathfrak{M}}(t, x) = u^{n+1}_{\mathscr{K}} \text{ if } x \in \mathscr{K} \quad \text{and} \quad u^{\Delta t}_{\partial\mathfrak{M}}(t, x) = u^{n+1}_{\mathbf{e}} \text{ if } x \in \mathbf{e}.$$

2.3 Inner Products and Norms

- *Discrete L^2 inner products*: For all $u_{\mathfrak{M}}, v_{\mathfrak{M}} \in \mathbb{R}^{\mathfrak{M}}$ and $u_{\partial\mathfrak{M}}, v_{\partial\mathfrak{M}} \in \mathbb{R}^{\partial\mathfrak{M}}$, we define:

$$(u_{\mathfrak{M}}, v_{\mathfrak{M}})_{\mathfrak{M}} = \sum_{\mathscr{K}\in\mathfrak{M}} m_{\mathscr{K}} u_{\mathscr{K}} v_{\mathscr{K}} \quad \text{and} \quad (u_{\partial\mathfrak{M}}, v_{\partial\mathfrak{M}})_{\partial\mathfrak{M}} = \sum_{\mathbf{e}\in\partial\mathfrak{M}} m_{\widetilde{\mathbf{e}}} u_{\mathbf{e}} v_{\mathbf{e}}.$$

  The associated discrete L^2 norms are noted $\|u_{\mathfrak{M}}\|_{0,\mathfrak{M}}$ and $\|u_{\partial\mathfrak{M}}\|_{0,\partial\mathfrak{M}}$.
- *Discrete H^1 semi-definite inner products*: For all $u_{\mathscr{T}}, v_{\mathscr{T}} \in \mathbb{R}^{\mathscr{T}}$ and $u_{\partial\mathfrak{M}}$, $v_{\partial\mathfrak{M}} \in \mathbb{R}^{\partial\mathfrak{M}}$:

$$[\![u_{\mathscr{T}}, v_{\mathscr{T}}]\!]_{1,\mathscr{T}} = \sum_{\sigma\in\mathscr{E}_{int}} \frac{m_\sigma}{d_{\mathscr{K},\mathscr{L}}}(u_{\mathscr{K}} - u_{\mathscr{L}})(v_{\mathscr{K}} - v_{\mathscr{L}}) + \sum_{\sigma\in\mathscr{E}_{ext}} \frac{m_{\widetilde{\mathbf{e}}}}{d_{\mathscr{K},\mathbf{e}}}(u_{\mathscr{K}} - u_{\mathbf{e}})(v_{\mathscr{K}} - v_{\mathbf{e}})$$

$$\text{and} \qquad [\![u_{\partial\mathfrak{M}}, v_{\partial\mathfrak{M}}]\!]_{1,\partial\mathfrak{M}} = \sum_{\mathbf{v}=\mathbf{e}|\mathbf{e}'\in\mathscr{V}} \frac{1}{d_{\mathbf{e},\mathbf{e}'}}(u_{\mathbf{e}} - u_{\mathbf{e}'})(v_{\mathbf{e}} - v_{\mathbf{e}'}).$$

  The associated seminorms are noted $|u_{\mathscr{T}}|_{1,\mathscr{T}}$ and $|u_{\partial\mathfrak{M}}|_{1,\partial\mathfrak{M}}$.

3 Numerical Scheme and Discrete Energy

3.1 Finite-Volume Scheme

In this section, we give the finite-volume scheme used to solve the Cahn-Hilliard equation (1). In the interior mesh $\mathfrak{M}$, we use the usual finite-volume approximation based on a consistent two-point flux approximation for Laplace operators. As regards the equation on the boundary mesh $\partial\mathfrak{M}$, we use a 1d-finite-volume scheme on a curved domain and a consistent two-point flux approximation for the Laplace-Beltrami operator.

We assume that $c^n_{\mathscr{T}} \in \mathbb{R}^{\mathscr{T}}$ is given, the scheme is then written as follows: Find $(c^{n+1}_{\mathscr{T}}, \mu^{n+1}_{\mathscr{T}}) \in \mathbb{R}^{\mathscr{T}} \times \mathbb{R}^{\mathscr{T}}$ such that $\forall u_{\mathscr{T}}, v_{\mathscr{T}} \in \mathbb{R}^{\mathscr{T}}$:

$$\begin{cases} \left(\frac{c^{n+1}_{\mathfrak{M}} - c^n_{\mathfrak{M}}}{\Delta t}, v_{\mathfrak{M}}\right)_{\mathfrak{M}} = -[\![\mu^{n+1}_{\mathscr{T}}, v_{\mathscr{T}}]\!]_{1,\mathscr{T}} \\ \left(\mu^{n+1}_{\mathfrak{M}}, u_{\mathfrak{M}}\right)_{\mathfrak{M}} = \sum_{\sigma\in\mathscr{E}_{int}} \frac{m_\sigma}{d_{\mathscr{K},\mathscr{L}}}(c^{n+1}_{\mathscr{K}} - c^{n+1}_{\mathscr{L}})(u_{\mathscr{K}} - u_{\mathscr{L}}) \\ \qquad\qquad + \boxed{\sum_{\sigma\in\mathscr{E}_{ext}} \frac{m_{\tilde{\mathbf{e}}}}{d_{\mathscr{K},\mathbf{e}}}(c^{n+1}_{\mathscr{K}} - c^{n+1}_{\mathbf{e}})u_{\mathscr{K}}} \\ \qquad\qquad + \sum_{\mathscr{K}\in\mathfrak{M}} m_{\widetilde{\mathscr{K}}} d^{f_b}(c^n_{\mathscr{K}}, c^{n+1}_{\mathscr{K}})u_{\mathscr{K}} \\ \left(\frac{c^{n+1}_{\partial\mathfrak{M}} - c^n_{\partial\mathfrak{M}}}{\Delta t}, u_{\partial\mathfrak{M}}\right)_{\partial\mathfrak{M}} = -[\![c^{n+1}_{\partial\mathfrak{M}}, u_{\partial\mathfrak{M}}]\!]_{1,\partial\mathfrak{M}} - \sum_{\mathbf{e}\in\partial\mathfrak{M}} m_{\tilde{\mathbf{e}}} d^{f_s}(c^n_{\mathbf{e}}, c^{n+1}_{\mathbf{e}})u_{\mathbf{e}} \\ \qquad\qquad - \boxed{\sum_{\sigma\in\mathscr{E}_{ext}} \frac{m_{\tilde{\mathbf{e}}}}{d_{\mathscr{K},\mathbf{e}}}(c^{n+1}_{\mathbf{e}} - c^{n+1}_{\mathscr{K}})u_{\mathbf{e}}} \end{cases} \tag{3}$$

With the aim of obtaining convergence result without any condition on the step time Δt, we use a semi-implicit discretization for nonlinear terms:

$$d^{f_b}(x,y) = \frac{f_b(y) - f_b(x)}{y - x} \quad \text{and} \quad d^{f_s}(x,y) = \frac{f_s(y) - f_s(x)}{y - x}, \quad \forall x, y. \tag{4}$$

We can note that we mostly use in practice polynomial functions for f_b and f_s. Then, the term $d^f(x,y)$ can be written as a polynomial function in the variables x, y. Thus, we do not have numerical instability if x is too close to y. If we choose non polynomial functions for nonlinear terms, we have to adapt our discretization (see [1] for more details).

We remark that we can also choose an implicit discretization for nonlinear terms but in that case the same results hold only for $\Delta t \leq \Delta t_0$, with a small enough Δt_0 which only depends on the parameters on the equation.

In each case, we have to use a Newton method at each time step; its convergence is achieved in a few inner iterations.

We can notice that the finite-volume scheme is a low-order method. Thus, the approximation of the boundary does not influence the order of the method and it is not necessary to use curved element to improve the convergence of the scheme (3).

The boxed terms give the coupling between interior and boundary unknowns: the one in the second equation comes from the Laplacian of c in Ω and the one in the third equation stems from the normal derivative term in the dynamic boundary condition on Γ.

In order to improve the presentation and the analysis, we have written the scheme (3) in a way that looks like a variational formulation. We easily recover the usual finite-volume flux balance equations if, for each control volume, we choose the indicator function of this particular control volume as a test function in (3).

3.2 Discrete Energy Estimate

The discrete energy estimate is one of the key points for the proofs of existence and convergence results.

Definition 1 (*Discrete free energy*) The discrete free energy corresponding to the continuous definition (2) is defined by:

$$\mathscr{F}_{\mathscr{T}}(c_{\mathscr{T}}) = \underbrace{\frac{1}{2}|c_{\mathscr{T}}|^2_{1,\mathscr{T}} + \sum_{\mathscr{K}\in\mathfrak{M}} m_{\mathscr{K}} f_b(c_{\mathscr{K}})}_{:=\mathscr{F}_{b,\mathscr{T}}(c_{\mathscr{T}})} + \underbrace{\frac{1}{2}|c_{\partial\mathfrak{M}}|^2_{1,\partial\mathfrak{M}} + \sum_{\mathbf{e}\in\partial\mathfrak{M}} m_{\tilde{\mathbf{e}}} f_s(c_{\mathbf{e}})}_{:=\mathscr{F}_{s,\partial\mathfrak{M}}(c_{\partial\mathfrak{M}})}, \quad \forall c_{\mathscr{T}} \in \mathbb{R}^{\mathscr{T}}.$$

Using the scheme (3) with $u_{\mathscr{T}} = c^{n+1}_{\mathscr{T}} - c^n_{\mathscr{T}}$ and $v_{\mathscr{T}} = \mu^{n+1}_{\mathscr{T}}$ as test functions and the discretization (4) for nonlinear terms, we obtain the following energy equality:

Proposition 1 (Discrete energy estimate) *Let $c^n_{\mathscr{T}} \in \mathbb{R}^{\mathscr{T}}$. We assume that there exists a solution $(c^{n+1}_{\mathscr{T}}, \mu^{n+1}_{\mathscr{T}})$ to Problem (3). Then, the following equality holds:*

$$\begin{aligned}\mathscr{F}_{\mathscr{T}}(c^{n+1}_{\mathscr{T}}) - \mathscr{F}_{\mathscr{T}}(c^n_{\mathscr{T}}) + \Delta t \left|\mu^{n+1}_{\mathscr{T}}\right|^2_{1,\mathscr{T}} + \tfrac{1}{\Delta t}\left\|c^{n+1}_{\partial\mathfrak{M}} - c^n_{\partial\mathfrak{M}}\right\|^2_{0,\partial\mathfrak{M}} \\ + \tfrac{1}{2}\left|c^{n+1}_{\mathscr{T}} - c^n_{\mathscr{T}}\right|^2_{1,\mathscr{T}} + \tfrac{1}{2}\left|c^{n+1}_{\partial\mathfrak{M}} - c^n_{\partial\mathfrak{M}}\right|^2_{1,\partial\mathfrak{M}} = 0.\end{aligned} \tag{5}$$

This estimate gives a $L^\infty(0,T;H^1(\Omega))$ bound on the discrete solution $c^{\Delta t}_{\mathscr{T}}$ and a $L^\infty(0,T;H^1(\Gamma))$ bound on its trace $c^{\Delta t}_{\partial\mathfrak{M}}$.

4 Convergence Analysis

By using the topological degree theory, we can prove that if $c^n_{\mathscr{T}} \in \mathbb{R}^{\mathscr{T}}$ is given, there exists at least one solution $(c^{n+1}_{\mathscr{T}}, \mu^{n+1}_{\mathscr{T}}) \in \mathbb{R}^{\mathscr{T}} \times \mathbb{R}^{\mathscr{T}}$ to discrete Problem (3) (see [8] for more details).

We recall the definition of a solution to Problem (1) in a weak sense:

Definition 2 (*Weak formulation*) We say that a couple $(c, \mu) \in L^\infty(0, T; H^1(\Omega)) \times L^2(0, T; H^1(\Omega))$ such that $\mathrm{Tr}(c) \in L^\infty(0, T; H^1(\Gamma))$ is solution to continuous Problem (1) in the weak sense if for all $\psi \in \mathscr{C}_c^\infty\left([0, T[\times\overline{\Omega}\right)$, the following identities hold:

$$\int_0^T \int_\Omega (-\partial_t \psi c + \nabla\mu \cdot \nabla\psi) = \int_\Omega c^0 \psi(0, .), \tag{6}$$

$$\int_0^T \int_\Omega \left(-\mu\psi + \nabla c \cdot \nabla\psi + f_b'(c)\psi\right) + \int_0^T \int_\Gamma \left(-\partial_t \psi c + \nabla_\parallel c \cdot \nabla_\parallel \psi + f_s'(c)\psi\right)$$
$$= \int_\Gamma \mathrm{Tr}(c^0)\psi(0, .). \tag{7}$$

Then, we have the following convergence result.

Theorem 1 (Convergence theorem) *Assuming that Assumptions 1 hold, let us consider Problem* (1) *with an initial condition* $c_0 \in H^1(\Omega)$ *such that* $\mathrm{Tr}(c_0) \in H^1(\Gamma)$. *Then, there exists a weak solution* (c, μ) *on* $[0, T[$ *(in the sense of Definition 2). Furthermore, let* $(c^{(m)}, c^{(m)}_{|\Gamma})_{m\in\mathbb{N}}$ *and* $(\mu^{(m)})_{m\in\mathbb{N}}$ *be a sequence of solutions to Problem* (3) *associated with a sequence of discretizations such that the space and time steps,* $h_{\mathscr{T}}^{(m)}$ *and* $\Delta t^{(m)}$ *respectively, tend to* 0. *Then, up to a subsequence, the following convergence properties hold, for all* $q \geq 1$*:*

$$c^{(m)} \to c \text{ in } L^2(0, T; L^q(\Omega)) \text{strongly}, \quad c^{(m)}_{|\Gamma} \to \mathrm{Tr}(c) \text{ in } L^2(0, T; L^q(\Gamma)) \text{strongly},$$
$$\text{and} \quad \mu^{(m)} \rightharpoonup \mu \text{ in } L^2(0, T; L^q(\Omega)) \text{weakly}.$$

The discrete initial concentration used is the mean-value projection.

The main difficulty of this proof is the passage to the limit in nonlinear terms both in Ω and on Γ. Indeed, the usual $L^2((0, T)\times\Omega)$ compactness is not sufficient and we need to have an additional compactness property of the trace of c in $L^2(]0, T[\times\Gamma)$.

Theorem 2 (Estimation of space translates) *There exists an extension operator* $\phi : \mathbb{R}^{\mathscr{T}} \to L^2(\mathbb{R}^2)$ *satisfying* $\phi(u_{\mathscr{T}}) = u_{\mathscr{T}}$ *in* Ω *such that the following identity holds for all* $\eta \in \mathbb{R}^2$ *with* $C > 0$ *independent of* $h_{\mathscr{T}}$ *and* η*: For all* $u_{\mathscr{T}} \in \mathbb{R}^{\mathscr{T}}$,

$$\|\phi(u_{\mathscr{T}})(. + \eta) - \phi(u_{\mathscr{T}})\|^2_{L^2(\mathbb{R}^2)} \leq C|\eta|\,(|\eta| + h_{\mathscr{T}})\left(|u_{\mathscr{T}}|^2_{1,\mathscr{T}} + |u_{\partial\mathfrak{M}}|^2_{1,\partial\mathfrak{M}} + \|u_{\partial\mathfrak{M}}\|^2_{0,\partial\mathfrak{M}}\right).$$

Corollary 1 *Let* $(u_{\mathscr{T}_i})_i$ *be a sequence of functions with uniform bounds on discrete* H^1*-norms on* Ω *and* Γ. *We can extract a subsequence, still referred to as* $(u_{\mathscr{T}_i})_i$ *for simplicity, which is strongly converging in* $L^2(\Omega)$ *towards a certain function* u *of* $H^1(\Omega)$ *whose trace belongs to* $H^1(\Gamma)$ *and such that* $(u_{\partial\mathfrak{M}_i})_i$ *is strongly converging in* $L^2(\Gamma)$ *towards* $\mathrm{Tr}(u)$.

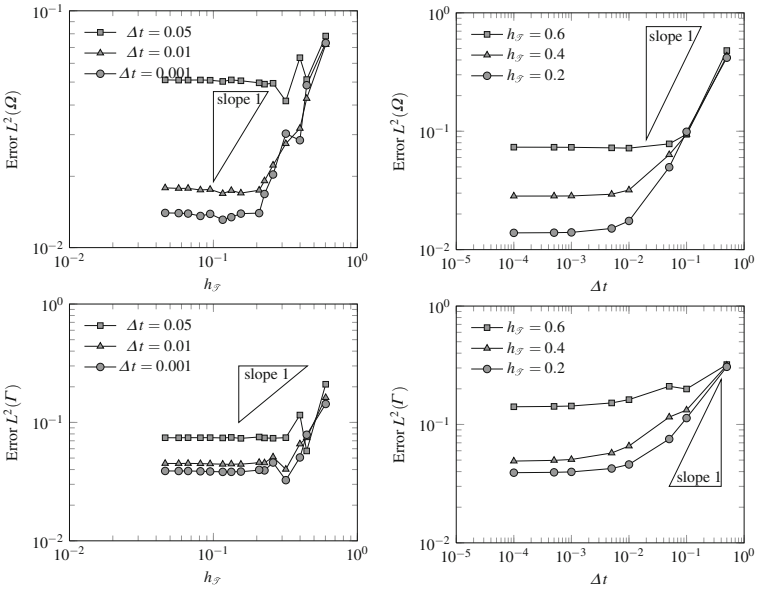
To obtain similar results with the sequence of functions which also depends on time, we have to consider the estimation of time translates. To this end, we adapt the

proof of Theorem A.2 in [5] and we use the particular form of the extension operator ϕ and the coupling between the domain Ω and its boundary Γ.

Then, thanks to the *a priori* estimates on the solutions (see [8]), there exists $c \in L^2(0, T, H^1(\Omega))$ with $\mathrm{Tr}(c) \in L^2(0, T, H^1(\Gamma))$ such that, up to a subsequence, $c_{\mathfrak{M}}^{\Delta t}$ strongly converges to c in $L^2(]0, T[\times\Omega)$ and moreover, also $c_{\partial\mathfrak{M}}^{\Delta t}$ strongly converges to $Tr(c)$ in $L^2(]0, T[\times\Gamma)$. It is now more or less standard to pass to the limit in the scheme and thus to prove the convergence result.

5 Numerical Tests

In [8], we give numerical experiments with different choices of parameters and surface potential f_s that show the different expected qualitative behavior of the solutions. In this paper, we focus on the numerical error estimates. Since no explicit non trivial solutions exist for our problem, we have to change the Problem (1). We add source term in the first equation of (1) and another one in the third equation of (1). We notice that μ then satisfies a non homogeneous Neumann boundary condition that can be easily handled in the FV setting. We consider the manufactured solution $c(t, (x, y)) = (1+\tanh(5*(x+t))$ with Ω the unit circle. We plot the error between the exact and approximate solutions at time $T = 0.5$ for the norm $L^2(\Omega)$ and $L^2(\Gamma)$.



As expected, we observe the first order convergence in the L^2 norm.

References

1. Boyer, F., Minjeaud, S.: Numerical schemes for a three component Cahn-Hilliard model. ESAIM Math. Model. Numer. Anal. **45**(4), 697–738 (2011)
2. Cherfils, L., Petcu, M., Pierre, M.: A numerical analysis of the Cahn-Hilliard equation with dynamic boundary conditions. Discrete Contin. Dyn. Syst. **27**(4), 1511–1533 (2010)
3. Eymard, R., Gallouët, T., Herbin, R.: Finite volume methods. In: Ciarlet, Ph, Lions, J.L. (eds.) Handbook of Numerical Analysis, pp. 713–1018. Marcel Dekker Inc., New York (2000)
4. Fischer, H.P., Maass, P., Dieterich, W.: Novel surface modes in spinodal decomposition. Phys. Rev. Lett. **79**, 893–896 (1997)
5. Gallouët, T., Herbin, R., Larcher, A., Latché, J.C.: Analysis of a fractional-step scheme for the p1 radiative diffusion model. http://hal.archives-ouvertes.fr/hal-00477086 (2009)
6. Goldstein, G., Miranville, A., Schimperna, G.: A Cahn-Hilliard model in a domain with non-permeable walls. Phys. D **240**(8), 754–766 (2011)
7. Kenzler, R., Eurich, F., Maass, P., Rinn, B., Schropp, J., Bohl, E., Dieterich, W.: Phase separation in confined geometries: solving the Cahn-Hilliard equation with generic boundary conditions. J. Comput. Phys. Commun. **133**, 139–157 (2001)
8. Nabet, F.: Finite volume method for the Cahn-Hilliard equation with dynamic boundary conditions. http://hal.archives-ouvertes.fr/hal-00872690 (2013)

Weak Convergence of Nonlinear Finite Volume Schemes for Linear Hyperbolic Systems

Michaël Ndjinga

Abstract We prove the weak convergence of nonlinear finite volume schemes applied to symmetric hyperbolic systems of linear partial differential equations on $\mathbb{R}^d$. The upwinding matrix can be any nonlinear non negative matrix valued function, the initial data any possibly discontinuous function in $L^2(\mathbb{R}^d)$ and the mesh may be smoothly unstructured.

1 Introduction

Using spectral techniques, [9] proved the L^2 stability of explicit and implicit finite volume schemes with a nonnegative nonlinear upwinding for linear hyperbolic systems on unstructured meshes. We cannot use Lax equivalence theorem to prove the convergence of these schemes because it requires a smooth solution, a linear scheme, as well as a rectangular grid, and all three assumptions are false in our case. More generally, proofs of convergence of finite volume methods for hyperbolic systems on unstructured meshes assume some smoothness of the solution (at least H^1 with compact support in [11], H^2 in [4], H^s with $s \in]0, 1]$ in [6]), and are restricted to the classical "upwind scheme" [4, 6, 11]. The mesh sequence is usually assumed to be quasiuniform: $\exists c_1, c_2 > 0$ such that $v_\alpha \geq c_1(\sup_\alpha d_\alpha)^d$ [4, 6, 11] and $s_{\alpha\beta} \leq c_2(\sup_\alpha d_\alpha)^{d-1}$ [4, 11], $s_{\alpha\beta} \geq c_2(\sup_\alpha d_\alpha)^{d-1}$ [6] for any cell α with diameter d_α, volume v_α and interface area $s_{\alpha\beta}$ for a neighbouring cell β. However the use of a different upwinding term can increase the precision of the numerical method by allowing less numerical diffusion. In the scalar case TVD methods often use a nonlinear upwinding term ([8] Chap. 16). To our knowledge there is no general result of convergence for nonlinear schemes applied to linear hyperbolic systems.
More precisely, we are interested in the use of centered type schemes for the

M. Ndjinga (✉)
CEA-Saclay, DEN,DM2S,STMF,LMEC, 91191 Gif-sur-Yvette, France
e-mail: michael.ndjinga@cea.fr

J. Fuhrmann et al. (eds.), *Finite Volumes for Complex Applications VII - Methods and Theoretical Aspects*, Springer Proceedings in Mathematics & Statistics 77,
DOI: 10.1007/978-3-319-05684-5_40,

accurate simulation of two phase flows in nuclear reactors thermalhydraulics (see [7]). Because these flows usually have low Mach numbers, we cannot use upwind type schemes (see [3]). Implicit centered schemes often give better results [1], provided they do not display checkerboard type oscillations (see [2]). Our first objective in this paper is to prove for linear problems using compactness methods that the centered scheme always converges weakly. Another issue is the capture of discontinuous solutions since vaporisation and condensation fronts are common features of our two phase systems. Purely centered schemes sharply violate the maximum principle around discontinuities so we need a scheme that would be centered in smooth parts of the flow and upwind around discontinuities. Our second objective is therefore to set some theoretical bases for such an adaptive approach.

We prove for any initial datum in $L^2(\mathbb{R}^d)$ the weak convergence of the familly of schemes (2–3) when the upwinding matrix valued functions $D_{\alpha\beta}(\mathscr{U})$ are non negative, uniformly bounded and Lipschitz continuous (theorem 2). The mesh cells are requested to have bounded aspect ratio : $\exists c \in \mathbb{R}$ so that for any cell α, $s_\alpha d_\alpha < c v_\alpha$, and be smoothly unstructured that is, neighbouring cells α, β should satisfy $\frac{d_\beta}{d_\alpha} \leq c$ and $||\frac{1}{2}(x_\alpha + x_\beta) - x_{\alpha\beta})|| \leq c d_\alpha^2$, where s_α denotes the area surrounding the cell, x_α the cell center of mass, and $x_{\alpha\beta}$ the interface center of mass. We however do not impose the quasiuniformity of the mesh sequence.

Since [9] proved the stability of time discrete explicit and implicit schemes, we use here the semi discrete setting of the schemes in order to obtain compact proofs. Unlike [5] who use compactness methods in L^∞ and L^1 for scalar conservation laws, we use L^2 compactness methods in the context of hyperbolic systems. First the Cauchy-Lipschitz theorem yields the existence of a discrete solution and the stability of the scheme (theorem 1). Then using the Banach-Alaoglu theorem we deduce the weak convergence of a subsequence, and then show that the weak limit is indeed a solution of the continuous problem (theorem 2). The uniqueness of the solution of the continuous problem yields the convergence of the entire discrete sequence.

2 Stability Theorem

We seek for a vector field $U(\mathbf{x}, t) \in \mathbb{R}^m$ with $\mathbf{x} \in \mathbb{R}^d$, $t \in \mathbb{R}_+$, satisfying the following linear system of conservation laws

$$\partial_t U(\mathbf{x}, t) + \nabla \cdot F(U)(\mathbf{x}, t) = 0, \tag{1}$$

where A_k are $m \times m$ real matrices and $F(U) = (A_1 U, \ldots, A_d U)$ is a linear function. If there exists a symmetric positive definite matrix E such that EA_k is symmetric for all k, then the Cauchy problem for system (1) is well posed in $L^2(\mathbb{R}^d)$ (see [10]). In the following we will make the simplifying assumption that the matrices A_k are symmetric and thus for any vector $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_d) \in \mathbb{R}^d$ the matrix $A(\omega) = \sum_{k=1}^d \omega_k A_k$ will be diagonalisable with real eigenvalues.

In order to approximate numerically the solutions of system (1), $\mathbb{R}^d$ is partitioned by a mesh $\mathscr{T}$ composed of a countable number of cells $\mathscr{T}_\alpha$ with indices $\alpha \in \mathbb{N}$, measure $0 < v_\alpha < \infty$, center of mass x_α and diameter d_α. Two neighbouring cells $\mathscr{T}_\alpha$ and $\mathscr{T}_\beta$ are separated by a smooth interface $f_{\alpha\beta}$ with an associated measure $0 < s_{\alpha\beta} < \infty$ ($s_{\alpha\beta} = s_{\beta\alpha}$), center of mass $x_{\alpha\beta}$ and average normal vector $\boldsymbol{\omega}_{\alpha\beta} = \frac{1}{s_{\alpha\beta}} \int_{f_{\alpha\beta}} \mathbf{ds} \in \mathbb{R}^d$ oriented from $\mathscr{T}_\alpha$ toward $\mathscr{T}_\beta$ ($\boldsymbol{\omega}_{\alpha\beta} = -\boldsymbol{\omega}_{\beta\alpha}$). The set of neighbours of a cell $\mathscr{T}_\alpha$ is denoted $v(\alpha)$ and the total area surrounding the cell α is $s_\alpha = \sum_{\beta\in v(\alpha)} s_{\alpha\beta}$. We recall that Stokes theorem yields $\forall\alpha$, $\sum_{\beta\in v(\alpha)} s_{\alpha\beta}\boldsymbol{\omega}_{\alpha\beta} = 0$. Using the finite volume framework (see [8]), the discrete unknown $U_\alpha(t)$ approximates the average value of the unknown U in the cell $\mathscr{T}_\alpha$ at time $t > 0$ and the global unknown vector is $\mathscr{U}(t) = {}^t(U_1(t), U_2(t), \dots)$. We introduce the set $L^2(\mathscr{T})^m$ of discrete functions $\mathscr{U} : \mathbb{R}^d \to \mathbb{R}^m$ that take constant values U_α on each cell α of $\mathscr{T}$, and are square integrable on $\mathbb{R}^d$: $||\mathscr{U}||_2^2 \equiv \int_{\mathbb{R}^d} \mathscr{U} \cdot \mathscr{U} \,\mathrm{d}v = \sum_\alpha v_\alpha {}^t U_\alpha U_\alpha < \infty$.

We consider flux schemes in the following semi-discrete form

$$U_\alpha(t)' + \frac{1}{v_\alpha} \sum_{\beta\in v(\alpha)} s_{\alpha\beta} F_{\alpha\beta}(\mathscr{U}(t)) = 0. \tag{2}$$

These schemes are designed to capture possibly discontinuous weak solutions of (1). They are said to be weakly consistent if $\forall\alpha, \beta \in \mathbb{N}$, $\forall U \in \mathbb{R}^m$, $F_{\alpha\beta}(U, U) = F(U)\omega_{\alpha\beta}$. We are interested in first order weakly consistent schemes where the interfacial flux takes the form

$$F_{\alpha\beta} = \frac{1}{2}(F(U_\alpha) + F(U_\beta))\boldsymbol{\omega}_{\alpha\beta} + D_{\alpha\beta}(\mathscr{U})\frac{U_\alpha - U_\beta}{2} \tag{3}$$

$$= F(U_\alpha)\boldsymbol{\omega}_{\alpha\beta} - \frac{1}{2}(A(\boldsymbol{\omega}_{\alpha\beta}) - D_{\alpha\beta}(\mathscr{U}))(U_\alpha - U_\beta) \tag{4}$$

where $D_{\alpha\beta} = D_{\beta\alpha}$ is the upwinding (or diffusion) matrix defined at each interface $f_{\alpha\beta}$ and assumed to depend on $\mathscr{U}$.

For a test function $\phi \in \mathscr{D}(\mathbb{R}^d \times \mathbb{R}_+)$ (the space of smooth functions with compact support in $\mathbb{R}^d \times \mathbb{R}_+$), and a mesh $\mathscr{T}$ we denote the mean value of ϕ on cell $\mathscr{T}_\alpha$ by $\phi_\alpha(t) = \frac{1}{v_\alpha} \int_{\mathscr{T}_\alpha} \phi(\mathbf{x}, t)\mathrm{d}v$ and on the face $f_{\alpha\beta}$ by $\phi_{\alpha\beta}(t) = \frac{1}{s_{\alpha\beta}} \int_{f_{\alpha\beta}} \phi(\mathbf{x}, t)\mathrm{d}s$.

Theorem 1 (Existence and Stability) *Assume that the mesh $\mathscr{T}$ satisfies $\sup_\alpha \frac{s_\alpha}{v_\alpha} < \infty$. Then for any initial datum $\mathscr{U}_0 \in L^2(\mathscr{T})^m$, the scheme (2–3) associated with symmetric non negative uniformly bounded and Lipschitz continuous upwinding matrices $D_{\alpha\beta}$ admits a solution $\mathscr{U}(t)$ defined for any $t \in \mathbb{R}_+$ and satisfying $\forall t \in \mathbb{R}_+$, $||\mathscr{U}(t)||_2 \le ||\mathscr{U}_0||_2$. Furthermore we have the following bound:*

$$\forall T \in \mathbb{R}_+, \quad \int_0^T \sum_{f_{\alpha\beta}} s_{\alpha\beta} {}^t(U_\alpha - U_\beta) D_{\alpha\beta}(U_\alpha - U_\beta) \,\mathrm{d}t = ||\mathscr{U}(0)||_2^2 - ||\mathscr{U}(T)||_2^2. \tag{5}$$

Proof Any discrete solution of the scheme (2) associated with the initial datum $\mathscr{U}(0) = \mathscr{U}_0$ is a solution of the Cauchy problem $\mathscr{U}' + \mathscr{F}(\mathscr{U}) = 0$, where $\mathscr{F}(\mathscr{U})_\alpha = -\frac{1}{v_\alpha}\sum_{\beta\in v(\alpha)} \frac{s_{\alpha\beta}}{2}(A(\boldsymbol{\omega}_{\alpha\beta}) - D_{\alpha\beta}(\mathscr{U}))(U_\alpha - U_\beta)$ from (4).
$\forall \mathscr{U}, \mathscr{V} \in L^2(\mathscr{T})^m$,

$$
\begin{aligned}
||\mathscr{F}(\mathscr{U}) - \mathscr{F}(\mathscr{V})||_2^2 &= \sum_\alpha v_\alpha || \frac{1}{v_\alpha} \sum_{\beta\in v(\alpha)} \frac{s_{\alpha\beta}}{2}(A(\boldsymbol{\omega}_{\alpha\beta})(U_\alpha - V_\alpha - U_\beta + V_\beta) \\
&\quad - D_{\alpha\beta}(\mathscr{U})(U_\alpha - U_\beta) + D_{\alpha\beta}(\mathscr{V})(V_\alpha - V_\beta))||_2^2 \\
&= \sum_\alpha v_\alpha || \frac{1}{v_\alpha} \sum_{\beta\in v(\alpha)} \frac{s_{\alpha\beta}}{2}((D_{\alpha\beta}(\mathscr{U}) - D_{\alpha\beta}(\mathscr{V}))(V_\alpha - V_\beta) \\
&\quad + (A(\boldsymbol{\omega}_{\alpha\beta}) - D_{\alpha\beta}(\mathscr{U}))(U_\alpha - V_\alpha - U_\beta + V_\beta))||_2^2.
\end{aligned}
$$

Using Minkowski inequality and the convexity of $x \to |x|^2$ yields

$$
\begin{aligned}
||\mathscr{F}(\mathscr{U}) - \mathscr{F}(\mathscr{V})||_2 &\le \sqrt{\sum_\alpha \frac{1}{v_\alpha} || \sum_{\beta\in v(\alpha)} \frac{s_{\alpha\beta}}{2}(A(\boldsymbol{\omega}_{\alpha\beta}) - D_{\alpha\beta}(\mathscr{U}))(U_\alpha - V_\alpha - U_\beta + V_\beta)||_2^2} \\
&\quad + \sqrt{\sum_\alpha \frac{1}{v_\alpha} || \sum_{\beta\in v(\alpha)} \frac{s_{\alpha\beta}}{2}(D_{\alpha\beta}(\mathscr{U}) - D_{\alpha\beta}(\mathscr{V}))(V_\alpha - V_\beta)||_2^2} \\
&\le \sqrt{\sum_\alpha \frac{s_\alpha}{v_\alpha} \sum_{\beta\in v(\alpha)} \frac{s_{\alpha\beta}}{4} |||A(\boldsymbol{\omega}_{\alpha\beta}) - D_{\alpha\beta}(\mathscr{U})|||_2^2 ||(U_\alpha - V_\alpha - U_\beta + V_\beta)||_2^2} \\
&\quad + \sqrt{\sum_\alpha \frac{s_\alpha}{v_\alpha} \sum_{\beta\in v(\alpha)} \frac{s_{\alpha\beta}}{4} |||D_{\alpha\beta}(\mathscr{U}) - D_{\alpha\beta}(\mathscr{V})|||_2^2 ||(V_\alpha - V_\beta)||_2^2} \\
&\le (\sigma_A + \sigma_D) \sqrt{\sum_\alpha v_\alpha \frac{s_\alpha}{4v_\alpha^2} \sum_{\beta\in v(\alpha)} s_{\alpha\beta} 2(||(U_\alpha - V_\alpha||_2^2 + ||U_\beta - V_\beta||_2^2)} \\
&\quad + K||\mathscr{U} - \mathscr{V}||_2 \sqrt{\sum_\alpha v_\alpha \frac{s_\alpha}{4v_\alpha^2} \sum_{\beta\in v(\alpha)} s_{\alpha\beta} 2(||V_\alpha||_2^2 + ||V_\beta||_2^2)} \\
&\le \sup_\alpha \frac{s_\alpha}{v_\alpha} \left((\sigma_A + \sigma_D) \frac{1}{\sqrt{2}} ||\mathscr{U} - \mathscr{V}||_2 + K||\mathscr{U} - \mathscr{V}||_2 \frac{1}{\sqrt{2}} ||\mathscr{V}||_2 \right),
\end{aligned}
$$

where K and σ_D are the Lipschitz constant and upper bound of the matrices $D_{\alpha\beta}(\mathscr{U})$, and $\sigma_A = \sup_{||\boldsymbol{\omega}||_2=1} |||A(\boldsymbol{\omega})|||_2$. Hence $\mathscr{F}$ is Lipschitz continuous, and $\mathscr{U}' + \mathscr{F}(\mathscr{U}) = 0$ admits a local solution according to the Cauchy-Lipschitz theorem. Now we derive the following energy estimate

$$
\begin{aligned}
\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}||\mathscr{U}||_2^2 &= \sum_\alpha v_\alpha {}^t U_\alpha U_\alpha' = -\sum_\alpha {}^t U_\alpha \sum_{\beta\in v(\alpha)} s_{\alpha\beta} F_{\alpha\beta} = -\sum_{f_{\alpha\beta}} s_{\alpha\beta} {}^t(U_\alpha - U_\beta) F_{\alpha\beta} \\
&= \frac{1}{2} \sum_{f_{\alpha\beta}} s_{\alpha\beta} {}^t(U_\alpha - U_\beta) A(\boldsymbol{\omega}_{\alpha\beta})(U_\alpha + U_\beta) - {}^t(U_\alpha - U_\beta) D_{\alpha\beta}(U_\alpha - U_\beta)
\end{aligned}
$$

$$= \frac{1}{2}\sum_{\alpha}\sum_{\beta\in\nu(\alpha)} s_{\alpha\beta}{}^t U_\alpha A(\boldsymbol{\omega}_{\alpha\beta})U_\alpha - \frac{1}{2}\sum_{f_{\alpha\beta}} s_{\alpha\beta}{}^t (U_\alpha - U_\beta) D_{\alpha\beta}(U_\alpha - U_\beta).$$

Since $\sum_{\beta\in\nu(\alpha)} s_{\alpha\beta}{}^t U_\alpha A(\boldsymbol{\omega}_{\alpha\beta})U_\alpha = {}^t U_\alpha A(\sum_{\beta\in\nu(\alpha)} s_{\alpha\beta}\omega_{\alpha\beta})U_\alpha = 0$ we obtain $\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}||\mathscr{U}||_2^2 = -\frac{1}{2}\sum_{f_{\alpha\beta}} s_{\alpha\beta}{}^t (U_\alpha - U_\beta) D_{\alpha\beta}(U_\alpha - U_\beta)$. Since the matrices $D_{\alpha\beta}$ are symmetric and non negative, $\forall t \in \mathbb{R}_+$, $||\mathscr{U}(t)||_2 \le ||\mathscr{U}(0)||_2$. Therefore the maximal solutions are bounded and hence defined for all $t \in \mathbb{R}_+$, and (5) follows. □

3 Weak Convergence of the Scheme

Lemma 1 *Consider an initial datum $U_0 \in L^1_{loc}(\mathbb{R}^d)^m$, a mesh $\mathscr{T}$. For any $\phi \in \mathscr{D}(\mathbb{R}^d \times \mathbb{R}_+)$, a discrete solution $\mathscr{U}(\mathbf{x}, t)$ of (2–3) satisfies*

$$-\int_{\mathbb{R}_+}\int_{\mathbb{R}^d} \partial_t\phi(\mathbf{x},t)\mathscr{U}(\mathbf{x},t)\mathrm{d}v\,\mathrm{d}t - \int_{\mathbb{R}_+}\sum_{\alpha} F(U_\alpha)\Big(\sum_{\beta\in\nu(\alpha)} s_{\alpha\beta}\frac{\phi_\alpha+\phi_\beta}{2}\boldsymbol{\omega}_{\alpha\beta}\Big)\,\mathrm{d}t$$
$$+\int_{\mathbb{R}^d}\phi(\mathbf{x},0)\mathscr{U}(\mathbf{x},0)\mathrm{d}v + \int_{\mathbb{R}_+}\frac{1}{2}\sum_{f_{\alpha\beta}} s_{\alpha\beta}\left(\phi_\alpha-\phi_\beta\right)D_{\alpha\beta}(U_\alpha-U_\beta)\,\mathrm{d}t = 0. \tag{6}$$

Proof Multiplying (2) by ϕ_α, and integrating in space, we obtain $\sum_\alpha v_\alpha\phi_\alpha U_\alpha(t)' + \sum_\alpha \phi_\alpha \sum_{\beta\in\nu(\alpha)} s_{\alpha\beta}F_{\alpha\beta}(\mathscr{U}(t)) = 0$. From (4) we deduce $\sum_\alpha v_\alpha\phi_\alpha U_\alpha(t)' - \frac{1}{2}\sum_\alpha\sum_{\beta\in\nu(\alpha)} s_{\alpha\beta}\phi_\alpha(A(\boldsymbol{\omega}_{\alpha\beta}) - D_{\alpha\beta}(\mathscr{U}))(U_\alpha - U_\beta) = 0$. Using $A(\boldsymbol{\omega}_{\alpha\beta}) = -A(\boldsymbol{\omega}_{\beta\alpha})$ and $D_{\alpha\beta}(\mathscr{U}) = D_{\beta\alpha}(\mathscr{U})$ we finally obtain

$$\sum_\alpha v_\alpha\phi_\alpha U_\alpha'(t) - \frac{1}{2}\sum_{f_{\alpha\beta}} s_{\alpha\beta}\left(\phi_\alpha+\phi_\beta\right)A(\boldsymbol{\omega}_{\alpha\beta})(U_\alpha-U_\beta)$$
$$+\frac{1}{2}\sum_{f_{\alpha\beta}} s_{\alpha\beta}\left(\phi_\alpha-\phi_\beta\right)D_{\alpha\beta}(U_\alpha-U_\beta) = 0. \tag{7}$$

The middle term in (7) can be rewritten into

$$\frac{1}{2}\sum_{f_{\alpha\beta}} s_{\alpha\beta}\left(\phi_\alpha+\phi_\beta\right)A(\boldsymbol{\omega}_{\alpha\beta})(U_\alpha-U_\beta) = \sum_\alpha\sum_{\beta\in\nu(\alpha)} s_{\alpha\beta}\frac{\phi_\alpha+\phi_\beta}{2}A(\boldsymbol{\omega}_{\alpha\beta})U_\alpha$$
$$= \sum_\alpha A\Big(\sum_{\beta\in\nu(\alpha)} s_{\alpha\beta}\frac{\phi_\alpha+\phi_\beta}{2}\boldsymbol{\omega}_{\alpha\beta}\Big)U_\alpha$$
$$= \sum_\alpha F(U_\alpha)\Big(\sum_{\beta\in\nu(\alpha)} s_{\alpha\beta}\frac{\phi_\alpha+\phi_\beta}{2}\boldsymbol{\omega}_{\alpha\beta}\Big).$$

Integrating (7) with regard to time and using integration by parts, we find

$$-\int_{\mathbb{R}_+}\sum_\alpha v_\alpha \phi'_\alpha(t)U_\alpha(t)\,\mathrm{d}t - \int_{\mathbb{R}_+}\sum_\alpha F(U_\alpha)(\sum_{\beta\in\nu(\alpha)} s_{\alpha\beta}\frac{\phi_\alpha+\phi_\beta}{2}\boldsymbol{\omega}_{\alpha\beta})\,\mathrm{d}t$$
$$+\sum_\alpha v_\alpha\phi_\alpha(0)U_\alpha(0) + \int_{\mathbb{R}_+}\frac{1}{2}\sum_{f_{\alpha\beta}} s_{\alpha\beta}\left(\phi_\alpha-\phi_\beta\right)D_{\alpha\beta}(U_\alpha-U_\beta)\,\mathrm{d}t = 0.$$

Since $\mathscr{U}(\mathbf{x},t)$ is piecewise constant, $\sum_\alpha v_\alpha\phi_\alpha(0)U_\alpha(0) = \int_{\mathbb{R}^d}\phi(\mathbf{x},0)\mathscr{U}(\mathbf{x},0)\mathrm{d}v$. Similarly, given that $\partial_t\phi\in\mathscr{D}(\mathbb{R}^d\times\mathbb{R}_+)$, and $\phi'_\alpha(t) = \frac{1}{v_\alpha}\int_{\mathscr{T}_\alpha}\partial_t\phi(\mathbf{x},t)\mathrm{d}v$ we find $\sum_\alpha v_\alpha\phi'_\alpha(t)U_\alpha(t) = \int_{\mathbb{R}^d}\partial_t\phi(\mathbf{x},t)\mathscr{U}(\mathbf{x},t)\mathrm{d}v$, which yields (6). □

Lemma 2 (Weak and strong convergence of discretised fields)
Let $U\in L^2(\mathbb{R}^d)^m$, and consider a mesh sequence $\mathscr{T}^n$, $n\in\mathbb{N}$ such that the maximum diameter d^n_α satisfies $\lim_{n\to\infty}\sup_\alpha d^n_\alpha = 0$.

- *The sequence of discretised fields $\mathscr{U}^n\in L^2(\mathscr{T}^n)^m$ with cell values $U^n_\alpha = \frac{1}{v^n_\alpha}\int_{\mathscr{T}^n_\alpha}U(\mathbf{x})\mathrm{d}v$ converges weakly to U.*
- *If U is smooth with compact support, then $\mathscr{U}^n$ converges strongly to U in $L^2(\mathbb{R}^d)^m$.*

We omit the proof which is a classical use of the Cauchy-Schwarz inequality.

Lemma 3 *Consider an initial datum $U_0\in L^2(\mathbb{R}^d)^m$, a mesh $\mathscr{T}$ with symmetric non negative upwinding matrices such that $D_{\alpha\beta}(\mathscr{U})\le\rho_D$ for some $\rho_D<\infty$, and a time dependent field $\mathscr{U}(t)\in L^2(\mathscr{T})^m$ satisfying (5). For any $\phi\in\mathscr{D}(\mathbb{R}^d\times\mathbb{R}_+)$ with compact support in $\Omega_\phi\times[0,T]$ and Lipschitz constant $K_\phi>0$ we have the bound*

$$\int_{\mathbb{R}_+}\sum_{f_{\alpha\beta}} s_{\alpha\beta}\left(\phi_\alpha-\phi_\beta\right)D_{\alpha\beta}(\mathscr{U})(U_\alpha-U_\beta)\,\mathrm{d}t \le (2\rho_D T K_\phi^2)^{\frac{1}{2}}\sqrt{\sum_{\mathscr{T}_\alpha\cap\Omega_\phi\neq\emptyset} s_\alpha d_\alpha^2}\,||\mathscr{U}(0)||_2.$$

Proof Cauchy-Schwarz inequality yields $\int_{\mathbb{R}_+}\sum_{f_{\alpha\beta}} s_{\alpha\beta}\left(\phi_\alpha-\phi_\beta\right)D_{\alpha\beta}(U_\alpha-U_\beta)\,\mathrm{d}t$ $\le\sqrt{\int_0^T\sum_{f_{\alpha\beta}} s_{\alpha\beta}(\phi_\alpha-\phi_\beta)^2\,\mathrm{d}t}\sqrt{\int_0^T\sum_{f_{\alpha\beta}} s_{\alpha\beta}||D_{\alpha\beta}(U_\alpha-U_\beta)||_2^2\,\mathrm{d}t}$, and from (5) $\int_0^T\sum_{f_{\alpha\beta}} s_{\alpha\beta}||D_{\alpha\beta}(U_\alpha-U_\beta)||_2^2\,\mathrm{d}t\le 2\rho_D||\mathscr{U}(0)||_2^2$. Now for each face $f_{\alpha\beta}$ we have $|\phi_\alpha(t)-\phi_\beta(t)|\le K_\phi(d_\alpha+d_\beta)$, and the lemma follows from

$$\sum_{f_{\alpha\beta}} s_{\alpha\beta}(\phi_\alpha-\phi_\beta)^2 \le \sum_{f_{\alpha\beta}\cap\Omega_\phi\neq\emptyset} K_\phi^2 s_{\alpha\beta}(d_\alpha+d_\beta)^2 \le 2K_\phi^2\sum_{f_{\alpha\beta}\cap\Omega_\phi\neq\emptyset} s_{\alpha\beta}(d_\alpha^2+d_\beta^2)$$
$$\le 2K_\phi^2\sum_{\mathscr{T}_\alpha\cap\Omega_\phi\neq\emptyset}\sum_{\beta\in\nu(\alpha)} s_{\alpha\beta}d_\alpha^2 = 2K_\phi^2\sum_{\mathscr{T}_\alpha\cap\Omega_\phi\neq\emptyset} s_\alpha d_\alpha^2.$$

□

Lemma 4 (Strong convergence of the discrete gradients) *Consider a mesh $\mathscr{T}$, and $\phi\in\mathscr{D}(\mathbb{R}^d\times\mathbb{R}_+)$, with compact support Ω_ϕ and Lipschitz constant K_ϕ. Then*

$$\sum_\alpha v_\alpha \left| \frac{1}{v_\alpha} \sum_{\beta\in\nu(\alpha)} s_{\alpha\beta} \frac{\phi_\alpha+\phi_\beta}{2} \boldsymbol{\omega}_{\alpha\beta} - \frac{1}{v_\alpha}\int_{\partial\mathscr{T}_\alpha} \phi \mathbf{ds} \right|^2 \le d_\alpha (\sup_\alpha \frac{s_\alpha^2 d_\alpha^2}{v_\alpha^2}) \left(\sum_{\mathscr{T}_\alpha\cap\Omega_\phi\neq\emptyset} v_\alpha \right)$$
$$\times \sup_{f_{\alpha\beta}} \left(K_{\phi'} \frac{||\frac{x_\alpha+x_\beta}{2} - x_{\alpha\beta})||}{d_\alpha^2} + \frac{K_{\phi''}}{2}(1+\frac{d_\beta^2}{d_\alpha^2}) \right). \tag{8}$$

Proof The convexity of $x \to |x|^2$ yields $\sum_\alpha \frac{1}{v_\alpha} \left| \sum_{\beta\in\nu(\alpha)} \frac{s_{\alpha\beta}}{s_\alpha} s_\alpha \left(\frac{\phi_\alpha+\phi_\beta}{2} - \phi_{\alpha\beta} \right) \omega_{\alpha\beta} \right|^2 \le \sum_\alpha \frac{1}{v_\alpha} \left(\sum_{\beta\in\nu(\alpha)} s_{\alpha\beta} s_\alpha \left| \frac{\phi_\alpha+\phi_\beta}{2} - \phi_{\alpha\beta} \right|^2 \right)$.
Using the Taylor expansion $\phi_\alpha = \frac{1}{v_\alpha}\int_{\mathscr{T}_\alpha} \phi(x_{\alpha\beta}) + \nabla\phi(x_{\alpha\beta})\cdot(x-x_{\alpha\beta}) + R_1^{x_{\alpha\beta}}(x)\mathrm{d}v$ and the Taylor-Lagrange inequality $|R_1^{x_{\alpha\beta}}(x)| \le K_\phi ||x_\alpha - x_{\alpha\beta}||^2$ we deduce

$$\left| \frac{\phi_\alpha+\phi_\beta}{2} - \phi_{\alpha\beta} \right| \le |\nabla\phi \cdot \left(\frac{x_\alpha+x_\beta}{2} - x_{\alpha\beta})| + \frac{K_{\phi''}}{2}(||x_\alpha - x_{\alpha\beta}||^2 + ||x_\beta - x_{\alpha\beta}||^2 \right).$$

(8) follows from $\sum_\alpha v_\alpha \left| \frac{1}{v_\alpha} \sum_{\beta\in\nu(\alpha)} s_{\alpha\beta} \frac{\phi_\alpha+\phi_\beta}{2} \boldsymbol{\omega}_{\alpha\beta} - \frac{1}{v_\alpha}\int_{\partial\mathscr{T}_\alpha} \phi \mathbf{ds} \right|^2 \le \sum_\alpha v_\alpha \frac{s_\alpha^2}{v_\alpha^2} \sup_{\beta\in\nu(\alpha)} (K_{\phi'} ||\frac{x_\alpha+x_\beta}{2} - x_{\alpha\beta})|| + \frac{K_{\phi''}}{2}(d_\alpha^2 + d_\beta^2))^2$. □

Theorem 2 (Weak convergence of semi discrete finite volume schemes)
Consider the Cauchy problem (1) associated to the initial data $U_0 \in L^2(\mathbb{R}^d)^m$. Consider a sequence of meshes $(\mathscr{T}^n)_{n\in\mathbb{N}}$, where each cell α of $\mathscr{T}^n$ has diameter d_α^n, surrounding area s_α^n, measure v_α^n and center of mass x_α^n, and each interface $f_{\alpha\beta}$ has area $s_{\alpha\beta}^n$ and center of mass $x_{\alpha\beta}^n$. We assume that $\lim_{n\to\infty} \sup_\alpha d_\alpha^n = 0$, $\forall n, \sup_\alpha \frac{s_\alpha^n}{v_\alpha^n} < \infty$ and $\exists c \in \mathbb{R}$ such that, $\forall n, \alpha, \beta \in \nu(\alpha)$

$$s_\alpha^n d_\alpha^n < c v_\alpha^n, \quad \frac{d_\beta^n}{d_\alpha^n} < c, \quad ||\tfrac{1}{2}(x_\alpha^n + x_\beta^n) - x_{\alpha\beta}^n)|| \le c(d_\alpha^n)^2.$$

Consider $(\mathscr{U}^n(\mathbf{x},t))_{n\in\mathbb{N}}$ a sequence of approximate solutions given by the scheme (2–3) on the meshes $(\mathscr{T}^n)_{n\in\mathbb{N}}$ with uniformly bounded and Lipschitz continuous symmetric non negative upwinding matrices $D_{\alpha\beta}^n(\mathscr{U}^n)$.

$(\mathscr{U}^n(\mathbf{x},t))_{n\in\mathbb{N}}$ converges weakly to the weak solution of (1) in $L^2(\mathbb{R}^d\times[0,T])^m$.

Proof From theorem (1), $(\mathscr{U}^n(\mathbf{x},t))_{n\in\mathbb{N}}$ is a bounded sequence in $L^2(\mathbb{R}^d\times[0,T])^m$ which is a Hilbert space for the inner product $< U, V >= \int_0^T \int_{\mathbb{R}^d} U\cdot V \,\mathrm{d}v\mathrm{d}t$. Therefore according to the Banach-Alaoglu theorem, the sequence $\mathscr{U}^n(\mathbf{x},t)$ is weakly convergent up to a subsequence to a function of $U(\mathbf{x},t) \in L^2(\mathbb{R}^d\times[0,T])^m$. Consider $\phi \in \mathscr{D}(\mathbb{R}^d\times\mathbb{R}_+)$, from lemma (1) we find

$$-\int_0^T\int_{\mathbb{R}^d}\partial_t\phi(\mathbf{x},t)\mathscr{U}^n(\mathbf{x},t)\mathrm{d}v\,\mathrm{d}t-\int_0^T\sum_\alpha F(U_\alpha^n)(\sum_{\beta\in v(\alpha)}s_{\alpha\beta}\frac{\phi_\alpha+\phi_\beta}{2}\boldsymbol{\omega}_{\alpha\beta})\,\mathrm{d}t$$
$$+\int_{\mathbb{R}^d}\phi(\mathbf{x},0)\mathscr{U}^n(\mathbf{x},0)\mathrm{d}v+\int_0^T\frac{1}{2}\sum_{f_{\alpha\beta}}s_{\alpha\beta}\left(\phi_\alpha-\phi_\beta\right)D_{\alpha\beta}(U_\alpha^n-U_\beta^n)\,\mathrm{d}t=0. \tag{9}$$

The last term in (9) is in $O(\sup_\alpha(d_\alpha^n)^{\frac{1}{2}})$ from lemma 3 since $s_\alpha^n d_\alpha^n \leq c v_\alpha^n$. The third term in (9) converges with order $O(\sup_\alpha d_\alpha^n)$ to $\int_{\mathbb{R}^d}\phi(\mathbf{x},0)U_0(\mathbf{x})\mathrm{d}v$ (lemma 2). The first term in (9) converges similarly with order $O(\sup_\alpha d_\alpha^n)$ to $\int_0^T\int_{\mathbb{R}^d}\partial_t\phi(\mathbf{x},t)U(\mathbf{x},t)\mathrm{d}v$ given that $\partial_t\phi\in\mathscr{D}(\mathbb{R}^d\times\mathbb{R}_+)$ (lemma 2).

As for the second term in (9), from the Green-Ostrogradski formula ($\int_{v_\alpha}\nabla\phi\mathrm{d}v=\int_{\partial\mathscr{T}_\alpha}\phi\mathrm{d}\mathbf{s}$), the discrete field $\alpha\to\frac{1}{v_\alpha}\int_{\partial\mathscr{T}_\alpha}\phi\mathrm{d}\mathbf{s}$ is the projection of $\nabla\phi$ on $\mathscr{T}^n$. Since $\nabla\phi$ is smooth with compact support, $\alpha\to\frac{1}{v_\alpha}\int_{\partial\mathscr{T}_\alpha}\phi\mathrm{d}\mathbf{s}$ converges strongly to $\nabla\phi$ in $L^2(\mathbb{R}^d)^m$ (lemma 1). From lemma 4, since $\lim_{n\to\infty}\sup_\alpha d_\alpha^n\sup_\alpha s_\alpha^n=0$, the field $\alpha\to\frac{1}{v_\alpha}\sum_{\beta\in v(\alpha)}s_{\alpha\beta}\frac{\phi_\alpha+\phi_\beta}{2}\boldsymbol{\omega}_{\alpha\beta}$ converges strongly to $\nabla\phi$.

Hence $\int_0^T\sum_\alpha F(U_\alpha^n)(\sum_{\beta\in v(\alpha)}s_{\alpha\beta}\frac{\phi_\alpha+\phi_\beta}{2}\boldsymbol{\omega}_{\alpha\beta})\,\mathrm{d}t$ shares the same limit with $\int_0^T\int_{\mathbb{R}^d}F(\mathscr{U}^n)\nabla\phi\,\mathrm{d}v\mathrm{d}t$. That limit is $\int_0^T\int_{\mathbb{R}^d}F(\mathscr{U})\nabla\phi\,\mathrm{d}v\mathrm{d}t$ since $\nabla\phi$ is a smooth function and $F(\mathscr{U}^n)$ converges weakly to $F(U)$ (F is linear).

Hence any limit point of $(\mathscr{U}^n(\mathbf{x},t))_{n\in\mathbb{N}}$ is a solution of (1), and the sequence $(\mathscr{U}^n(\mathbf{x},t))_{n\in\mathbb{N}}$ converges weakly in $L^2(\mathbb{R}^d\times[0,T])^m$ up to extracting a subsequence. From the uniqueness of the solution to (1) (see [10]), we deduce the convergence of the entire sequence of approximated solutions. □

References

1. Dao Thu, H., Ndjinga, M., Magoules, F.: Comparison of upwind and centered schemes for low Mach number flows. Finite Volumes Complex Appl VI Prob Perspect **4**, 303–311 (2011)
2. Dellacherie, S.: Checkerboard modes and the wave equation. Proceedings of ALGORITMY 2009 conference on scientific computing, Vysoke Tatry, Podbanske, Slovakia, pp. 71–80 (2009)
3. Dellacherie, S.: Analysis of godunov type schemes applied to the compressible Euler system at low Mach number. J. Comput. Phys. **229**(6), 978–1016 (2010)
4. Després, B.: Lax theorem and finite volume schemes. Math. Comput. **73**(247), 1203–1234 (2004)
5. Eymard, R., Gallouët, T., Herbin, R.: The finite volume method. In: Ciarlet, P.H., Lions, J.L. (eds.) Handbook for Numerical Analysis, North Holland (2000)
6. Jovanovic, V., Rohde, C.: Finite-volume schemes for Friedrichs systems in multiple space dimensions: *a priori* and *a posteriori* estimates. Numer. Methods Partial Differ. Equ. **21**(1), 104–131 (2005)
7. Kumbaro, A., Ndjinga, M.: Application of the simplified eigenstructure decomposition solver to the simulation of general multifield models. Nuc. Eng. Des. **261**, 56–65 (2013)
8. Leveque, R.J.: Numerical Methods for Conservation Laws. ETH Zürich, Birkhäuser, Basel (1990)

9. Ndjinga, M.: L^2 stability of nonlinear finite volume schemes for linear hyperbolic systems. C. R. Acad. Sci. Paris, Ser. **I 351**, 707–711 (2013)
10. Serre, D.: Systems of Conservation Laws I. Cambridge University Press, Cambridge (1999)
11. Vila, J.P., Villedieu, P.: Convergence de la méthode des volumes finis pour les systèmes de Friedrichs. C. R. Acad. Sci. Paris Série **I (325)**, 671–676 (1997)

A-Posteriori Error Estimates for the Localized Reduced Basis Multi-Scale Method

Mario Ohlberger and Felix Schindler

Abstract We present a localized a-posteriori error estimate for the Localized Reduced Basis Multi-Scale (LRBMS) method [1]. The LRBMS is a combination of numerical multi-scale methods and model reduction using reduced basis methods to efficiently reduce the computational complexity of parametric multi-scale problems with respect to the multi-scale parameter ε and the online parameter μ simultaneously. We formulate the LRBMS based on a generalization of the SWIPDG discretization presented in [2] on a coarse partition of the domain that allows for any suitable discretization on the fine triangulation inside each coarse grid element. The estimator is based on the idea of a conforming reconstruction of the discrete diffusive flux, presented in [2], that can be computed using local information only. It is offline/online decomposable and can thus be efficiently used in the context of model reduction.

1 Introduction

We are interested in efficient and reliable numerical approximations of parametric elliptic multi-scale problems for given parameters $\mu \in \mathscr{P} \subset \mathbb{R}^p$, for $p \in \mathbb{N}$, i.e.

$$-\nabla \cdot ({}_{\mu}\lambda^{\varepsilon}\kappa \cdot \nabla^{\varepsilon}_{\mu} p) = f \quad \text{in } \Omega, \tag{1}$$

with homogeneous Dirichlet boundary conditions, where ε indicates the multi-scale nature of the quantities in prefix notation. Equation (1) arises e.g. in the context of two-phase flow in porous media, where it needs to be solved in every timestep for different μ to obtain the global pressure ${}^{\varepsilon}_{\mu}p : \Omega \to \mathbb{R}$ (see [1, Sect. 1]). A discretization of (1) usually consists in finding an approximation ${}^{\varepsilon}_{\mu}p_h \in V_h$ by a Galerkin projection onto

M. Ohlberger · F. Schindler (✉)
Applied Mathematics, University of Münster, Einsteinstr 62, 48149 Münster, Germany
e-mail: felix.schindler@wwu.de

J. Fuhrmann et al. (eds.), *Finite Volumes for Complex Applications VII - Methods and Theoretical Aspects,* Springer Proceedings in Mathematics & Statistics 77,
DOI: 10.1007/978-3-319-05684-5_41,

a fine triangulation τ_h of Ω resolving the ε scale. Two traditional approaches exist to reduce the computational complexity of the discrete problem: numerical multi-scale methods and model order reduction techniques. Numerical multi-scale methods reduce the complexity of multi-scale problems with respect to ε, while model order reduction techniques reduce the complexity of parametric problems with respect to μ (see [3] for an overview). It is well known that solving parametric heterogeneous multi-scale problems accurately can be challenging and computationally costly, in particular for strongly varying scales and parameter ranges. In general, numerical multi-scale methods capture the macroscopic behavior of the solution in a coarse approximation space, e.g., $V_H \subset V_h$, usually associated with a coarse triangulation $\mathscr{T}_H$ of Ω, and recover the microscopic behavior of the solution by local fine-scale corrections. Model order reduction using reduced basis (RB) methods, on the other hand, is based on the idea to introduce a reduced space $V_{\text{red}} \subset V_h$, spanned by solutions of (4) for a limited number of parameters μ. These training parameters are iteratively selected by an adaptive Greedy procedure (see [1] and the reference therein). The idea of the recently presented localized reduced basis multi-scale (LRBMS) approach (see [1]) is to combine numerical multi-scale and RB methods and to generate a local reduced space $V_{\text{red}}^T \subset V_h^T$ for each coarse element of $\mathscr{T}_H$, given a tensor product type decomposition of the fine approximation space, $V_h = \oplus_{T \in \mathscr{T}_H} V_h^T$. The coarse reduced space is then given as $V_{H,\text{red}} := \oplus_{T \in \mathscr{T}_H} V_{\text{red}}^T \subset V_h$, resulting in a multiplicative decomposition of the solution into ${}^{\varepsilon}_{\mu}p_{H,\text{red}}(x) = \sum_{n=1}^{\dim(V_{H,\text{red}})} {}_{\mu}p_n(x)\, {}^{\varepsilon}\varphi_n(x)$, where the RB functions ${}^{\varepsilon}\varphi_n$ capture the microscopic behavior of the solution and the coefficient functions ${}_{\mu}p_n$ only vary on the coarse triangulation.

It is vital for an efficient and reliable use of RB as well as LRBMS methods to have access to an estimate on the model reduction error. Such an estimate is used to drive the adaptive Greedy basis generation during the offline phase of the computation and to ensure the quality of the reduced solution during the online phase. It is usually given by a residual based estimator involving the stability constant and the residual in a dual norm. It was shown in [1] that such an estimator can be successfully applied in the context of the LRBMS, but it was also pointed out that an estimator relying on global information might not be computationally feasible since too much work is required in the offline part of the computation.

The novelty of this contribution lies in a completely different approach to error estimation—at least in the context of RB methods. We make use of the ansatz of local error estimation presented in [2] which measures the error by a conforming reconstruction of the physical quantities involved, specifically the diffusive flux $-{}_{\mu}\lambda\, {}^{\varepsilon}\kappa \nabla {}^{\varepsilon}_{\mu}p$. This kind of local error estimation was proven to be very successful in the context of multi-scale problems and robust with respect to ε. We show in this work how we can transfer those ideas to the framework of the LRBMS to obtain an estimate of the error ${}_{\mu}|||\, {}^{\varepsilon}_{\mu}p - {}^{\varepsilon}_{\mu}p_{H,\text{red}} \,|||$. We would like to point out that we are able to estimate the error against the weak solution ${}^{\varepsilon}_{\mu}p$ in a parameter dependent energy norm while traditional RB-approaches only allow to estimate the model reduction error in a parameter independent norm and only against the discrete solution. In principle, this approach is able to turn the LRBMS method into a full multi-scale

approximation scheme, while traditional RB methods can only be seen as a model reduction technique. We would also like to point out that, to the best of our knowledge, this is the first work that makes use of local error information in the context of RB methods.

This work is organized as follows. Section 2 introduces the notation and presents the overall setting, the discretization and the LRBMS framework. We then carry out the error analysis for our multi-scale SWIPDG discretization in the parametric setting as well as the LRBMS method in Sect. 3 and state our main result in Thm. 1.

2 Problem Formulation, Discretization and Model Reduction

We consider linear elliptic problems of the form (1) in a bounded connected domain $\Omega \subset \mathbb{R}^d$, $d = 2, 3$, with polygonal boundary $\partial\Omega$ for a set of admissible parameters $\mathscr{P} \subset \mathbb{R}^p$, $p \in \mathbb{N}$.

Triangulations We require two nested partitions of Ω, a coarse one, $\mathscr{T}_H$, and a fine one, τ_h. Let τ_h be a simplicial triangulation of Ω with elements $t \in \tau_h$. In the context of multi-scale problems we call τ_h a *fine triangulation* if it resolves all features of the quantities involved in (1), specifically if ${}^{\varepsilon}\kappa^t := {}^{\varepsilon}\kappa\big|_t \in [L^\infty(t)]^{d\times d}$ is constant for all $t \in \tau_h$. We only require the coarse elements $T \in \mathscr{T}_H$ to be shaped such that a local Poincaré inequality in $H^1(T)$ is fulfilled (see Thm. 1) and collect in $\tau_h^T \subset \tau_h$ the fine elements of τ_h that cover the coarse element T. In addition, we collect all fine faces in $\mathscr{F}_h$, all coarse faces in $\mathscr{F}_H$ and denote by ${\mathscr{F}_H}^T \subset \mathscr{F}_H$ the faces of a coarse element $T \in \mathscr{T}_H$ and by ${\mathscr{F}_h}^E \subset \mathscr{F}_h$ the fine faces that cover a coarse face $E \in \mathscr{F}_H$.

The continuous problem We define the *broken Sobolev space* $H^1(\tau_h) \subset L^2(\Omega)$ by $H^1(\tau_h) := \left\{q \in L^2(\Omega) \;\middle|\; q|_t \in H^1(t) \quad \forall t \in \tau_h\right\}$, with $H_0^1(\Omega) \subset H^1(\Omega) \subset H^1(\tau_h)$, where H^1 denotes the usual Sobolev space of weakly differentiable functions and H_0^1 its elements which vanish on the boundary in the sense of traces. In the same manner we denote the local broken Sobolev spaces $H^1(\tau_h^T) \subset L^2(T)$ for all $T \in \mathscr{T}_H$. We also denote by $\nabla_h : H^1(\tau_h) \to [L^2(\Omega)]^d$ the *broken gradient operator* which is locally defined by $(\nabla_h q)|_t := \nabla(q|_t)$ for all $t \in \tau_h$ and $q \in H^1(\tau_h)$. Given $f \in L^2(\Omega)$, ${}_{\mu}\lambda \in C^0(\Omega)$ strictly positive and ${}^{\varepsilon}\kappa \in [L^\infty(\Omega)]^{d\times d}$ symmetric and uniformly positive definite, such that ${}_{\mu}\lambda\, {}^{\varepsilon}\kappa \in [L^\infty(\Omega)]^{d\times d}$ is bounded from below (away from 0) and above for all $\mu \in \mathscr{P}$, we define the parameter dependent bilinear form ${}^{\varepsilon}_{\mu}b : H^1(\tau_h) \times H^1(\tau_h) \to \mathbb{R}$ and the linear form $l : H^1(\tau_h) \to \mathbb{R}$ by ${}^{\varepsilon}_{\mu}b(p, q) := \sum_{T\in\mathscr{T}_H} {}^{\varepsilon}_{\mu}b^T(p, q)$ and $l(q) := \sum_{T\in\mathscr{T}_H} l^T(q)$, respectively, and their local counterparts ${}^{\varepsilon}_{\mu}b^T := H^1(\tau_h^T) \times H^1(\tau_h^T) \to \mathbb{R}$ and $l^T : H^1(\tau_h^T) \to \mathbb{R}$ for all $T \in \mathscr{T}_H$ and $\mu \in \mathscr{P}$ by

$$ {}^{\varepsilon}_{\mu}b^T(p, q) := \int_T ({}_{\mu}\lambda\, {}^{\varepsilon}\kappa \cdot \nabla_h p) \cdot \nabla_h q \,\mathrm{d}x \qquad \text{and} \qquad l^T(q) := \int_T f q \,\mathrm{d}x. $$

Definition 1 *(Weak solution)* Given a parameter $\mu \in \mathscr{P}$ we define the *weak solution* of (1) by ${}^{\varepsilon}_{\mu}p \in H^1_0(\Omega)$, such that

$$ {}^{\varepsilon}_{\mu}b({}^{\varepsilon}_{\mu}p, q) = l(q) \qquad \text{for all } q \in H^1_0(\Omega). \tag{2} $$

Note that, since ${}^{\varepsilon}_{\mu}b$ is continuous and coercive for all $\mu \in \mathscr{P}$ (due to the assumptions on ${}_{\mu}\lambda\, {}^{\varepsilon}\kappa$) and since l is bounded, there exists a unique solution of (2) due to the Lax-Milgram Theorem.

A note on parameters In addition to the assumptions we posed on ${}_{\mu}\lambda$ above we also demand it to be *affinely decomposable* with respect to $\mu \in \mathscr{P}$, i.e. there exist $\Xi \geq 1$ strictly positive *coefficients* ${}_{\xi}\theta : \mathscr{P} \to \mathbb{R}$ for $0 \leq \xi \leq \Xi - 1$ and ${}_{\Xi}\theta \in \{0, 1\}$ and $\Xi + 1$ nonparametric *components* ${}_{\xi}\lambda \in C^0(\Omega)$, such that ${}_{\mu}\lambda = \sum_{\xi=0}^{\Xi} {}_{\xi}\theta(\mu)\, {}_{\xi}\lambda$. We can then compare λ for two parameters $\mu, \overline{\mu} \in \mathscr{P}$ by ${}_{\mu,\overline{\mu}}\alpha\, {}_{\overline{\mu}}\lambda \leq {}_{\mu}\lambda \leq {}_{\mu,\overline{\mu}}\gamma\, {}_{\overline{\mu}}\lambda$, where ${}_{\mu,\overline{\mu}}\alpha := \min_{\xi=0}^{\Xi-1} {}_{\xi}\theta(\mu)\, {}_{\xi}\theta(\overline{\mu})^{-1}$ and ${}_{\mu,\overline{\mu}}\gamma := \max_{\xi=0}^{\Xi-1} {}_{\xi}\theta(\mu)\, {}_{\xi}\theta(\overline{\mu})^{-1}$ denote the positive equivalence constants. This assumption on the data function ${}_{\mu}\lambda$ is a common assumption in the context of RB methods and covers a wide range of physical problems. If ${}_{\mu}\lambda$ does not exhibit such a decomposition one can replace ${}_{\mu}\lambda$ by an arbitrary close approximation using Empirical Interpolation techniques (see [1] and the references therein) which does not impact our analysis. All quantities that linearly depend on ${}_{\mu}\lambda$ inherit the above affine decomposition in a straightforward way. Since we would like to estimate the error in a problem dependent norm we also need the notion of a *parameter dependent energy norm* ${}^{\varepsilon}_{\mu}|||\cdot||| : H^1(\tau_h) \to \mathbb{R}$ for any $\mu \in \mathscr{P}$, defined by ${}^{\varepsilon}_{\mu}|||q||| := \left(\sum_{T \in \mathscr{T}_H} {}^{\varepsilon}_{\mu}|||q|||_T{}^2\right)^{1/2}$ with ${}^{\varepsilon}_{\mu}|||q|||_T := \left({}^{\varepsilon}_{\mu}b^T(q, q)\right)^{1/2}$, for all $T \in \mathscr{T}_H$. Note that ${}^{\varepsilon}_{\mu}|||\cdot|||$ is a norm only on $H^1_0(\Omega)$. We can compare these norms for any two parameters $\mu, \overline{\mu} \in \mathscr{P}$ using the above decomposition of ${}_{\mu}\lambda$:

$$ \sqrt{{}_{\mu,\overline{\mu}}\alpha}\ {}^{\varepsilon}_{\overline{\mu}}|||\cdot||| \quad \leq \quad {}^{\varepsilon}_{\mu}|||\cdot||| \quad \leq \quad \sqrt{{}_{\mu,\overline{\mu}}\gamma}\ {}^{\varepsilon}_{\overline{\mu}}|||\cdot||| \tag{3} $$

We denote by $0 < {}^{\varepsilon}_{\mu}c_t \leq {}^{\varepsilon}_{\mu}C_t$ the smallest and largest eigenvalue of ${}_{\mu}\lambda^t\, {}^{\varepsilon}\kappa^t$ and additionally define $0 < {}^{\varepsilon}c^t := \min_{\mu \in \mathscr{P}} {}^{\varepsilon}_{\mu}c^t$, ${}^{\varepsilon}c^t < {}^{\varepsilon}C^t := \max_{\mu \in \mathscr{P}} {}^{\varepsilon}_{\mu}C^t$ for all $t \in \tau_h$. From here on we denote an a-priori chosen parameter by $\hat{\mu} \in \mathscr{P}$ while $\overline{\mu} \in \mathscr{P}$ denotes an arbitrary parameter and $\mu \in \mathscr{P}$ denotes the parameter during the online phase of the simulation.

The generalized SWIPDG discretization We discretize (2) by allowing for a suitable discretization of at least first order inside each coarse element $T \in \mathscr{T}_H$ and by coupling those with a symmetric weighted interior penalty (SWIP) discontinuous Galerkin (DG) discretization along the coarse faces of $\mathscr{T}_H$. We give a very brief definition of the SWIPDG bilinear form, see [2, Sect. 2.3] and the references therein for a detailed discussion and the definition of $[\![\cdot]\!]_e$, $\{\!\{\cdot\}\!\}_\omega$ and γ_e. For any two-valued function $q \in H^1(\tau_h)$, we define its multi-scale jump $[\![q]\!]_E$ and its multi-scale average ${}^{\varepsilon}\{\!\{q\}\!\}_E$ for all coarse faces $E \in \mathscr{F}_H$ locally by $[\![q]\!]_E\big|_e := [\![q]\!]_e$ and ${}^{\varepsilon}\{\!\{q\}\!\}_E\big|_e := \{\!\{q\}\!\}_\omega$ for all $e \in \mathscr{F}_h{}^E$. In addition we define the multi-scale penalty parameter ${}^{\varepsilon}_{\mu}\sigma_E$ locally

by ${}^{\varepsilon}_{\mu}\sigma_E\big|_e := {}_{\mu}\lambda\gamma_e$ for all fine faces $e \in \mathscr{F}_h{}^E$ on all coarse faces $E \in \mathscr{F}_H$. With these definitions at hand we define the *multi-scale SWIPDG* bilinear form ${}^{\varepsilon}_{\mu}b_h : H^1(\tau_h) \times H^1(\tau_h) \to \mathbb{R}$ by

$$ {}^{\varepsilon}_{\mu}b_h(p,q) := \sum_{T\in\mathscr{T}_H} {}^{\varepsilon}_{\mu}b_h^T(p,q) + \sum_{E\in\mathscr{F}_H} {}^{\varepsilon}_{\mu}b_h^E(p,q), $$

with the coupling bilinear forms ${}^{\varepsilon}_{\mu}b_h^E : H^1(\tau_h^T) \times H^1(\tau_h^S) \to \mathbb{R}$ given by

$$ {}^{\varepsilon}_{\mu}b_h^E(p,q) := \int_E -\left\{\!\!\left\{ ({}_{\mu}\lambda\, {}^{\varepsilon}\kappa \cdot \nabla_h q)\cdot n_E \right\}\!\!\right\}_E [\![p]\!]_E - \left({}^{\varepsilon}\left\{\!\!\left\{ ({}_{\mu}\lambda^{\varepsilon}\kappa \cdot \nabla_h p)\cdot n_E \right\}\!\!\right\}_E - {}^{\varepsilon}_{\mu}\sigma_E [\![p]\!]_E \right) [\![q]\!]_E \,\mathrm{dx} $$

for all $E = \partial T \cap \partial S \in \mathscr{F}_H$. To complete the definition of the discretization we only demand the local bilinear forms ${}^{\varepsilon}_{\mu}b_h^T$ to be an approximation of ${}^{\varepsilon}_{\mu}b^T$ (with homogeneous Neumann boundary values) and the local discrete ansatz spaces $V_h^{k,T}$ to be locally polynomial of order $k \geq 1$, i.e. $q|_t \in \mathbb{P}_k(t)$ for all $t \in \tau_h^T$ and $q \in V_h^{k,T}$ on all $T \in \mathscr{T}_H$. We then define the multi-scale DG approximation space as $V_h^k(\mathscr{T}_H) := \left\{ q \in H^1(\tau_h) \middle|\, q|_T \in V_h^{k,T} \ \ \forall T \in \mathscr{T}_H \right\} \subset H^1(\tau_h)$ for $k \geq 1$.

Definition 2 (*Multi-scale DG approximation*) Given a parameter $\mu \in \mathscr{P}$ we define the *multi-scale DG approximation* of (2) by ${}^{\varepsilon}_{\mu}p_h \in V_h^1(\mathscr{T}_H)$, such that

$$ {}^{\varepsilon}_{\mu}b_h({}^{\varepsilon}_{\mu}p_h, q_h) = l(q_h) \qquad \text{for all } q_h \in V_h^1(\mathscr{T}_H). \tag{4} $$

The bilinear form ${}^{\varepsilon}_{\mu}b_h$ is continuous and coercive if the penalty parameter is chosen large enough and if the sum of the local bilinear forms is continuous and coercive. If those are chosen accordingly the discrete problem (4) thus has a unique solution. Possible choices for the local bilinear forms ${}^{\varepsilon}_{\mu}b_h^T$ and the local approximation spaces $V_h^{k,T}$ include continuous Finite Elements and variants of the IPDG and the SWIPDG discretizations.

The localized reduced basis multi-scale method Since the global (in a spatial sense) model reduction ansatz of classical RB methods does not always fit in the context of multi-scale problems, the LRBMS introduced in [1] takes a more localized approach to model reduction. We refer to [1] for a detailed definition of the LRBMS and only state what is needed for the error analysis here. The main idea of the LRBMS is to restrict solutions of (4) for some μ to the elements of the coarse triangulation and to form local reduced spaces $V_{\mathrm{red}}^T \subset V_h^{k,T}$ by a local compression of those solution snapshots. Given these local reduced spaces we define the broken reduced space by $V_{H,\mathrm{red}} := \oplus_{T\in\mathscr{T}_H} V_{\mathrm{red}}^T \subset V_h^k(\mathscr{T}_H)$. The LRBMS approximation is then given by a standard Galerkin projection of (4).

Definition 3 (*LRBMS approximation*) Given a parameter $\mu \in \mathscr{P}$ we define the *LRBMS approximation* of (2) by ${}^{\varepsilon}_{\mu}p_{H,\mathrm{red}} \in V_{H,\mathrm{red}}$, such that

$$ {}^{\varepsilon}_{\mu} b_h({}^{\varepsilon}_{\mu} p_{H,\mathrm{red}}, q_H) = l(q_H) \qquad \text{for all } q_H \in V_{H,\mathrm{red}}. \tag{5} $$

3 Error Analysis

Our error analysis is a generalization of the ansatz presented in [2] to provide an estimator for our multi-scale DG approximation solving (4) as well as for our LRBMS approximation solving (5). We transfer the idea of a conforming reconstruction of the nonconforming discrete diffusive flux $-{}_{\mu}\lambda^{\varepsilon}\kappa\nabla_h p_h$ to our setting. Our error analysis shares some similarities with the general multi-scale ansatz presented in [4], which is stated for a wide range of discretizations but for a different coupling strategy.

We obtain the mild requirement for the local approximation spaces that the constant function $\mathbb{1}$ is present, which is obvious for traditional discretizations and can be easily achieved for the LRBMS approximation by incorporating the DG basis with respect to $\mathscr{T}_H$. The estimates are fully offline/online decomposable and can thus be used for efficient model reduction in the context of the LRBMS.

We begin by stating an *abstract energy norm estimate* (see [2, Lemma 4.1]) that splits the difference between the weak solution ${}^{\varepsilon}_{\mu} p \in H^1_0(\Omega)$ solving (2) and any function $p_h \in H^1(\tau_h)$ into two contributions. This abstract estimate does not depend on our discretization and thus leaves the choice of s and u open. Note that we formulate the following Lemma with separate parameters for the energy norm and the weak solution. The price we have to pay for this flexibility are the additional constants involving ${}_{\mu,\overline{\mu}}\alpha$ and ${}_{\mu,\overline{\mu}}\gamma$, that vanish if $\overline{\mu}$ and μ coincide.

Lemma 1 (Abstract energy norm estimate) *Given any $\mu, \overline{\mu} \in \mathscr{P}$ let ${}^{\varepsilon}_{\mu} p \in H^1_0(\Omega)$ be the weak solution solving (2) and let $p_h \in H^1(\tau_h)$ be arbitrary. Then*

$$ \begin{aligned} {}^{\varepsilon}_{\overline{\mu}}\!\left|\!\left|\!\left| {}^{\varepsilon}_{\mu} p - p_h \right|\!\right|\!\right| &\leq \frac{1}{\sqrt{{}_{\mu,\overline{\mu}}\alpha}} \Bigg(\inf_{s \in H^1_0(\Omega)} \sqrt{{}_{\mu,\overline{\mu}}\gamma}\, {}^{\varepsilon}_{\overline{\mu}}\!\left|\!\left|\!\left| p_h - s \right|\!\right|\!\right| \\ &\quad + \inf_{u \in H_{div}(\Omega)} \Bigg\{ \sup_{\substack{\varphi \in H^1_0(\Omega) \\ {}^{\varepsilon}_{\mu}\!\left|\!\left|\!\left| \varphi \right|\!\right|\!\right| = 1}} \left\{ \left(f - \nabla\cdot u, \varphi\right)_{L^2} - \left({}_{\mu}\lambda^{\varepsilon}\kappa \cdot \nabla_h p_h + u, \nabla\varphi\right)_{L^2} \right\} \Bigg) \Bigg) \\ &\leq \frac{\sqrt{{}_{\mu,\overline{\mu}}\gamma}}{\sqrt{{}_{\mu,\overline{\mu}}\alpha}}\; 2\; {}^{\varepsilon}_{\overline{\mu}}\!\left|\!\left|\!\left| {}^{\varepsilon}_{\mu} p - p_h \right|\!\right|\!\right|. \end{aligned} $$

The above Lemma is proven by applying the norm equivalence (3), following the arguments in the proof of [2, Lemma 4.1] and applying the norm equivalence again.

The next Thm. states the main localization result and gives an indication on how to proceed with the choice of u: it allows us to localize the estimate of the above Lemma, if $u_h \in H_{\mathrm{div}}(\Omega) := \left\{ v \in [L^2(\Omega)]^{d\times d} \;\middle|\; \nabla\cdot v \in L^2(\Omega) \right\}$ fulfills a *local conservation property*.

Theorem 1 (Locally computable abstract energy norm estimate) *Let* ${}^{\varepsilon}_{\mu}p \in H^1_0(\Omega)$ *be the weak solution solving* (2)*, let* $s \in H^1_0(\Omega)$ *and* $p_h \in H^1(\tau_h)$ *be arbitrary, let* $u \in H_{div}(\Omega)$ *fulfill the* local conservation property $(\nabla\cdot u, \mathbb{1})_T = (f, \mathbb{1})_T$ *and let* $C_P^T > 0$ *denote the constant from the Poincaré inequality* $\|\varphi - \Pi_0^T\varphi\|^2_{L^2,T} \leq C_P^T h_T^2 \|\nabla\varphi\|^2_{L^2,T}$ *for all* $\varphi \in H^1(T)$ *on all* $T \in \mathscr{T}_H$*, where* Π_l^ω *denotes the* L^2*-orthogonal projection onto* $\mathbb{P}_l(\omega)$ *for* $l \in \mathbb{N}$ *and* $\omega \subseteq \Omega$*. It then holds that*

$$
\begin{aligned}
{}^{\varepsilon}_{\overline{\mu}}|||{}^{\varepsilon}_{\mu}p - p_h||| \quad &\leq \quad \eta[s,u], \qquad \textit{with the} \text{ global estimator } \eta[s,u] \textit{ defined as} \\
\eta[s,u] &:= \frac{\sqrt{\mu,\overline{\mu}^{\gamma}}}{\sqrt{\mu,\overline{\mu}^{\alpha}}}\Big(\sum_{T\in\mathscr{T}_H} \eta_{nc}^T[s]^2\Big)^{1/2} \\
&\quad + \frac{1}{\sqrt{\mu,\overline{\mu}^{\alpha}}}\Big(\sum_{T\in\mathscr{T}_H} \eta_r^T[u]^2\Big)^{1/2} \\
&\quad + \frac{\max\left(\sqrt{\mu,\hat{\mu}^{\gamma}}, \sqrt{\mu,\hat{\mu}^{\alpha}}^{-1}\right)}{\sqrt{\mu,\overline{\mu}^{\alpha}}}\Big(\sum_{T} \in \mathscr{T}_H \eta_{df^T}[u]^2\Big)^{1/2}
\end{aligned}
$$

and the local nonconformity estimator *given by* $\eta_{nc}^T[s] := \frac{\varepsilon}{\overline{\mu}}|||p_h - s|||_T$*, the local* residual estimator *given by* $\eta_r^T[u] := ({C_P^T}/{}^{\varepsilon}c^T)^{1/2} h_T \|f - \nabla\cdot u\|_{L^2,T}$ *and the local* diffusive flux estimator *given by* $\eta_{df}^T[u] := \left\|(_{\hat{\mu}}\lambda\,{}^{\varepsilon}\kappa)^{1/2}\nabla_h p_h + (_{\hat{\mu}}\lambda\,{}^{\varepsilon}\kappa)^{-1/2}u\right\|_{L^2,T}$ *for all coarse elements* $T \in \mathscr{T}_H$*, where* ${}^{\varepsilon}c^T := (\max_{t\in\tau_h^T} 1/{}^{\varepsilon}c^t)^{-1}$.

The above Thm. is proven by loosely following the proof of [2, Thm. 3.1], i.e. by starting from Lem. 1, localizing with respect to $\mathscr{T}_H$, using the local conservation property and the norm equivalence (3).

What is left now in order to turn the abstract estimate of Thm. 1 into a fully computable one is to specify s and u. We will do so in the following paragraphs.

Oswald interpolation Given any nonconforming approximation $p_h \in V_h^k(\mathscr{T}_H) \not\subset H^1_0(\Omega)$ we will choose $s \in H^1_0(\Omega)$ as a conforming reconstruction of p_h by the *Oswald Interpolation operator* $I_{\text{os}} : V_h^1(\mathscr{T}_H) \to V_h^1(\mathscr{T}_H) \cap H^1_0(\Omega)$ which we define by prescribing its values on the Lagrange nodes of the triangulation (see [2, Sect. 2.5] and the references therein): we define $I_{\text{os}}[p_h](\nu) := p_h^t(\nu)$ inside any $t \in \tau_h$ and

$$
I_{\text{os}}[p_h](\nu) := \frac{1}{|\tau_h^\nu|}\sum_{t\in\tau_h^\nu} p_h^t(\nu) \quad \text{for all inner nodes of } \tau_h \text{ and } \quad I_{\text{os}}[p_h](\nu) := 0
$$

for all boundary nodes of τ_h, where $\tau_h^\nu \subset \tau_h$ denotes the set of all simplices of the fine triangulation which share ν as a node.

Diffusive flux reconstruction As mentioned above we will reconstruct a conforming diffusive flux approximation $u_h \in H_{\text{div}}(\Omega)$ of the nonconforming discrete diffu-

sive flux $-{}_{\mu}\lambda^{\varepsilon}\kappa\nabla_h p_h \notin H_{\mathrm{div}}(\Omega)$ in a conforming discrete subspace $RTN_h^l(\tau_h) \subset H_{\mathrm{div}}(\Omega)$, namely the *Raviart-Thomas-Nédélec* space of vector functions (see [2] and the references therein), defined for $k-1 \leq l \leq k$ by

$$RTN_h^l(\tau_h) := \left\{ v \in H_{\mathrm{div}}(\Omega) \middle|\ v|_t \in RTN^l(t) := [\mathbb{P}_l(t)]^d + \mathbf{x}\mathbb{P}_l(t) \quad \forall t \in \tau_h \right\}.$$

See [2, Sect. 2.4] and the references therein for a detailed discussion of the role of the polynomial degree l, the properties of elements of $RTN_h^l(\tau_h)$ and the origin of the use of diffusive flux reconstructions in the context of error estimation in general. Now, given any $p_h \in H^1(\tau_h)$ and any $\mu \in \mathscr{P}$ we define the diffusive flux reconstruction ${}^{\varepsilon}_{\mu}u_h[p_h] \in RTN_h^l(\tau_h)$ locally by demanding

$$\left({}^{\varepsilon}_{\mu}u_h[p_h]\cdot n_E, q\right)_{L^2,E} = \left(-{}^{\varepsilon}\left\{\left({}_{\mu}\lambda^{\varepsilon}\kappa\cdot\nabla_h p_h\right)\cdot n_E\right\}_E + \sigma_E \left[\!\left[p_h\right]\!\right]_E, q\right)_{L^2,E}$$

for all $q \in \mathbb{P}_l(e)$ for all $e \in \mathscr{F}_h{}^E$ and all $E \in \mathscr{F}_H{}^T$ and by

$$\left({}^{\varepsilon}_{\mu}u_h[p_h], \nabla_h q\right)_{L^2,T} = -{}^{\varepsilon}_{\mu}b_h^T(p_h, q) + \sum_{E\in\mathscr{F}_H{}^T}\left({}^{\varepsilon}\omega_E^+({}_{\mu}\lambda^{T\,\varepsilon}\kappa^T\cdot\nabla_h q)\cdot n_E, \left[\!\left[p_h\right]\!\right]_E\right)_{L^2,E}$$

for all $q \in V_h^{k,T}$ such that $\nabla q|_t \in [\mathbb{P}_{l-1}(t)]^d$ for all $t \in \tau_h^T$ and all $T \in \mathscr{T}_H$. The next Lemma shows that this reconstruction of the diffusive flux is sensible for a multi-scale approximation as well as an LRBMS approximation, since the reconstructions of both fulfill the requirements of Thm. 1.

Lemma 2 (Local conservativity) *Let ${}^{\varepsilon}_{\mu}p_* \in H^1(\tau_h)$ either denote a multi-scale DG approximation ${}^{\varepsilon}_{\mu}p_h \in V_h^1(\mathscr{T}_H)$ given by (4) or an LRBMS approximation ${}^{\varepsilon}_{\mu}p_{H,red} \in V_{H,red}$ given by (5). Let ${}^{\varepsilon}_{\mu}u_h[{}^{\varepsilon}_{\mu}p_*] \in RTN_h^l(\tau_h)$ denote its diffusive flux reconstruction and let $\mathbb{1} \in V^{*,T}$, where $V^{*,T}$ either denotes the local approximation space $V_h^{1,T}$ or the local reduced space V_{red}^T, for all $T \in \mathscr{T}_H$. It then holds that ${}^{\varepsilon}_{\mu}u_h[{}^{\varepsilon}_{\mu}p_*]$ fulfills the local conservation property of Thm. 1.*

The above Lemma is proven by applying the ideas of [2, Lemma 2.1] to our setting while accounting for $\mathscr{T}_H$, i.e. by using the local conservation property, the definition of the discrete bilinear form and the fact, that $\mathbb{1} \in V^{*,T}$. At this points some remarks are in order. If we drop the parameter dependency and set $\mathscr{T}_H = \tau_h$, we obtain the discretization proposed in [2] and the estimators of Thm. 1 and [2, Thm.3.1] coincide. The estimators defined in Thm. 1 can be efficiently offline/online decomposed, even if we choose $\overline{\mu} = \mu$. A more elaborate work containing the proofs and the efficiency of the estimator (using standard arguments) is in preparation.

We finally obtain a fully computable and fully specified estimate by combining the definition of the Oswald interpolant and the diffusive flux reconstruction with Thm. 1 for both our multi-scale DG discretization and the LRBMS method.

References

1. Albrecht, F., Haasdonk, B., Kaulmann, S., Ohlberger, M.: The localized reduced basis multiscale method. In: Proceedings of Algoritmy 2012. Conference on Scientific Computing, Vysoke Tatry, Podbanske, 9–14 September, pp. 393–403. Slovak University of Technology in Bratislava, Publishing House of STU (2012)
2. Ern, A., Stephansen, A.F., Vohralík, M.: Guaranteed and robust discontinuous galerkin a posteriori error estimates for convection-diffusion-reaction problems. J. comput. Appl. Math. **234**(1), 114–130 (2010)
3. Ohlberger, M.: Error control based model reduction for multiscale problems. In: Proceedings of Algoritmy 2012. Conference on Scientific Computing, Vysoke Tatry, Podbanske, 9–14 September, pp. 1–10. Slovak University of Technology in Bratislava, Publishing House of STU (2012) (2012)
4. Pencheva, G.V., Vohralík, M., Wheeler, M.F., Wildey, T.: Robust a posteriori error control and adaptivity for multiscale, multinumerics, and mortar coupling. SIAM J. Numer. Anal. **51**(1), 526–554 (2013)

Chapter 42
Positivity Preserving Implicit and Partially Implicit Time Integration Methods in the Context of the DG Scheme Applied to Shallow Water Flows

Sigrun Ortleb

Abstract This contribution is concerned with the development of unconditionally positive implicit time integration schemes in the context of shallow water flows discretized by the DG scheme. For explicit time integration—which is mostly applied in combination with wetting and drying shallow water flows—both linear stability and positivity preservation require very small time steps. Also for implicit Runge-Kutta schemes, positivity preservation generally leads to additional time step restrictions. In this work, we discuss two possible extensions to implicit time integration schemes that guarantee non-negativity of the water height for any time step size while still preserving the conservativity of the space discretization.

1 Introduction

The shallow water equations (SWE) represent an important model in many scientific and engineering applications. They can be used to provide realistic simulations of flows in rivers, lakes or coastal areas, where the incorporation of arbitrary non-flat bottom topography is absolutely necessary. If the bottom topography is assumed to be constant with respect to time, the SWE are given by

$$\begin{aligned} \partial_t \varphi + \partial_{x_1}(\varphi v_1) + \partial_{x_2}(\varphi v_2) &= 0, \\ \partial_t(\varphi v_1) + \partial_{x_1}\left(\varphi v_1^2 + \frac{1}{2}\varphi^2\right) + \partial_{x_2}(\varphi v_1 v_2) &= -g\varphi\, \partial_{x_1} b, \\ \partial_t(\varphi v_2) + \partial_{x_1}(\varphi v_1 v_2) + \partial_{x_2}\left(\varphi v_2^2 + \frac{1}{2}\varphi^2\right) &= -g\varphi\, \partial_{x_2} b, \end{aligned} \tag{1}$$

S. Ortleb (✉)
Fachbereich Mathematik und Naturwissenschaften, Universität Kassel, Heinrich Plett Str. 40, 34132 Kassel, Germany
e-mail: ortleb@mathematik.uni-kassel.de

J. Fuhrmann et al. (eds.), *Finite Volumes for Complex Applications VII - Methods and Theoretical Aspects*, Springer Proceedings in Mathematics & Statistics 77
DOI: 10.1007/978-3-319-05684-5_42,

where b denotes the bottom elevation, the geopotential $\varphi = gH$ is composed of the water height H above the bottom and the gravitational constant $g = 9.812$ and $\mathbf{v} = (v_1, v_2)^T$ denotes the velocity vector. Here, we consider the numerical solution of this system of balance laws by the discontinuous Galerkin (DG) method [6, 10].

The SWE as in (1) contain source terms due to arbitrary bottom topography. Therefore, an important challenge is posed by steady states that need to be preserved by the numerical method. If the non-zero flux gradients are exactly balanced by the source term, a well-balanced scheme is required. In addition, the DG scheme has to guarantee non-negativity of the water height H. Regarding these demands, a positivity preserving and well-balanced third-order DG scheme on unstructured triangular grids using explicit time integration and adaptive modal filtering as a damping procedure was developed in [12] based on the method of Xing et al. [17].

However, for explicit time integration, linear stability requires very small time steps, especially for high polynomial degrees. In addition, for locally refined grids, e.g. refinement in wetting and drying regions, the system resulting from space discretization becomes even more stiff which makes implicit time stepping absolutely necessary in such cases. Unfortunately, the positivity preserving approach of Xing et al. [17] can not be applied in a straigthforward manner as it needs to enforce non-negative cell means of water height under rather restrictive time step constraints that also depend on the order of the DG discretization. This is due to the fact that the scheme exploits so-called strong stability preserving (SSP) properties of certain Runge-Kutta time integrations schemes. Also for implicit Runge-Kutta schemes, this generally leads to additional time step restrictions. As a lot of computational time is required to solve large systems of nonlinear equations within each time step, e.g. three nonlinear systems per time step for a third-order time integration scheme, the implicit scheme needs to be allowed much larger time steps to be able to beat explicit time stepping in terms of CPU time.

In the literature, for a computational treatment of wetting and drying, explicit time stepping is implemented in the majority of previous work. A special unconditionally positive implicit time integration scheme is used by Casulli [5]. This approach leads to a mildly nonlinear system to be solved each time step but is mass conserving and guarantees nonnegative water height for any time step size. However, this method is only first order accurate in space and time. A different approach is taken by Kärnä et al. [11]. There, the bottom topography is allowed to move in time as water elevation drops, i.e. a user-defined function is introduced which redefines the bottom topography. However, this function has to fulfill certain conditions and hence has to be carefully chosen prior to numerical computation. In the context of stabilized residual distribution schemes, Ricchiuto and Bollermann [15] developed a well-balanced and positivity preserving scheme for shallow water flows also considering implicit time integration via the second order trapezoidal rule. In this case, the time step size can be chosen twice as large as for the explicit Euler scheme.

In this contribution, we discuss two possible extensions to the strategy of positivity preservation in [17]. Both modifications are applied to a third-order SDIRK method and guarantee non-negativity of the water height for any time step size while still preserving conservativity. The first approach is the MPSDIRK3 scheme described

in [13] which is based on the so-called Patankar trick [3, 14]. The second approach, called IRSDIRK3, is a new contribution using an iterative redistribution of the water column. This novel idea extends the work of [2] to implicit schemes. The proposed methods are not restricted to unstructured grids but can easily be adapted to structured cartesian or triangular grids.

A numerical comparison of the above approaches is carried out, confronting them to the third-order TVD-RK explicit scheme [16]. Due to the proposed modifications, the implicit scheme can take full advantage of larger time steps and is therefore able to beat explicit time stepping in terms of CPU time. The new approach by iterative redistribution is also suitable for IMEX time integration. In this context, we give preliminary numerical results for the first order finite volume discretization, i.e. the DG scheme for $N = 0$, combined with a first order IMEX time integration method.

2 The DG Space Discretization

We rewrite the shallow water equations (1) in the more compact form

$$\frac{\partial}{\partial t}\mathbf{u}(\mathbf{x},t) + \nabla \cdot \mathbf{F}(\mathbf{u}(\mathbf{x},t)) = \mathbf{s}(\mathbf{u}(\mathbf{x},t),\mathbf{x}), \tag{2}$$

for $(\mathbf{x}, t) \in \Omega \times \mathbb{R}_+$, where $\Omega \subset \mathbb{R}^2$ is an open polygonal domain and the conservative variables are now collected in $\mathbf{u} = (\varphi, \varphi v_1, \varphi v_2)^T$, while $\mathbf{F}$ contains the fluxes and $\mathbf{s}$ the sources due to bottom topography. Let now T^h be a conforming triangulation of $\overline{\Omega}$ and let W^h be the piecewise polynomial space defined by $W^h = \{w_h \in L^\infty(\Omega) \mid w_h|_{\tau_i} \in P^N(\tau_i) \quad \forall\, \tau_i \in T^h\}$, where $P^N(\tau_i)$ denotes the space of all polynomials on τ_i of degree $\leq N$. The well-balanced DG-approximation $\mathbf{u}_h : \Omega \times \mathbb{R}_+ \to \mathbb{R}^3$, $\mathbf{u}_h(\cdot, t) \in (W^h)^3$ is now constructed according to the approach of Xing et al. in [17], i.e. we solve

$$\begin{aligned}\frac{d}{dt}\int_{\tau_i} \mathbf{u}_h \cdot \mathbf{w}\, d\mathbf{x} \;=\; & \int_{\tau_i} \mathbf{F}(\mathbf{u}_h) \cdot \nabla \mathbf{w}\, d\mathbf{x} - \int_{\partial\tau_i} \mathbf{F}^{WB}(\mathbf{u}_{i,*}^-, \mathbf{u}_{i,*}^+, \varphi_i^-, \mathbf{n}) \cdot \mathbf{w}\, d\sigma \\ & + \int_{\tau_i} \mathbf{s}_h(\mathbf{u}_h, \mathbf{x}) \cdot \mathbf{w}\, d\mathbf{x},\end{aligned} \tag{3}$$

for any $\tau_i \in T^h$, $\mathbf{w} \in (W^h)^3$, where $\mathbf{F}^{WB}$ is the well-balanced correction of a suitable numerical flux function to be specified below and $\mathbf{u}_i^-$, $\mathbf{u}_i^+$ denote the approximate solution within τ_i and an adjacent element, respectively. Furthermore, the source term is discretized by $\mathbf{s}_h(\mathbf{u}_h, \mathbf{x}) = -g \cdot (0,\ \varphi_h \cdot \partial_{x_1} b_h,\ \varphi_h \cdot \partial_{x_2} b_h)^T$, where b_h is the projection of the bottom b to W^h. Given a numerical flux $\mathbf{F}^{num}$, which is the HLL flux [8] in our computations, the well-balanced flux $\mathbf{F}^{WB}$ is

$$\mathbf{F}^{WB}(\mathbf{u}_{i,*}^-, \mathbf{u}_{i,*}^+, \varphi_i^-, \mathbf{n}) = \mathbf{F}^{num}(\mathbf{u}_{i,*}^-, \mathbf{u}_{i,*}^+, \mathbf{n}) + \frac{1}{2}\left((\varphi_i^-)^2 - (\varphi_{i,*}^-)^2\right)\begin{pmatrix}0\\ \mathbf{n}\end{pmatrix},$$

where the modified (starred) left and right states are obtained from a hydrostatic reconstruction according to [1], i.e. we set $\mathbf{u}_{i,*}^{\pm} = (\varphi_{i,*}^{\pm}, \varphi_{i,*}^{\pm} \cdot (v_1)_i^{\pm}, \varphi_{i,*}^{\pm} \cdot (v_2)_i^{\pm})^T$ and $\varphi_{i,*}^{\pm} = \max\left\{0, \varphi_i^{\pm} + g\left(b_i^{\pm} - \max\left\{b_i^-, b_i^+\right\}\right)\right\}$.

Instead of limiters, modal filtering as described in [12] is used for shock capturing.

3 Separation of Production and Destruction Terms for Unconditional Positivity

The semidiscrete DG scheme (3) can compactly be written as the system of ODEs $\frac{d\mathbf{U}(t)}{dt} = \mathscr{L}_h(\mathbf{U}(t), t)$, where the vector $\mathbf{U}$ collects the complete set of DOFs of the spatial discretization. Neglecting boundary terms, in the DG scheme for cell means of water height, $\bar{H}_i(t) = \frac{1}{|\tau_i|}\int_{\tau_i} H_h(\mathbf{x}, t)\, d\mathbf{x}$, we now distinguish between positive and negative flux contributions over element boundaries. Collecting the indices of neighbor elements in the set $N(\tau_i)$, we have

$$\frac{d}{dt}(|\tau_i|\bar{H}_i) = -\sum_{j\in N(\tau_i)}\int_{\Gamma_{ij}} F_1^{WB}(\mathbf{u}_*^-, \mathbf{u}_*^+, \mathbf{n})\, d\sigma = \sum_{j\in N(\tau_i)} p_{ij} - \sum_{j\in N(\tau_i)} d_{ij},$$

where the properties $p_{ij} - d_{ij} = -\int_{\Gamma_{ij}} F_1^{WB}(\mathbf{u}_*^-, \mathbf{u}_*^+, \mathbf{n})\, d\sigma$ and $p_{ij} = d_{ji}$ are guaranteed by choosing the production term $p_{ij} = \max\left\{0, -\int_{\Gamma_{ij}} F_1^{WB}(\mathbf{u}_*^-, \mathbf{u}_*^+, \mathbf{n})\, d\sigma\right\}$ and the destruction term $d_{ij} = \max\left\{0, \int_{\Gamma_{ij}} F_1^{WB}(\mathbf{u}_*^-, \mathbf{u}_*^+, \mathbf{n})\, d\sigma\right\}$.

3.1 The MPSDIRK3 Scheme

In the similar context of production-destruction equations, Burchard et al. [3] developed the so-called modified Patankar-Euler (MPE) scheme based on the non-conservative Patankar scheme [14], which is of first order in time. For the cell means of water height in the DG SWE code it has the form

$$|\tau_i|\bar{H}_i^{n+1} = |\tau_i|\bar{H}_i^n + \Delta t\left(\sum_{i=1}^{I} p_{ij}^n \frac{\bar{H}_j^{n+1}}{\bar{H}_j^n} - \sum_{i=1}^{I} d_{ij}^n \frac{\bar{H}_i^{n+1}}{\bar{H}_i^n}\right).$$

In [3], a second order modified Patankar scheme was constructed as well. These schemes are positivity preserving and conservative for any time step size, see [3].

A *modified Patankar SDIRK3* (MPSDIRK3) scheme for the DG SWE code has been constructed in [13] based on Cash's third-order SDIRK method [4] given by

$$\begin{array}{c|ccc} \gamma & \gamma & & \\ \gamma+\delta & \delta & \gamma & \\ 1 & \alpha & \beta & \gamma \\ \hline & \alpha & \beta & \gamma \end{array} \quad \text{with} \quad \begin{aligned} \alpha &= 1.2084966491760101, \\ \beta &= -0.6443631706844691, \\ \gamma &= 0.4358665215084580, \\ \delta &= 0.2820667392457705. \end{aligned}$$

In the MPSDIRK3 scheme, only the last stage is modified as the implicit Euler scheme is unconditionally positive, see [9], and we have $\delta < \gamma$. In combination with the PP limiter by Xing et al. [17], the first two stages hence yield non-negative water height. Only in the last stage, we need to modify the vector

$$\mathbf{S} = \mathbf{U}^n + \alpha \Delta t \mathscr{L}_h \left(\mathbf{U}^{(1)}\right) + \beta \Delta t \mathscr{L}_h \left(\mathbf{U}^{(2)}\right)$$

by a vector containing non-negative cell means of water-height as described in [13].

3.2 The IRSDIRK3 Scheme

In [2], Bollermann et al. suggest to choose a local time step for each triangle edge in order to prevent negative water height. This approach still preserves conservativity. However, this technique does not allow the movement of the wet-dry front over more than one cell within one time step. Hence, also for implicit methods, the time step is still restricted by the positivity requirement. We will therefore generalize the idea in [2] to an iterative redistribution of the water height. Thus, we obtain a novel procedure which is unconditionally positive and conservative as the technique in [2] but also allows for larger time steps in the implicit case. To illustrate the idea, we consider again the intermediate quantities $\bar{H}_i^{[0]} = \bar{H}_i^n$ and $\bar{H}_i^{[1]} = \bar{H}_i^{[0]} + \frac{1}{|\tau_i|} \cdot \sum_{j \in N(\tau_i)} \Delta t_{ij}^{[0]} \left(p_{ij}^{[0]} - d_{ij}^{[0]}\right)$, where the initial production and destruction terms are $p_{ij}^{[0]} = \max\left\{0, -\alpha \int_{\Gamma_{ij}} F_1^{WB}(\mathbf{U}^{(1)}, \mathbf{n})\, d\sigma - \beta \int_{\Gamma_{ij}} F_1^{WB}(\mathbf{U}^{(2)}, \mathbf{n})\, d\sigma\right\} = d_{ji}^{[0]}$. Defining $\Delta t_i^{[0]} = |\tau_i| \bar{H}_i^{[0]} / \left(\sum_{j \in N(\tau_i)} d_{ij}^{[0]}\right)$, we set $\Delta t_{ij}^{[0]} = \min\left\{\Delta t, \Delta t_i^{[0]}\right\}$ if $d_{ij}^{[0]} > 0$ and else $\Delta t_{ij}^{[0]} = \min\left\{\Delta t, \Delta t_j^{[0]}\right\}$. A similar kind of modified cell means $\bar{H}_i^{[1]}$ is also computed within the scheme of Bollermann et al. Corresponding to that work, we would then set $\bar{H}_i^{n+1} = \bar{H}_i^{[1]}$. The basic idea of the wet-dry treatment of Bollermann et al. is hence to reduce the time step locally only for edges that contribute to the outflow of cells with possibly negative cell mean of water height. These are the cells that violate the condition $0 \le \bar{H}_i^n - \frac{\Delta t}{|\tau_i|} \cdot \sum_{j \in N(\tau_i)} d_{ij}^{[0]}$. However, cells that violate this do not necessarily contain negative cell means of water height in the next

time step. In fact, for an implicit scheme with a large time step, only a small amount of water may have moved from one cell to the next due to only a small amount of water contained in the donator cell at the old time level. Now, after calculating $\bar{H}_i^{[1]}$, water may have come in from another cell, so the local time step can be chosen larger. Hence, in the proposed *iterative redistribution SDIRK3* (IRSDIRK3) scheme we iteratively compute the quantities $\bar{H}_i^{[l+1]} = \bar{H}_i^{[l]} + \frac{1}{|\tau_i|} \cdot \sum_{j \in N(\tau_i)} \Delta t_{ij}^{[l]} \left(p_{ij}^{[l]} - d_{ij}^{[l]} \right)$ for $l = 1, 2, 3, \ldots$. Here, the destruction and production iterates are given by $d_{ij}^{[l]} = \left(1 - \Delta t_{ij}^{[l-1]} / \Delta t\right) d_{ij}^{[l-1]} = p_{ji}^{[l]}$ and the local time steps are $\Delta t_{ij}^{[l]} = \min \left\{ \Delta t, \Delta t_i^{[l]} \right\}$ if $d_{ij}^{[l]} > 0$ and else $\Delta t_{ij}^{[l]} = \min \left\{ \Delta t, \Delta t_j^{[l]} \right\}$, with $\Delta t_i^{[l]} = |\tau_i| \bar{H}_i^{[l]} / \left(\sum_{j \in N(\tau_i)} d_{ij}^{[l]} \right)$. This algorithm is stopped when $\sum_i (\bar{H}_i^{[l]} - \bar{H}_i^{[l-1]})^2 < tol$ for a given tolerance tol which we set to $tol = \Delta t \cdot 10^{-12}$.

4 Numerical Experiments for Shallow Water Flows

Numerical results for MPSDIRK3 and IRSDIRK3 are presented for a polynomial degree of $N = 2$. The nonlinear systems of equations arising due to the implicit time integration were solved using a Jacobian-free Newton-GMRES scheme. Figure 1 shows the DG solution for the oscillating-lake test proposed in [7] on a computational grid consisting of $K = 23138$ elements. In Tables 1 and 2, we compare the CPU times obtained by the TVD-RK3 scheme to those of the MPSDIRK3 and IRSDIRK3 scheme, respectively, on increasingly stiff computational grids. The stiffness of the grids, measured by $S = \max_i |\tau_i| / \min_i |\tau_i|$, is increased via local refinement. According to the results, the implicit schemes beat the explicit one by a factor up to 3.8. Tables 1 and 2 furthermore list the mass conservation errors and L^2 errors in water height committed by the implicit scheme. Full conservation can obviously only be achieved if the accuracy within the iterative solver is set to zero, which is neglected due to practical reasons as usual. However, the results show that for the specific tolerances chosen in this study, the corresponding conservation error can be neglected. In this computation, the IRSDIRK3 scheme shows slightly better results, both in terms of CPU time and in terms of conservation error. In addition, the IR technique is easier to adapt to IMEX time integration. Preliminary computations have been carried out for the first order IMEX scheme

$$\mathbf{U}^{n+1} = \mathbf{U}^n + \Delta t \left(\mathscr{L}_h^E(\mathbf{k}^{(1)}, t^n) + \mathscr{L}_h^I(\mathbf{k}^{(1)}, t^{n+1}) \right), \quad \mathbf{k}^{(1)} = \mathbf{U}^n + \Delta t \mathscr{L}_h^I(\mathbf{k}^{(1)}, t^{n+1})$$

and $N = 0$, i.e. a first-order finite volume scheme. The IMEX splitting is obtained by defining implicit cells as small ones with $|\tau_i| \leq 0.35 \cdot (\max_i |\tau_i| + \min_i |\tau_i|)$. Now, all fluxes between implicit cells are collected in $\mathscr{L}_h^I$ and the remaining terms in $\mathscr{L}_h^E$. In Table 3, we compare the results of the IR-IMEX scheme to those obtained by implicit Euler time integration. Here, the IMEX scheme beats the explicit one

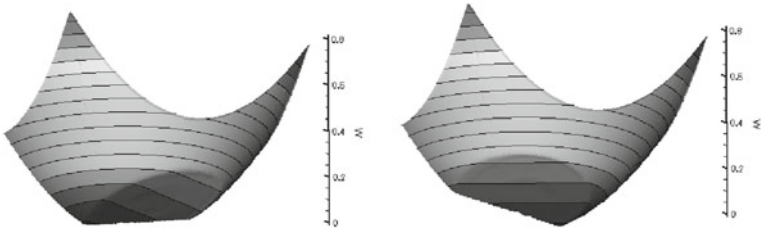


Fig. 1 Water surface $w = H + b$ at time $T = 0.75$ (*left*) and $T = 1.5$ (*right*)

Table 1 CPU time comparison and conservation error for implicit MPSDIRK3 scheme

Stiffness S	Avg. Δt_{EX}	Avg. Δt_{IM}	$\frac{\text{CPU}_{EX}}{\text{CPU}_{IM}}$	err_{cons}	L^2 error in H
6.5	2.99e-4	1.07e-2	0.65	2.31e-14	2.371e-03
25.9	1.51e-4	5.42e-3	0.82	1.11e-14	2.419e-03
103.4	7.55e-5	2.71e-3	1.29	8.88e-15	2.438e-03
413.7	3.77e-5	1.36e-3	1.52	6.93e-14	2.132e-03
1654.6	1.89e-5	6.79e-4	1.34	2.25e-13	1.902e-03
105894.6	2.40e-6	8.57e-5	3.51	5.42e-13	7.195e-03

Table 2 CPU time comparison and conservation error for implicit IRSDIRK3 scheme

Stiffness S	Avg. Δt_{EX}	Avg. Δt_{IM}	$\frac{\text{CPU}_{EX}}{\text{CPU}_{IM}}$	err_{cons}	L^2 error in H
6.5	2.99e-4	1.07e-2	0.65	4.89e-15	2.373e-03
25.9	1.51e-4	5.42e-3	0.83	3.02e-14	2.418e-03
103.4	7.55e-5	2.71e-3	1.36	3.38e-14	2.439e-03
413.7	3.77e-5	1.36e-3	1.69	1.51e-14	2.112e-03
1654.6	1.89e-5	6.78e-4	1.63	1.55e-14	1.887e-03
105894.6	2.40e-6	8.57e-5	3.81	8.88e-16	7.454e-03

Table 3 CPU time comparison: Explicit Euler versus implicit IR-IMEX scheme

Stiffness S	Avg. Δt_{EX}	Avg. Δt_{IM}	$\frac{\text{CPU}_{EX}}{\text{CPU}_{IM}}$	err_{cons}	L^2 error in H
413.7	1.14e-4	9.43e-4	1.43	1.78e-15	4.003e-03
1654.6	5.69e-5	9.43e-4	1.84	7.55e-15	4.004e-03
105894.6	7.25e-6	9.43e-4	8.28	4.44e-16	4.005e-03

by a factor of 8.28. So far, no preconditioner was used within the Newton-GMRES solver. An additional speed-up using preconditioning strategies will be the aim of future work.

Acknowledgments The author S. Ortleb gratefully acknowledges financial support by the Deutsche Forschungsgemeinschaft (DFG) through grant ME 1889/3-1.

References

1. Audusse, E., Bouchut, F., Bristeau, M.-O., Klein, R., Perthame, B.: A fast and stable well-balanced scheme with hydrostatic reconstruction for shallow water flows. SIAM J. Sci. Comput. **25**, 2050–2065 (2004)
2. Bollermann, A., Noelle, S., Lukáčová-Medvidová, M.: Finite volume evolution galerkin methods for the shallow water equations with dry beds. Commun. Comput. Phys. **10**, 371–404 (2011)
3. Burchard, H., Deleersnijder, E., Meister, A.: A higher-order conservative patankar-type discretization for stiff systems of production-destruction equations. Appl. Numer. Math. **47**, 1–30 (2003)
4. Cash, J.R.: Diagonally implicit runge-kutta formulae with error estimates. J. Inst. Maths Applics **24**, 293–301 (1979)
5. Casulli, V.: A high-resolution wetting and drying algorithm for free-surface hydrodynamics. Int. J. Numer. Meth. Fluids **60**, 391–408 (2009)
6. Cockburn, B., Shu, C.-W.: Runge-kutta discontinuous galerkin methods for convection-dominated problems. J. Sci. Comp. **16**, 173–261 (2001)
7. Gallardo, J.M., Parés, C., Castro, M.: On a well-balanced high-order finite volume scheme for shallow water equations with topography and dry areas. J. Comput. Phys. **227**, 574–601 (2007)
8. Harten, A., Lax, P.D., van Leer, B.: On upstream differencing and godunov-type schemes for hyperbolic conservation laws. SIAM Rev. **25**, 35–61 (1983)
9. Higueras, I.: Representations of runge-kutta methods and strong stability preserving methods. SIAM J. Numer. Anal. **43**, 924–948 (2005)
10. Karniadakis, G.E., Sherwin, S.: "Spectral/hp Element Methods for Computational Fluid Dynamics", 2^{nd} edn. Oxford University Press, New York (2005)
11. Kärnä, T., de Brye, B., Gourgue, O., Lambrechts, J., Comblen, R., Legat, V., Deleersnijder, E.: A fully implicit wetting-drying method for dg-fem shallow water models, with an application to the scheldt estuary. Comp. Meth. Appl. Mech. Eng. **200**, 509–524 (2011)
12. Meister, A., Ortleb, S.: The DG scheme on triangular grids with adaptive modal and variational filtering routines applied to shallow water flows. In: Ansorge, R., Bijl, H., Meister, A., Sonar, T. (eds.) Recent Developments in the Numerics of Nonlinear Hyperbolic Conservation Laws, NNFM 120, pp. 253–266. Springer, New York (2013)
13. Ortleb, S., Meister, A.: On unconditionally positivity preserving dg schemes for shallow water flows with shock capturing by adaptive filtering procedures. PAMM **13**, 505–506 (2013)
14. Patankar, S.V.: Numerical Heat Transfer and Fluid Flow. McGraw-Hill, New York (1980)
15. Ricchiuto, M., Bollermann, A.: Stabilized residual distribution for shallow water simulations. J. Comput. Phys. **228**, 1071–1115 (2009)
16. Shu, C.-W., Osher, S.: Efficient implementation of essentially non-oscillatory shock capturing schemes ii. J. Comput. Phys. **83**, 32–78 (1989)
17. Xing, Y., Zhang, X., Shu, C.-W.: Positivity-preserving high order well-balanced discontinuous galerkin methods for the shallow water equations. Adv. Water Resour. **33**, 1476–1493 (2010)

Convergence of a Nonlinear Scheme for Anisotropic Diffusion Equations

Christophe Le Potier

Abstract We study a nonlinear correction depending on a parameter ν to eliminate oscillations appearing in the discretization of diffusion operators. For $\nu = 1$, it satisfies the LMP structure (see definition 1.1 in [6]). For $\nu < 1$, with a few non restrictive assumptions on the mesh, we prove the convergence of this scheme. Using an analytical solution, we show the robustness and the accuracy of this algorithm in comparison with results obtained by linear schemes which do not satisfy the minimum principle on this test.

1 Statement of the Problem

Let Ω be an open bounded connected polygonal subset of $\mathbb{R}^d$ (d = 2 or d = 3). We consider the following elliptic problem:

$$\begin{cases} -\operatorname{div}(D\nabla\bar{u}) = f & \text{in } \Omega, \\ \bar{u} = 0 & \text{on } \partial\Omega; \end{cases} \tag{1}$$

with:

- $f \in L^2(\Omega)$, the source term;
- $\bar{u}$ the concentration ;
- D, the permeability, a symmetric tensor-valued function such that (a) D is piecewise Lipschitz-continuous on Ω and (b) the set of the eigenvalues is included in $[\lambda_{min}, \lambda_{max}]$ with $\lambda_{min} > 0$ for all $x \in \Omega$.

It is well known that classical linear methods discretizing diffusion operators do not always satisfy a maximum principle for distorted meshes or high anisotropy ratios

C. Le Potier (✉)
Commissariat à l'énergie atomique, CEA-Saclay DEN, DM2S, STMF, LMEC,
91191 Gif-Sur-Yvette Cedex, France
e-mail: clepotier@cea.fr

J. Fuhrmann et al. (eds.), *Finite Volumes for Complex Applications VII - Methods and Theoretical Aspects*, Springer Proceedings in Mathematics & Statistics 77,
DOI: 10.1007/978-3-319-05684-5_43,

[8]. In [11], we proposed a general approach to correct a cell-centered scheme. For example, it can be applied to schemes developed in [1, 2, 7, 9, 14]. This method has been analyzed in [4] in the one-dimensional case for the heat equation. In more general cases, proofs of convergence are shown for monotone corrections in [3]. However, the required assumptions can be difficult to verify since they depend on the numerically computed solution. That is the reason why, we propose a proof of convergence for a nonlinear correction with only few assumptions on the mesh.

2 Basics for Numerical Schemes

We take the notations and the assumptions on the discretization of Ω given in ([3], Sect. 2). We just recall that $\mathcal{M}$ is a family of non-empty open polygonal connected disjoint subsets of Ω (the *control volumes*) such that $\overline{\Omega} = \cup_{K\in\mathcal{M}} \overline{K}$. To study the convergence of the schemes, we will use the following quantities: the size of the mesh

$$\mathrm{size}(\mathcal{D}) = \sup_{K\in\mathcal{M}} \mathrm{diam}(K)$$

and the regularity of the mesh

$$\mathrm{regul}(\mathcal{D}) = \sup_{\substack{K\in\mathcal{M}\\ \sigma\in\mathcal{E}_K}} \left\{ \frac{\mathrm{diam}(K)}{d_{K,\sigma}} \right\} + \sup_{\substack{K,L\in\mathcal{M}\\ \sigma\in\mathcal{E}_K\cap\mathcal{E}_L}} \left\{ \frac{d_{L,\sigma}}{d_{K,\sigma}} \right\}.$$

We recall that we consider cell-centered schemes consisting in finding :

$$\forall K \in \mathcal{M}, \quad \mathcal{S}_K(u) = |K|\, f_K, \tag{2}$$

where f_K denotes the mean value of f on the cell K, and $(\mathcal{S}_K(u), K \in \mathcal{M})$ is a system of $\mathrm{Card}(\mathcal{M})$ equations on some unknowns $(u_K)_{K\in\mathcal{M}}$. We will also use the following quantity (slightly changed compared to ([3], Sect. 3.3) because we only consider symmetric schemes) :

$$\mathrm{reg}(\mathcal{D}) = \mathrm{regul}(\mathcal{D}) + \max_{K\in\mathcal{M}, L\in V(K)} \frac{\mathrm{diam}(L)}{\mathrm{diam}(K)} + \max_{K\in\mathcal{M}} \mathrm{Card}\, V(K).$$

3 A Correction Depending on Parameters ν, ν_f, ε

We denote by $\mathcal{H}_\mathcal{M}$ the set of functions which are constant on each control volume of $\mathcal{M}$. We consider the discrete linear operator $\mathcal{A}^\mathcal{D}$ which can be written in the following form

$$\forall u \in \mathscr{H}_{\mathscr{M}}, \forall K \in \mathscr{M}, \quad \mathscr{A}_K(u) = \sum_{Z \in V(K)} \alpha_{K,Z}(u_Z - u_K) \tag{3}$$

where $\alpha_{K,Z}$ does not depend on u. A correction is a family $\beta^{\mathscr{D}} = (\beta_{K,Z})_{K \in \mathscr{M}, Z \in V(K)}$ of functions $\beta_{K,Z} : \mathscr{H}_{\mathscr{M}} \to \mathbb{R}$. For a given correction β:

- the corrected scheme $\mathscr{S}^{\mathscr{D}} = (\mathscr{S}_K, K \in \mathscr{M})$ is defined by

$$\forall u \in \mathscr{H}_{\mathscr{M}}, \forall K \in \mathscr{M}, \quad \mathscr{S}_K(u) = -\mathscr{A}_K(u) + \mathscr{R}_K(u) \tag{4}$$

- the corrective term is the function $\mathscr{R}^{\mathscr{D}} : \mathscr{H}_{\mathscr{M}} \to \mathscr{H}_{\mathscr{M}}$ defined by

$$\forall u \in \mathscr{H}_{\mathscr{M}}, \forall K \in \mathscr{M}, \quad \mathscr{R}_K(u) = \sum_{Z \in V(K)} \beta_{K,Z}(u)(u_K - u_Z). \tag{5}$$

We assume that the original scheme is symmetric, coercive and consistent in the sense $A2$ and $A3$ described in ([3], Sect. 3.1). We study a correction depending on positive parameters ν, ν_f and ε with $\nu \leq 1$, $\nu_f < 1$ and $\varepsilon > 0$. It satisfies $\forall K \in \mathscr{M}$, $Z \in V(K)$ $\beta_{K,Z} \geq 0$ and $\beta_{K,Z} = \beta_{Z,K}$. We also assume that for all $K \in \mathscr{M}$, $f_K \geq 0$ to be able to apply the proposition 1.4 in [6]. We deduce that the corrected scheme is still coercive as shown in [11]. For $\nu = 1$, it satisfies the LMP structure (see definition 1.1 in [6]). For $\nu < 1$, we use a few non restrictive assumptions on the mesh to prove the convergence of this scheme.

As detailed in ([13], proposition 5.1), it is possible to change the value of f_K such that for all $K \in \mathscr{M}$, f_K is different from zero and such that the modified scheme converges.

We slightly change the regularized correction in ([3], Sect. 4.2). We recall the definition of $\mathrm{Card}_\varepsilon V(K,u)^*$ for $u \in \mathscr{H}_{\mathscr{M}}$ and $K \in \mathscr{M}$:

$$\mathrm{Card}_\varepsilon V(K,u)^* = \sum_{Z \in V(K)} \frac{|u_K - u_Z|}{|u_K - u_Z| + \varepsilon}.$$

We also define the expressions $\mathrm{sgn}_\varepsilon(u_K - u_Z) = \dfrac{(u_K - u_Z)}{|u_K - u_Z| + \varepsilon}$ and

$\rho_{K,Z}(u) = \nu_f min\left(\dfrac{|K|\, f_K}{\mathrm{Card}\, V(K)}, \dfrac{|Z|\, f_Z}{\mathrm{Card}\, V(Z)}\right)$ with the convention $\rho_{K,Z}(u) = \rho_{K,K}(u)$ if $Z = \sigma \in \mathscr{E}_{\mathrm{ext}}$. The correction $\beta_\varepsilon^{\mathscr{D}}$ is defined, for all $u \in \mathscr{H}_{\mathscr{M}}$, all $K \in \mathscr{M}$ and all $Z \in V(K)$, by:

$$\begin{aligned} \beta^\varepsilon_{K,Z}(u) &= \nu \max\left(\frac{|\mathscr{A}_K(u)|}{\mathrm{Card}_\varepsilon V(K,u)^*}, \frac{|\mathscr{A}_Z(u)|}{\mathrm{Card}_\varepsilon V(Z,u)^*}\right) \frac{1}{|u_K - u_Z| + \varepsilon} \\ &\quad + \frac{\rho_{K,Z}(u)}{|u_K - u_Z| + \varepsilon} = \tilde{\beta}^\varepsilon_{K,L}(u) + \frac{\rho_{K,Z}(u)}{|u_K - u_Z| + \varepsilon} \end{aligned} \tag{6}$$

with the convention $\dfrac{|\mathscr{A}_Z(u)|}{\mathrm{Card}_\varepsilon V(Z,u)^*} = 0$ if $Z = \sigma \in \mathscr{E}_{\mathrm{ext}}$.

Comparing (6) to the regularized correction in ([3], Sect. 4.2), we remark that the terms $\rho_{K,Z}(u)$ have been changed. Indeed, to prove the proposition 4, we use that for all $K \in \mathscr{M}$,

$$|\sum_{Z\in V(K)} \rho_{K,Z}(u)\mathrm{sgn}_\varepsilon(u_K - u_Z)| \leq |K| f_K.$$

The corresponding corrected scheme $\mathscr{S}^\varepsilon$ can be written, for all $u \in \mathscr{H}_\mathscr{M}$ and all $K \in \mathscr{M}$,

$$\begin{aligned}\mathscr{S}^\varepsilon_K(u) = &- \mathscr{A}_K(u) \\ &+ \nu \sum_{Z\in V(K)} \max\Big(\frac{|\mathscr{A}_K(u)|}{\mathrm{Card}_\varepsilon V(K,u)^*}, \frac{|\mathscr{A}_Z(u)|}{\mathrm{Card}_\varepsilon V(Z,u)^*}\Big)\mathrm{sgn}_\varepsilon(u_K - u_Z) \\ &+ \sum_{Z\in V(K)} \rho_{K,Z}(u)\mathrm{sgn}_\varepsilon(u_K - u_Z). \end{aligned} \tag{7}$$

We recall the sets $V(K,u)^+$ and $V(K,u)^-$ defined by:

$$\begin{aligned} V(K,u)^+ &= \{Z \in V(K)\ ;\ \mathscr{A}_K(u)(u_Z - u_K) > 0\}, \\ V(K,u)^- &= \{Z \in V(K)\ ;\ \mathscr{A}_K(u)(u_Z - u_K) < 0\}. \end{aligned}$$

Proposition 1 *There exists one solution to the corrected scheme* (7).

Proof According to Proposition 6 in [3], there exists one solution because the corrective term $\mathscr{R}^\mathscr{D} : \mathscr{H}_\mathscr{M} \to \mathscr{H}_\mathscr{M}$ is continuous.

Proposition 2 *For* $\nu = 1$, *the corrected scheme* (7) *satisfies the LMP structure.*

Proof We refer to Sects. 4.2 and 3.2.4 in [3] for the proof of the LMP structure.

Proposition 3 *We assume* $\nu < 1$. *Let* ε_{min} *defined by* $\varepsilon_{min} = (1-\nu_f) min_{K\in\mathscr{M}} \frac{|K| f_K}{2\sum_{Z\in V(K)} |\alpha_{K,Z}|}$ *and* $0 < \varepsilon \leq \varepsilon_{min}$. *Let* u *be a solution to* $\mathscr{S}^\varepsilon = 0$ *and let* $K_0 \in \mathscr{M}$ *be such that*

$$\frac{|\mathscr{A}_{K_0}(u)|}{\mathrm{Card}_\varepsilon V(K_0,u)^*} = \max_{K\in\mathscr{M}} \frac{|\mathscr{A}_K(u)|}{\mathrm{Card}_\varepsilon V(K,u)^*}. \tag{8}$$

Then, there exists $Z \in V(K_0,u)^+$ *such that* $|u_{K_0} - u_Z| \geq \varepsilon$. *Moreover, for all* $K \in \mathscr{M}$

$$\frac{|\mathscr{A}_K(u)|}{\mathrm{Card}_\varepsilon V(K,u)^*} \leq 2|K_0| \frac{f_{K_0}(1+\nu_f)}{(1-\nu)}. \tag{9}$$

Proof The K_0 component of $\mathscr{S}^{\varepsilon}(u)$ reduces to

$$-\mathscr{A}_{K_0}(u) + \nu \sum_{Z \in V(K_0)} \frac{\left|\mathscr{A}_{K_0}(u)\right|}{\mathrm{Card}_{\varepsilon} V(K_0, u)^*} \mathrm{sgn}_{\varepsilon}(u_{K_0} - u_Z) + \sum_{Z \in V(K_0)} \rho_{K_0,Z}(u) \mathrm{sgn}_{\varepsilon}(u_{K_0} - u_Z) = |K_0|\, f_{K_0}. \tag{10}$$

The corrected scheme becomes:

$$-(1-\nu)\mathscr{A}_{K_0}(u) - \frac{(2\nu)\mathscr{A}_{K_0}(u)}{\mathrm{Card}_{\varepsilon} V(K_0, u)^*} \sum_{Z \in V(K_0,u)^+} \left|\mathrm{sgn}_{\varepsilon}(u_{K_0} - u_Z)\right| + \sum_{Z \in V(K_0)} \rho_{K_0,Z}(u) \mathrm{sgn}_{\varepsilon}(u_{K_0} - u_Z) = |K_0|\, f_{K_0} \tag{11}$$

As $|\sum_{Z \in V(K_0)} \rho_{K_0,Z}(u) \mathrm{sgn}_{\varepsilon}(u_{K_0} - u_Z)| \leq \nu_f |K_0| f_{K_0}$, we get that $-\mathscr{A}_{K_0}(u) \geq 0$. We obtain :

$$|\mathscr{A}_{K_0}(u)| \leq \frac{|K_0| f_{K_0}(1+\nu_f)}{(1-\nu)}.$$

On the other hand, we can deduce from equality (10) that there exists a constant C_1 ($0 \leq C_1 \leq 2$) such that $-\mathscr{A}_{K_0}(u) C_1 = (1-\nu_f)\,|K_0|\, f_{K_0}$. Since $-\mathscr{A}_{K_0}(u) \leq \sum_{Z \in V(K_0)} |\alpha_{K_0,Z}||u_{K_0} - u_Z|$, we deduce that there exists $Z \in V(K_0)$ such that $|u_{K_0} - u_Z| \geq \dfrac{(1-\nu_f)\,|K_0|\, f_{K_0}}{C_1 \sum_{Z \in V(K_0)} |\alpha_{K_0,Z}|}$.

Using the assumption on ε, we get that $\mathrm{Card}_{\varepsilon} V(K_0, u)^* \geq \frac{1}{2}$. We finally obtain

$$\frac{\left|\mathscr{A}_{K_0}(u)\right|}{\mathrm{Card}_{\varepsilon} V(K_0, u)^*} \leq 2 \frac{|K_0|\, f_{K_0}(1+\nu_f)}{(1-\nu)}. \tag{12}$$

We conclude using the definition of K_0.

We prove now the convergence of the scheme.

Proposition 4 *Assume $f \in L^d(\Omega)$ and $\nu < 1$. Let $(\mathscr{D}^n)_{n \geq 1}$ be a sequence of admissible meshes of Ω (in the sense given in [3]) such that* $\mathrm{size}(\mathscr{D}^n) \to 0$ *as $n \to \infty$ and* $(\mathrm{reg}(\mathscr{D}^n))_{n \geq 1}$ *is bounded; assume that there exists $C_1 > 0$ satisfying*

$$\forall n \geq 1, \forall K, L \in \mathscr{M}^n, \quad |K| \leq C_1 |L|, \tag{13}$$

$$\forall n \geq 1, \forall K \in \mathscr{M}^n, \quad \mathrm{diam}(K)^d \leq C_1 |K|. \tag{14}$$

Let $(\varepsilon_n)_{n\geq 1}$ be a sequence of positive real numbers satisfying $0 < \varepsilon_n \leq \varepsilon_{min}$ and let $(u^n)_{n\geq 1}$ be a sequence of discrete functions satisfying $u^n \in \mathscr{H}_{\mathscr{M}^n}$ and the equation $\mathscr{S}^{\varepsilon_n} = 0$.

Then, as $n \to \infty$, u^n converges in $L^2(\Omega)$ to the unique solution of (1).

Proof We remark that $\mathrm{Card}_\varepsilon V(K,u)^* \leq \mathrm{Card}\ V(K)$. It means that $\mathrm{Card}_\varepsilon V(K,u)^*$ is bounded as $\mathrm{size}(\mathscr{D}^n) \to 0$. Applying Proposition 3 and the proof given in proposition 10 in [3], as $\mathrm{size}(\mathscr{D})$ tends to 0, we obtain

$$\sum_{K\in\mathscr{M}} \mathrm{diam}(K) \sum_{Z\in V(K)} \tilde{\beta}^{\varepsilon}_{K,Z}(u)\,|u_K - u_Z| \to 0.$$

For a given $\varphi \in \mathscr{C}^\infty_c(\Omega)$, we set $\varphi_{\mathscr{D}} = (\varphi_K)_{K\in\mathscr{M}} \in \mathscr{H}_{\mathscr{M}}$ with $\varphi_K = \varphi(x_k)$. Applying proposition 7 in [3], we deduce, since $\mathrm{size}(\mathscr{D})$ tends to 0,

$$\Big|\sum_{K\in\mathscr{M}} \varphi_K \sum_{Z\in V(K)} \tilde{\beta}^{\varepsilon}_{K,Z}(u)(u_K - u_Z)\Big| \to 0. \tag{15}$$

Then we consider the term $T_K = \sum_{Z\in V(K)} \rho_{K,Z}(u_K - u_Z)$. Multiplying the term T_K by φ_K and summing over $K \in \mathscr{M}$ we get:

$$\Big|\sum_{K\in\mathscr{M}} T_K\varphi_K\Big| \leq \sum_{K\in\mathscr{M}} \sum_{Z\in V(K)} \rho_{K,Z} \frac{|u_K - u_Z|}{|u_K - u_Z| + \varepsilon} |\varphi_K - \varphi_Z|\,.$$

Thanks to the regularity of φ, there exists C_2 not depending on $\mathscr{D}$ such that

$$|\varphi_K - \varphi_Z| \leq C_2\,\mathrm{size}(\mathscr{D})$$

for all $K \in \mathscr{M}$. We deduce

$$\Big|\sum_{K\in\mathscr{M}} T_K\varphi_K\Big| \leq \nu_f C_2 \sum_{K\in\mathscr{M}} |K| f_K\,\mathrm{size}(\mathscr{D}) \leq C_2\nu_f \Big|\int_\Omega f(x)dx\Big|\,\mathrm{size}(\mathscr{D}) \tag{16}$$

Therefore $|\sum_{K\in\mathscr{M}} T_K\varphi_K| \to 0$ as $\mathrm{size}(\mathscr{D}) \to 0$. We deduce

$$\Big|\sum_{K\in\mathscr{M}} \varphi_K \sum_{Z\in V(K)} \beta^{\varepsilon}_{K,Z}(u)(u_K - u_Z)\Big| \to 0$$

as $\mathrm{size}(\mathscr{D}) \to 0$. Using the proof of proposition 7 in [3], we obtain the desired result.

4 Numerical Results

We start from the conservative and consistent original operator $\mathscr{A}^{\mathscr{D}}$ developed in [9]. Moreover, we use highly perturbed grids of triangular cells given in the test 1 of the benchmark [8]. To deal with the nonlinear terms, we perform a Picard algorithm. Let us denote u^i the value of the solution where i is a fixed point iteration. We fix $u = u^i$ in $\beta_{K,Z}(u)$ in (4) and the iterative scheme can be written :

$$\forall K \in \mathscr{M}, \quad -\mathscr{A}_K(u^{i+1}) + \sum_{Z \in V(K)} \beta_{K,Z}(u^i)(u_K^{i+1} - u_Z^{i+1}) = |K| f_K.$$

Moreover, we use a BICGSTAB (Biconjuguate gradient stabilized) algorithm to solve the previous linear system.

Some notations used to present the numerical results are given in Table 1. In order to numerically evaluate the convergence of the scheme, let us consider the following elliptic problem:

$$\begin{cases} -\operatorname{div}(D\nabla\bar{u}) = f \text{ in } \Omega =]0, 0.5[\times]0, 0.5[\\ \bar{u}(x, y) = \sin(\pi x)\sin(\pi y) \text{ for } (x, y) \in \partial\Omega \end{cases} \tag{17}$$

with

$$D = \frac{1}{x^2 + y^2} \begin{pmatrix} y^2 + \alpha x^2 & -(1-\alpha)xy \\ -(1-\alpha)xy & x^2 + \alpha y^2 \end{pmatrix}$$

and

$$\begin{cases} u_{\text{ana}} = \sin(\pi x)\sin(\pi y), \\ f = -\operatorname{div} D\nabla u_{\text{ana}}. \end{cases} \tag{18}$$

The parameter α is equal to 10^{-3} and the anisotropy ratio is equal to 10^3. As $\Omega =]0, 0.5[\times]0, 0.5[$, we check that $f \geq 0$.

We show the results obtained in Table 2 with the scheme developed in [9] (denoted S. ori), and with the nonlinear corrections denoted (S.1, $\nu = 1$, $\nu_f = 0$) and (S. 2, $\nu = 0.99$, $\nu_f = 0$). Normally, we should have used the value of ε_{min} defined in Proposition 4, but we observe that the number of iterations nit is high. For each grid, we take $\varepsilon = 50(max_{K \in \mathscr{M}}|K|)$ and we check that $\text{Card}_\varepsilon V(K_0, u)^* \geq \frac{1}{2}$. It is clear that the original scheme is second order in space but we observe large oscillations. Let us remark that one also obtains these large oscillations [12] with the schemes DDFV [5] and SUSHI [7]. Concerning schemes S.1 and S.2, we observe the expected results (positive solution for scheme S.1, and convergence for scheme S.2). We also remark some unexpected results (convergence for scheme S.1, and positive solution for scheme S.2). So, we will have to understand why the scheme S.1 is convergent. However, the results presented in this paper seem to us interesting from a software development point of view. We know that for $\nu < 1$, the corrected scheme must

Table 1 Notations

h	Length of the edges on the boundary of the square $]0, 0.5[\times]0, 0.5[$
L^2 error	L^2 error of the computed solution with respect to the analytical solution
ratiol2	Order of convergence, in L^2 norm, of the method
nit	Number of iterations needed in the Picard method to compute the approximate solution of $\mathscr{S}^\varepsilon$
Min. Val.	$\min\{u_K\,;\ K \in \mathscr{M}\}$
Max. Val.	$\max\{u_K\,;\ K \in \mathscr{M}\}$

Table 2 Numerical results for (17) with the original scheme and the nonlinear schemes S.1, S.2 as a function of the discretization step

h	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{64}$	$\frac{1}{128}$
L^2 error (S. ori)	1.19×10^{0}	3.62×10^{-1}	8.84×10^{-2}	2.20×10^{-2}	5.41×10^{-3}
ratiol2 (S. ori)		1.71	2.03	2.01	2.02
Undershoots (S. ori)	16 %	9 %	4 %	2 %	0.6 %
Min. Val. (S. ori)	-1.51×10^{0}	-1.16×10^{0}	-3.2×10^{-1}	-8.06×10^{-2}	-2.02×10^{-2}
L^2 error (S. 1)	1.15×10^{-1}	5.77×10^{-2}	2.45×10^{-2}	1.29×10^{-2}	6.58×10^{-3}
ratiol2 (S. 1)		1.00	1.23	0.92	0.97
nit	15	20	24	19	17
$\mathrm{Card}_\varepsilon V(K_0, u)^*$	0.70	0.74	0.74	0.74	0.74
ε	4.92×10^{-1}	1.23×10^{-1}	3.07×10^{-2}	7.70×10^{-3}	1.92×10^{-3}
L^2 error (S. 2)	1.17×10^{-1}	5.83×10^{-2}	2.45×10^{-2}	1.28×10^{-2}	6.51×10^{-3}
ratiol2 (S. 2)		0.99	1.25	0.93	0.98
nit	15	19	23	19	18
$\mathrm{Card}_\varepsilon V(K_0, u)^*$	0.71	0.74	0.74	0.74	0.74
ε	4.92×10^{-1}	1.23×10^{-1}	3.07×10^{-2}	7.70×10^{-3}	1.92×10^{-3}

converge in the numerical results. Concerning the CPU time, the nonlinear schemes are of course slower than the linear scheme. If it is too long for a given application, one can use the linear algorithm developed in [10].

References

1. Aavatsmark, I., Barkve, T., Boe, O., Mannseth, T.: Discretization on unstructured grids for inhomogeneous, anisotropic media. Part I: derivation of the methods. SIAM J. Sci. Comput. **19**(5), 1700–1716 (1998)
2. Agelas, L., Masson, R.: Convergence of the finite volume MPFA O scheme for heterogeneous anisotropic diffusion problems on general meshes. C. R. Acad. Sci. Paris Ser. I **346**(17–18), 1007–1012 (2008)
3. Cancès, C., Cathala, M., Le Potier, C.: Monotone corrections for generic cell-centered Finite Volume approximations of anisotropic diffusion equations. Numer. Math. **125**, 387–417 (2013)

4. Després, B.: Non linear finite volume schemes for the heat equation in 1D, HAL: hal-00714781, to appear in M2AN
5. Domelevo, K., Omnes, P.: A finite volume method for the Laplace equation on almost arbitrary two-dimensional grids, M2AN **39**(6), 1203–1249 (2005)
6. Droniou, J., Le Potier, C.: Construction and convergence study of schemes preserving the elliptic local maximum principle. SIAM J. Numer. Anal. **49**(2), 459–490 (2011)
7. Eymard, R., Gallouët, T., Herbin, R.: A cell-centred finite-volume approximation for anisotropic diffusion operators on unstructured meshes in any space dimension. IMA J. Numer. Anal. **26**(2), 326–353 (2006)
8. Herbin, R., Hubert, F.: Benchmark on discretization schemes for anisotropic diffusion problems on general grids. In: Eymard, R., Hérard, J.-M. (eds.) 5th International Symposium on Finite Volumes for Complex Applications, pp. 659–692. ISTE/Wiley, London/Hoboken (2008)
9. Le Potier, C.: Schéma volumes finis pour des opérateurs de diffusion fortement anisotropes sur des maillages non structurés, C. R. Acad. Sci. Paris Ser. I **340**(12), 921–926 (2005)
10. Le Potier, C.: Un schéma linéaire vérifiant le principe du maximum pour des opérateurs de diffusion très anisotropes sur des maillages déformés, C. R. Acad. Sci. Paris Ser. I **347**, 105–110 (2009)
11. Le Potier, C.: Correction non linéaire et principe du maximum pour la discrétisation d'opérateurs de diffusion avec des schémas volumes finis centrés sur les mailles. C. R. Acad. Sci. Paris **348**(11–12), 691–695 (2010)
12. Le Potier, C., Mahamane, A., Omnes, P.: Internal report
13. Le Potier, C.: Theoretical convergence of monotone and positive schemes for anisotropic diffusion equation, in preparation
14. Lipnikov, K., Shashkov, M., Yotov, I.: Local flux mimetic finite difference methods. Numer. Math. **112**(1), 115–152 (2009)

A Hydrodynamic Model for Dispersive Waves Generated by Bottom Motion

S. R. Pudjaprasetya and S. S. Tjandra

Abstract A numerical scheme based on the staggered finite volume method is presented at the aim of simulating surface waves generated by a bottom motion. Here, we address the 2D Euler equations in which the vertical domain is resolved only by one layer. Under the assumption of horizontally dominant flow, we enhance the conservative scheme for shallow water equations to include bottom motion and to account take into the hydrodynamic pressure term. The resulting scheme can simulate free surface wave generated by downward motion of a bed-section. The result demonstrates the evolution of a negative wave displacement followed by a dispersive wave train. Our numerical results show good agreement with results from the KdV model and experiment by Hammack [3].

1 Introduction

Motivated by the origin of tsunami, this paper investigates surface wave generation by bottom motion. The study of bottom motion generating surface wave has long been an interesting subject of researches, see for instance [1, 2, 4, 5, 9]. Hammack experiment in [3] is used as one of the benchmark test for tsunami generation codes. In the experiment, part of a bottom wave tank was shifted downwards, as result, surface wave is generated which is then propagate to the right. The generated wave produces a long wave of depression followed by a series of short-waves. In the far

S. R. Pudjaprasetya (✉)
Industrial and Financial Mathematics Research Group, Bandung Institute of Technology, Jalan Ganesha 10, Bandung 40132, Indonesia
e-mail: sr_pudjap@math.itb.ac.id

S. S. Tjandra
Industrial Engineering, Parahyangan Catholic University, Jalan Ciumbuleuit 94, Bandung 40141, Indonesia
e-mail: ssudharmatjandra@yahoo.com

J. Fuhrmann et al. (eds.), *Finite Volumes for Complex Applications VII - Methods and Theoretical Aspects*, Springer Proceedings in Mathematics & Statistics 77,
DOI: 10.1007/978-3-319-05684-5_44,

field downstream region, the effects of nonlinearities and frequency dispersion are of the same order. Thus, numerical models for this simulation should combine those two effects. Fuhrman and Madsen in [2] use Boussinesq model for simulating this experiment. Kervella et. al. in [7] study linear and nonlinear 3D models of tsunami generation.

Here, we enhance the staggered conservative scheme for the nonlinear SWE described previously in Stelling and Duinmeijer [10], to incorporate dispersion effect by solving the Euler equations. Here, we implement one layer approximation for the vertical axis. In this article, numerical scheme for the nonlinear SWE is called hydrostatic model, whereas the scheme for solving Euler equations is called hydrodynamic model.

2 Mathematical Model

Consider the Euler equations for the flow of incompressible and inviscid fluid with constant density

$$u_x + w_z = 0 \tag{1}$$

$$u_t + uu_x + wu_z = -g\eta_x - P_x \tag{2}$$

$$w_t + uw_x + ww_z = -P_z \tag{3}$$

with $(u\ \ w)^T$ is the fluid particle velocity, $P(x, z, t)$ the hydrodynamic pressure term. Let $\eta(x, t)$ denotes the surface elevation measured from the undisturbed water level. And to calculate the bottom motion, we let the bottom topography to depend also on time t, and we denote it as $-d(x, t)$. For horizontally dominant flow, continuity equation appears as a dynamic equation in terms of η and d, which will be formulated below. Integrating (1) with respect to z from $z = -d(x, t)$ to $z = \eta(x, t)$ yields

$$\int_{-d(x,t)}^{\eta(x,t)} u_x \, dz + w]_{-d(x,t)}^{\eta(x,t)} = 0.$$

Kinematic boundary conditions along the free surface $z = \eta(x, t)$ and along the impermeable bottom $z = -d(x, t)$ are $w = \eta_t + u\eta_x$ and $w = -d_t - ud_x$, respectively. Substituting those two conditions, and neglecting the non-linear term yields

$$(\eta + d)_t + u(\eta + d)_x + \int_{-d(x,t)}^{\eta(x,t)} u_x \, dz = 0.$$

Under shallow water assumption, in which horizontal velocity u is independent of z, the integral term can be approximated by $u_x(\eta + d)$, and the continuity equation reads

$$h_t + (hu)_x = 0, \tag{4}$$

where $h = \eta + d$ denoting the water thickness. Recapitulating, the governing equations for hydrodynamic model that will be used in further discussion are (1–4). Without hydrodynamic pressure and for horizontally dominant flow, the equations can be reduced to

$$h_t + (hu)_x = 0, \tag{5}$$
$$u_t + uu_x + g\eta_x = 0, \tag{6}$$

which is the shallow water equations (SWE). In [8] we discuss the staggered finite volume scheme to solve the nonlinear SWE (5, 6). The conservative properties of this staggered scheme and its accuracy and robustness for simulation of rapidly varied flows are discussed in Stelling and Duinmeijer [10], see also [6]. This scheme is then modified in [11] to solve Euler equations with a small number of vertical grid points. As a result, the scheme is able to simulate nonlinear wave phenomena with dispersion. The key issue of this paper is implementing the conservative scheme for the nonlinear shallow water equation with dispersion for simulation of surface wave generated by bottom motion.

3 Hydrostatic Model

In this section we first discuss numerical scheme for the hydrostatic model (5, 6). Consider a computational domain $[0, L]$ with a staggered grid and partition points $x_{1/2} = 0, x_1, \ldots, x_{i-1/2}, x_i, x_{i+1/2}, \ldots, x_{Nx+1/2} = L$. Continuity equation (5) is approximated at cell $[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$ and momentum equation (6) is approximated at cell $[x_i, x_{i+1}]$. The approximate equations are

$$\frac{dh_i^n}{dt} + \frac{{}^*h_{i+\frac{1}{2}}^n u_{i+\frac{1}{2}}^n - {}^*h_{i-\frac{1}{2}}^n u_{i-\frac{1}{2}}^n}{\Delta x} = 0, \tag{7}$$
$$\frac{du_{i+1/2}^n}{dt} + g\frac{\eta_{i+1}^{n+1} - \eta_i^{n+1}}{\Delta x} + (uu_x)_{i+1/2}^n = 0. \tag{8}$$

In this approximation h is calculated at every full grid points x_i, whereas u at every half grid points $x_{i+\frac{1}{2}}$, see Fig. 1. Since $\eta = h - d$, hence η is also calculated at every full grid points x_i. In (7), terms h are written with * because it needs approximation, and we implement the upwind approximation

$${}^*h_{i+\frac{1}{2}} = \begin{cases} h_i & \text{if } u_{i+1/2} \geq 0 \\ h_{i+1} & \text{if } u_{i+1/2} < 0. \end{cases} \tag{9}$$

Hence, the term ${}^*h_{i+\frac{1}{2}} u_{i+\frac{1}{2}}$ expresses the first order approximation of mass flux at edge $x_{i+\frac{1}{2}}$ for $i = 0, 1, 2, \ldots, Nx$. When the flow is going to the right $u_{i+1/2} \geq 0$,

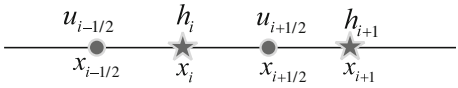


Fig. 1 Staggered grid with configuration of calculated variables h and u

we take the left flux $h_i u_{i+1/2}$, and when the flow is going to the left $u_{i+1/2} < 0$, we take the right flux $h_{i+1} u_{i+1/2}$, and hence mass conservation is always retained in the approximation (7) for any direction of the flow. We note here that bottom motion can be accommodated automatically in this scheme.

For the advection term $(uu_x)_{i+1/2}$ we implement the momentum conservative approximation as introduced by Stelling and Duijnmeijer in [10]. Since $uu_x = \frac{1}{h}\left(\frac{\partial(qu)}{\partial x} - u\frac{\partial q}{\partial x}\right)$ with $q = hu$ the horizontal momentum, a consistent approximation for the advection term is

$$(uu_x)_{i+1/2} = \frac{1}{\bar{h}_{i+\frac{1}{2}}}\left(\frac{\bar{q}_{i+1}\,{}^*u_{i+1} - \bar{q}_i\,{}^*u_i}{\Delta x} - u_{i+\frac{1}{2}}\frac{\bar{q}_{i+1} - \bar{q}_i}{\Delta x}\right), \tag{10}$$

$$\bar{h}_{i+1/2} = \frac{1}{2}(h_i + h_{i+1}), \quad \bar{q}_i = \frac{1}{2}(q_{i+\frac{1}{2}} + q_{i-\frac{1}{2}}), \quad q_{i+\frac{1}{2}} = {}^*h_{i+\frac{1}{2}} u_{i+\frac{1}{2}},$$

with an upwind approximation for ${}^*u_i$

$${}^*u_i = \begin{cases} u_{i-\frac{1}{2}}, & \text{if } \bar{q}_i \geq 0 \\ u_{i+\frac{1}{2}}, & \text{if } \bar{q}_i < 0 \end{cases} \tag{11}$$

Recapitulating, the hydrostatic scheme for the nonlinear SWE are (7, 9) for continuity equation and (8, 10, 11) for the momentum balance. The scheme is of second order accurate for the linear parts, but it is of order one for the non-linear parts, see [8] for details.

4 Hydrodynamic Model

In hydrodynamic formulation, variation in the vertical z-axis is considered, and hence we consider the full set of equations (1–4). The first approximation uses only one layer to resolve the vertical interval, and configuration will be described below. Along the free surface, hydrodynamic pressure is set to zero, and it is increasing with depth. Here, we assume P to be linearly depends on z, and let $P(x_i, z = -d(x_i, t^n), t^n) \equiv P_i^n$. Next, along the impermeable flat bottom holds $w(x, z = -d_0, t) = 0$, and we assume further w to be linearly depends on z. Let $w(x_i, z = \eta(x_i, t^n), t^n) \equiv w_i^n$. Hence, in this hydrodynamic scheme we only need one vector array for dynamic pressure P_i^n and one vector array for vertical velocity w_i^n, which is very efficient. In

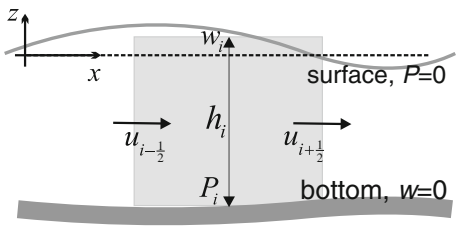


Fig. 2 Configuration of the calculated variables in the staggered grid of hydrodynamic model

the discrete hydrostatic model as explained previously, h are computed at full grid points x_i, whereas u are computed at half grid points $x_{i+1/2}$. In this hydrodynamic model, variables w and P are calculated at full grid points x_i, see Fig. 2.

A way to incorporate the hydrodynamic pressure term is described below. Suppose at any time step, we have calculated η^{n+1}, $\bar{u}$ and $\bar{w}$ from the hydrostatic model, in which $\bar{u}, \bar{w}$ are written in bars since they need corrections. Incorporating the hydrodynamic term, their values are corrected as follows

$$u_{i+\frac{1}{2}}^{n+1} = \bar{u}_{i+\frac{1}{2}} - \Delta t \frac{P_{i+1}^{n+1} - P_i^n}{2\Delta x}, \tag{12}$$

$$w_i^{n+1} = \bar{w}_i + \Delta t \frac{2P_i^{n+1}}{h_i^{n+1}}. \tag{13}$$

But values of P_i should be calculated first. And this can be obtained from one layer approximation of the continuity equation, read as

$$\frac{w_i^{n+1} - 0}{h_i^{n+1}} + \frac{u_{i+\frac{1}{2}}^{n+1} - u_{i-\frac{1}{2}}^{n+1}}{\Delta x} = 0. \tag{14}$$

Substituting (12) and (13) into (14) yields

$$\Delta x \left(w_i^* + \Delta t \frac{2P_i^{n+1}}{h_i^{n+1}} \right) + h_i^{n+1} \left(u_{i+\frac{1}{2}}^* - u_{i-\frac{1}{2}}^* + \Delta t \frac{-P_{i-1}^{n+1} + 2P_i^{n+1} - P_{i-1}^{n+1}}{2\Delta x} \right) = 0 \tag{15}$$

which is a tridiagonal system of equations for P_i^{n+1}.

Finally, the computational procedure when stepping from t_n to t_{n+1} is as follows

1. From the hydrostatic model (7, 8, 10), we calculate η_i^{n+1}, $\bar{u}_{i+\frac{1}{2}}$, and $\bar{w}_i$.
2. Solve the tridiagonal system (15) to calculate P_i^{n+1}.
3. Make correction for $u_{i+\frac{1}{2}}^{n+1}$ using (12).
4. Make correction for w_i^{n+1} using (13).

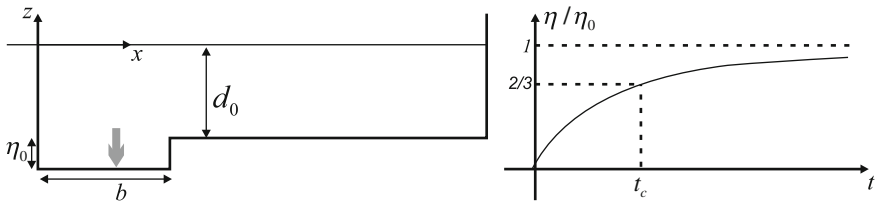


Fig. 3 Bed deformation model: (*left*) spatial deformation (*right*) time deformation

4.1 Hammack Experiment (1973)

The experiment was conducted in a closed wave tank with length 31.6 m, width 39.4 cm. Water depth is $d_0 = 5$ cm. A section of the bed, length b, at the left edge of the water tank was shifted η_0 downwards. This bed motion generates surface elevation in the form of wave depression. This wave is then propagate to the right. Here, we mimic the above experimental set up in our numerical simulation.

The motion of bottom is modelled by

$$d(x,t) = d_0 - \eta_0(1 - e^{-\alpha t})\mathscr{H}(b - x), \tag{16}$$

where $\mathscr{H}$ is the Heaviside function. In [3] this motion is called exponential bottom motion. It depends on parameters η_0 and b which are the amplitude and length of the moving part of the bed, see Fig. 3. Another parameter is the characteristic time t_c which is defined such that $\eta/\eta_0 = \frac{2}{3}$ and parameter α relates with the characteristic time t_c as $\alpha = 1.11/t_c$.

For simulation we take still water level as the initial condition, and a downward bed disturbance according to (16) with $b = 61$ cm, $\eta_0 = -0.5$ cm and $t_c = 0.093b/\sqrt{gd_0}$. Since our scheme can calculate bottom motion, here we perform computations considered as active generation. We take $\Delta x = b/12$ and $\Delta t = 0.001$ s. The choice of Δx is such that $x = b$ is located at a half grid point where we have u value. By doing this, we keep the mass conserved. At the right and left boundaries fully reflecting walls are prescribed, however the simulations are stopped before any reflections occur. As result from a downward bottom motion, water surface moves to a maximum displacement $-\eta_0$. After reaching its maximum, the surface returns to the still water level but it produces an oscillating tail (Fig. 4).

Figure 5 illustrates the downstream behavior of waves resulting from a downward bed displacement. Time series of the waves are recorded at four locations $(x-b)/d_0 = 0, 20, 180, 400$ and the results are plotted w.r.t $t\sqrt{g/d_0} - (x-b)/d_0$. The results of our hydrodynamic scheme are plotted together with Hammack results from two approaches, i.e. KdV model and experimental data. In Fig. 5 (top left) there is no result from the KdV model because it uses passive generation. We observe that the results of our hydrodynamic scheme show a good agreement with results from KdV model as calculated by Hammack [3]. When compared with Hammack

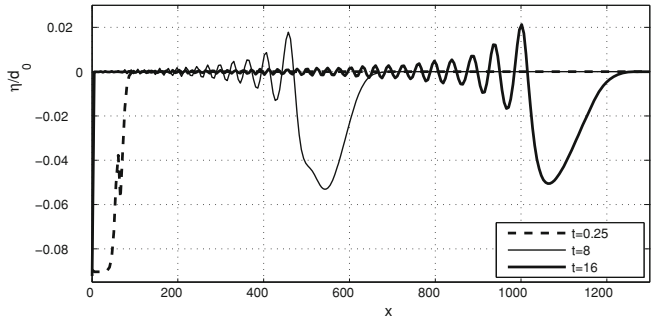


Fig. 4 Snapshots of surface elevation at subsequent time $t = 0.25$, $t = 8$, $t = 16$

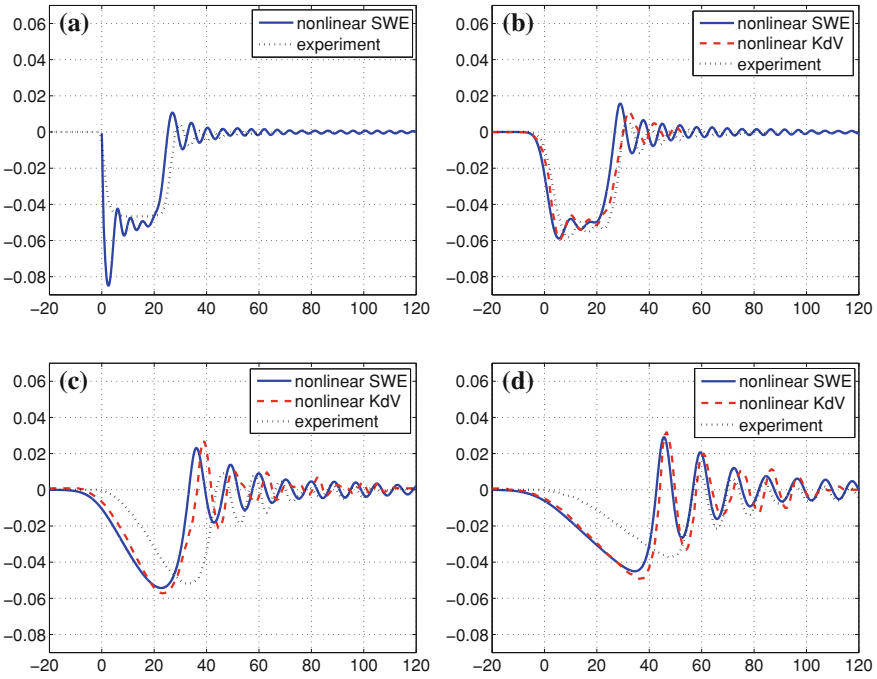


Fig. 5 Time series of surface elevation η/d_0 w.r.t $t\sqrt{g/d_0} - (x-b)/d_0$ resulting from a downward bottom motion. The waves are recorded at locations **a** $(x-b)/d_0 = 0$, **b** 20, **c** 180, **d** 400

experiment [3], the experimental waves have the same shape as predicted by the numerical models. However, in the far field the experimental waves noticeably faster than the numerical models. Further study and comparisons using various bottom motions are still under research, and will be reported in a separate paper.

5 Conclusions

We have presented the non-hydrostatic numerical scheme to calculate surface wave dynamics by solving the 2D Euler equations for horizontally dominant flow. The scheme was used for simulating surface wave generation due to a downward bottom motion. By resolving the vertical axis as just one layer, our hydrodynamic scheme can produce a negative wave displacement followed with a dispersive wave train. Our results are in a good agreement with results from KdV model, also comparable with experimental data. Both are taken from Hammack [3]. Considering these good agreements, we expect our hydrodynamic model is suitable for simulating wave generation by bottom motion. Moreover, it is expected that the proposed method can be computationally competitive with dispersive models like KdV or Boussinesq.

Acknowledgments Financial support from Riset Desentralisasi ITB 2014 and DIKTI scholarship are greatly acknowledged.

References

1. Dutykh, D., Dias, F.: Energy of tsunami waves generated by bottom motion. Proc. R. Soc. A **465**, 725–744 (2009)
2. Fuhrman, D.R., Madsen, P.A.: Tsunami generation, propagation, and run up with high order Boussinesq model. Coast. Eng. **56**, 747–758 (2009)
3. Hammack, J.L.: A note on tsunamis: their generation and propagation in an ocean of uniform depth. J. Fluid Mech. **4**, 769–799 (1973)
4. Hammack, J.L., Segur, H.: The Korteweg-de Vries equation and water waves, part 2. Comparison with experiment. J. Fluid Mech. **65**, 289–314 (1974)
5. Hammack, J.L., Segur, H.: The Korteweg-de Vries equation and water waves, part 3. Oscillatory waves. J. Fluid Mech. **84**, 337–358 (1978)
6. Jochen, K.: Ocean Modelling for Beginners. Springer, London (2009).
7. Kervella, Y., Dutykh, D., Dias, F.: Comparison between three-dimensional linear and nonlinear tsunami generation models. Theor. Comput. Fluid Dyn. **21**, 245–269 (2007)
8. Pudjaprasetya, S.R., Magdalena, I.: Momentum conservative scheme for shallow water flows. East Asian J. Appl. Math. (EAJAM). **4**(2), 152–165 (2014)
9. Saito, T., Furumura, T.: Three-dimensional tsunami generation simulation due to sea-bottom deformation and its interpretation based on the linear theory. Geophys. J. Int. **178**, 877–888 (2009)
10. Stelling, G.S., Duinmeijer, S.P.A.: A staggered conservative scheme for every Froude number in rapidly varied shallow water flows. Int. J. Numer. Methods Fluids **43**, 1329–1354 (2003)
11. Stelling, G.S., Zijlema, M.: An accurate and efficient finite difference algorithm for non hydrostatic free surface flow with application to wave propagation. Int. J. Numer. Methods Fluids **43**, 1–23 (2003)

A Conservative Coupling of Level-Set, Volume-of-Fluid and Other Conserved Quantities

Matthias Waidmann, Stephan Gerber, Michael Oevermann and Rupert Klein

Abstract A conservative level-set volume-of-fluid synchronization strategy including coupling to other conserved quantities such as mass or momentum is presented. The scheme avoids mass loss/gain of fluidic structures in zero Mach number two-phase flow while keeping the interface between the two fluid phases sharp. Local level-set correction and a consistent discretization error control using information from the energy equation based divergence constraint allow for application of the presented method to both constant and variable density zero Mach number two-phase flow with or without interfacial mass transport.

1 Introduction

Capturing methods representing an interface implicitly via a scalar field are very popular for simulation of fluidic interfaces since topological changes and severe interface movement can be handled much easier than in interface tracking methods, where the computational grid has to adjust according to the changing interface. The most common capturing methods are the level-set and the volume-of-fluid method. While stand-alone versions of both methods suffer from drawbacks concerning maintenance of physical properties at fluidic interfaces, hybrid approaches are able to combine the advantages and overcome the drawbacks of each method as shown below.

In **Level-Set Methods** [10] a sharp continuous moving interface is given implicitly as approximation to the iso-surface—usually the zero level—of the space ($\mathbf{x}$) and

M. Waidmann (✉) · S. Gerber · R. Klein
Institut für Mathematik, Freie Universität Berlin, Arnimallee 6, 14195 Berlin, Germany
e-mail: waidmann@math.fu-berlin.de

M. Oevermann
Division of Combustion, Department of Applied Mechanics, Chalmers University of Technology, Gothenburg, Schweden

J. Fuhrmann et al. (eds.), *Finite Volumes for Complex Applications VII - Methods and Theoretical Aspects,* Springer Proceedings in Mathematics & Statistics 77,
DOI: 10.1007/978-3-319-05684-5_45,

time (t) dependent scalar level-set function $G(\mathbf{x}, t)$. The interface Γ separates the domain Ω into two parts occupied by fluidic phases (+) and (−). The change of the interface via the level-set function (which is smooth around the interface) is governed by

$$\frac{\mathrm{D}G}{\mathrm{D}t} \equiv G_t + \mathbf{v} \cdot \nabla G = 0, \qquad G(\mathbf{x}, t) \begin{cases} > 0 & \forall\ \mathbf{x} \in \Omega^{(+)}(t) \\ = 0 & \forall\ \mathbf{x} \in \Gamma(t) \\ < 0 & \forall\ \mathbf{x} \in \Omega^{(-)}(t) \end{cases} \tag{1}$$

with velocity field $\mathbf{v}(\mathbf{x}, t)$. The advantages of obtaining an accurate continuous but simple interface representation capable of handling topological changes by default face the drawback of a missing mechanism to maintain mass conservation resulting in significant local non-physical mass transition across the interface.

Volume-of-fluid methods [7], in contrast, offer mass conservation properties. The common conservative form describing the change of the discontinuous phase indicator function $\varphi(\mathbf{x}, t)$ is given by

$$(\rho\varphi)_t + \nabla \cdot (\rho\varphi\mathbf{v}) = 0, \qquad \varphi(\mathbf{x}, t) = \begin{cases} \varphi^{(+)} := 1 & \forall\ \mathbf{x} \in \Omega^{(+)}(t) \\ \varphi^{(-)} := 0 & \forall\ \mathbf{x} \in \Omega^{(-)}(t) \end{cases} \tag{2}$$

whereat $\rho(\mathbf{x}, t)$ is the fluid density. The integral average $\overline{\varphi}$ of the phase indicator over a control volume ω is the volume fraction of the reference fluid phase (+) in ω with $0 \leq \overline{\varphi} \leq 1$. The major drawback of volume-of-fluid methods is the extensive interface reconstruction procedure due to ambiguity of the interface position within the control volume if no surrounding information is used in addition.

Meanwhile many **Hybrid Level-Set Volume-of-Fluid Methods** have been developed (e.g. [14] or recently [9]), aiming to overcome the drawbacks of both stand-alone methods combining the respective advantages based on the two different available interface representations. Many of them, however, still suffer from conservation issues. The presented conservative method in principle follows the corresponding strategy in [12] and [11] of avoiding volume-of-fluid related interface reconstruction while correcting the level-set based interface representation supported by the conservatively transported volume-of-fluid indicator distribution (as introduced first in [1]). The method presented in [12] and [11] is modified to be consistently extendable to variable density flows with or without interfacial mass transport and to different asymptotic limits of the underlying equations, e.g. with atmospheric density and/or pressure stratification (see e.g. [8]). The resulting procedure, including the necessary two-way coupling between level-set and volume-of-fluid indicator (as well as all other conserved quantities), is part of the framework of a generalized Cartesian grid finite volume projection method ([2, 8]) for zero Mach number variable density two-phase flow.

Here the strategy for coupling level-set, volume-of-fluid and other conserved quantities is focused on. While presented on a Cartesian grid, it is expected to be applicable to other than Cartesian grids as well, assuming the geometric features to be sufficiently resolved by the respective grid. The detailed presentation of the used surrounding Cartesian grid flow solver framework is deferred to a future publication.

2 Governing Equations and Numerical Solution Strategy

In short, the flow solver framework consists of both a predictor and a corrector step, the latter including two conservative projections and the conservative coupling of level-set, volume-of-fluid and conserved quantities. It solves the zero Mach number variable density equations for immiscible viscous two-phase flow which are based on the corresponding equations for inviscid single phase flow as given in [8] for length scales which are small compared to the atmospheric pressure scale height. For explanation of the conservative corrector step which couples level-set, volume-of-fluid and other conserved quantities it is sufficient to focus on the auxiliary system

$$(\rho\phi)_t + \nabla\cdot\left(P\mathbf{v}\,\frac{\rho\phi}{P}\right) = 0 \tag{3}$$

$$(P\varphi)_t + \nabla\cdot(P\mathbf{v}\,\varphi) = 0 \tag{4}$$

$$P_t + \nabla\cdot(P\mathbf{v}) = 0 \tag{5}$$

of the predictor step hyperbolic part in conservative formulation.

Equation (3) represents advection of any scalar ϕ, Eq. (4) is another conservative form of the additional phase indicator conservation law (2) and Eq. (5) provides information on the predictor step divergence error stuck to the predicted values of each of the conserved quantities. This error is collected in the time derivative P_t of the spatially homogenous entropy related auxiliary variable P for computation of correction fluxes during the corrector step. For consistent discretization error generation and correction the fluxes of all quantities are referred to fluxes of P instead of the mass flux as visible in Eqs. (3)–(5), enabling the presented scheme to handle e.g. variable density flows as well. The auxiliary system is solved in space-time integral form using a method of lines with an explicit second order accurate finite volume discretization on a Cartesian grid and a second order accurate strong stability preserving Runge-Kutta time integrator (SSP-RK2) from [5]. In addition, the level set Eq. (1) is integrated using a spatially unlimited third order upstream central finite difference scheme (see e.g. [6]) in a narrow band $\mathcal{N}$ around the interface. A two stage SSP-RK2 serves as time integrator as well.

For grid cells with homogenous phase indicator φ, constant throughout the entire time interval $\Delta T := \left[t^n, t^{n+1}\right]$ considered, upwind grid cell face flux approximation and time integration can be done second order accurate as in standard single phase finite volume methods with the fluid properties of the corresponding fluid phase including either $\overline{\varphi}(t) = 1$ or $\overline{\varphi}(t) = 0$ with $t \in \Delta T$. For the subset of grid cells with $0 < \overline{\varphi}(t) < 1$ the same strategy is applied, however, accurate flux computation and underlying data reconstruction via a ghost fluid approach [4] in a near interface sub-set of $\mathcal{N}$ is more complex in grid cells with faces intersected by the interface.

The numerical flux $\mathcal{F} = F\Delta A$ over such a cut Cartesian grid cell face $\partial\omega$ of size ΔA is determined based on the weighted sum

$$F = b\,F^{(+)} + (1-b)\,F^{(-)} \tag{6}$$

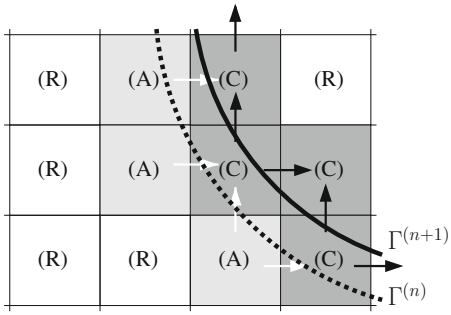


Fig. 1 Cartesian grid cell types: (A) abandoned, (C) cut, (R) regular, and possible correction fluxes ($\longrightarrow$; *white* (A), *black* (C)) at time level $n+1$ in the vicinity of a moving interface Γ

of the numerical approximation to the grid cell face normal space-time average flux densities $F^{(+)}$ and $F^{(-)}$ for any conserved quantity. $F^{(+)}$ and $F^{(-)}$ are determined separately within the corresponding fluid phases as proposed in [13] and the weight b is the time average Cartesian grid cell face fraction in reference fluid phase $(+)$.

Discretization errors in both level-set advection and zero level approximation on the one hand and phase indicator transport, e.g. due to truncation errors in the approximation of the space-time weights b in (6) on the other hand lead to non-physical mass transport between the fluid phases, to physically not necessarily reasonable integral average values for the conserved quantities in the vicinity of the interface and to diverging of the two different interface representations over time.

Distinguishing between (R) "regular", (A) "abandoned" and (C) "cut" grid cells at time level $n+1$ as shown in Fig. 1, the approach presented in the next section aims for overcoming these issues by correction of the conserved quantities based on restoring the violated boundedness of the volume fraction $\overline{\varphi}$ in cells (A) and (C) and synchronizing the transport of the two interface representations (1) and (4) while keeping the method extendable from constant to zero Mach number variable density flow, whereat mass and phase indicator Eq. (4) are not redundant anymore.

3 Conservative Level-Set Volume-of-Fluids Synchronization

The correction procedure consists of **(1)** separate volume fraction based adjustment of conserved quantities in ***(a)*** abandoned and ***(b)*** cut grid cells at the new time level $n+1$ based on (divergence error free) predicted data, relying on the topology as given via the spatially higher order accurate level-set function, followed by **(2)** local level-set correction, based on corrected volume fraction values $\overline{\varphi}$.

(1) Volume Fraction Based Adjustment of Conserved Quantities: Since Eqs. (3)–(5) are treated consistently with the same numerical scheme, discretization errors becoming evident in the volume fraction of the reference phase are inherent in all other conserved quantities with jump at the interface as well. Correction fluxes for such a conserved quantity $\rho\phi$ can be derived once the necessary correction $\Delta\overline{\varphi}$

for the volume fraction is determined. The correction flux for the integral average $\overline{\rho\phi}$ yields

$$\mathscr{F}_{\rho\phi,j} = \frac{1}{[\![\varphi]\!]} \left[\left(\frac{\phi^{(+)}}{\theta^{(+)}} \right)_j \left[\mathscr{F}_{ref}^{(+)} \right]_j + \left(\frac{\phi^{(-)}}{\theta^{(-)}} \right)_j \left[\mathscr{F}_{ref}^{(-)} \right]_j \right] \tag{7}$$

on grid cell face j with $\theta^{(\pm)} := \frac{P}{\rho^{(\pm)}}$, $[\![\varphi]\!] = \varphi^{(+)} - \varphi^{(-)} \equiv 1$ and

$$\left[\mathscr{F}_{ref}^{(+)} \right]_j = w_j \Delta\overline{P\varphi} = \mathscr{F}_{P\varphi,j}, \qquad \left[\mathscr{F}_{ref}^{(-)} \right]_j = -\left[\mathscr{F}_{ref}^{(+)} \right]_j \tag{8}$$

Since P is spatially homogeneous with $[\![P]\!] = P^{(+)} - P^{(-)} \equiv 0$ as well as $\Delta\overline{P} = 0$ concerning the volume fraction based correction procedure, correction fluxes for P vanish. This has already been considered in (8), and, thus, $\Delta\overline{P\varphi} = \Delta\overline{P}\overline{\varphi} = P\Delta\overline{\varphi}$. The quantities $\frac{\phi}{\theta}$ on grid cell face j, advected by the reference correction fluxes in Eq. (7), are determined using upwind values with respect to the direction of the corresponding reference flux $\mathscr{F}_{ref}$. The latter remains to be determined:

***(a) Abandoned Grid Cells* (A):** In grid cells which are left by the interface during ΔT with only fluid phase $(\pm)$ remaining in the abandoned grid cell at the end of ΔT, the target phase indicator based volume fraction $\overline{\varphi} = \varphi^{(\pm)}$ is known immediately allowing for a *local* correction approach due to $\alpha^{(\pm)} = 1$. Here, α is the volume fraction of the corresponding fluid phase $(\pm)$ based on the level-set zero level approximation. The weights w_j for reference flux determination in Eq. (8) need to satisfy

$$\sum_j w_j = 1 \qquad w_j := \frac{W_j}{\sum_i W_i} \tag{9}$$

while only grid cell faces cut or run over by the interface during ΔT are allowed to have non-zero weights. This avoids influence on other abandoned or regular neighboring grid cells as only cut grid cells are possible exchange partners (see Fig. 1). Due to the CFL stability condition [3] for the explicit predictor part, an abandoned grid cell has at least one cut neighbor cell. Besides of (9) the weights w_j are arbitrary. Non-physical over- or undershoots in neighboring cut grid cells could only be avoided in general if exactly the error causing weighting was used. This, however, is not possible since fluxes across grid cell faces, which do not have a cut neighbor cell at the end of ΔT anymore, might have contributed to the present errors. Yet, to keep possible over- and undershoots in cut grid cells (which are corrected in the subsequent step) as small and the resulting error distribution in cut grid cells as smooth as possible, W_j is chosen to be

$$W_j := \alpha_j^{(\pm)} \tag{10}$$

with $\alpha_j^{(\pm)}$ as the level-set based volume fraction of the neighboring cut grid cell sharing grid cell face j . Investigations of different weightings W_j are presented in a future publication.

***(b) Cut Grid Cells* (C):** Now both errors in the distribution of the phase indicator φ and related errors in the conserved quantity $\rho\phi$ are focused to the cut grid cells only, while $\Delta\overline{P\varphi} = P\Delta\overline{\varphi} = 0$ in all un-cut grid cells. The sum of all wrongly distributed volume fractions $\overline{\varphi}^\star$ in cut cells c, however, equals the sum of the corrected yet unknown distribution, $\sum_c \overline{\varphi}_c^\star = \sum_c \overline{\varphi}_c$. However, in cut grid cells the target value $\overline{\varphi}$ is not as trivially available as for the abandoned cells due to unknown target values for the level-set based volume fraction $\alpha^{(\pm)}$ as the latter has not yet reached its final value in cut cells as well. Further, the distribution weight w for flux computation (8) cannot be determined locally uniquely since values in possible neighboring exchange partner cells, which are cut cells as well, are—in contrast to the correction of abandoned grid cells—also subject to be corrected (see Fig. 1). This leads to a *non-local spatial coupling* limited to cut cells only. The resulting Poisson-type problem for obtaining correction fluxes $\mathscr{F}_{P\varphi}$, similar to the one solved in [12] and [11] for both cut and abandoned grid cells (called "mixed cells" there) at once, yields

$$\sum_j \mathscr{F}_{P\varphi,j} = \Delta\overline{P\varphi} = P\left(\overline{\varphi}^\star - \overline{\varphi}\right), \qquad \mathscr{F}_{P\varphi,j} := \frac{\Delta t}{\Delta V}\Delta A_j \left(\nabla\psi\cdot\mathbf{n}\right)_j \quad (11)$$

with unknown scalar ψ for each cut grid cell. In contrast to $\overline{\varphi}^\star$, the level-set based volume fraction α will always stay within the physically reasonable value range $0 < \alpha < 1$. The total relative volume deviation $\left(\sum_c \overline{\varphi}_c - \sum_c \alpha_c\right) = \left(\sum_c \overline{\varphi}_c^\star - \sum_c \alpha_c\right) \equiv \sum_c \left(\overline{\varphi}_c^\star - \alpha_c\right) = \sum_c \Delta\overline{\varphi}_c$ between the two different interface representations with local difference $\Delta\overline{\varphi} := \left(\overline{\varphi}^\star - \alpha\right)$ needs conservative redistribution among the cut cells in order to guarantee $0 < \overline{\varphi} < 1$. This is achieved via Eq. (11), turning into

$$\frac{\Delta t}{\Delta V}\sum_j \Delta A_j \left(\nabla\psi\cdot\mathbf{n}\right)_j = P\left(\Delta\overline{\varphi} - \chi\sum_c \Delta\overline{\varphi}_c\right) \quad (12)$$

after introduction of the target volume fraction $\overline{\varphi} := \alpha + \chi\sum_c \Delta\overline{\varphi}_c$. With $\sum_c \chi_c = 1$ and $\chi := \frac{\vartheta}{\sum_c \vartheta_c}$ the choice of $\vartheta := \alpha\,(1-\alpha) \geq 0$ controls the conservative redistribution process such that cut grid cells remain cut and generation of both new abandoned and new cut cells is avoided. Correction fluxes between pairs of grid cells different from cut/cut are set to zero to decouple the cut grid cells from the surrounding. After solving the resulting pseudo-Neumann Poisson problem (12) in the entire domain for ψ enforcing a trivial zero solution in all un-cut grid cells,

correction fluxes (11) are available for Eq. (8) and its application to Eq. (7). The solvability condition for (12) is satisfied by default as the sum of its right hand side is zero due to only non-zero contributions in cut grid cells.

(2) Local Correction of Level-Set Function: After redistribution of conserved quantities including volume fractions $\overline{\varphi}$ in cut grid cells, the level-set function G is corrected *locally*. A local interface segment normal level-set correction velocity is determined as in [12], based on the discrepancy between the volume fractions α and the adjusted $\overline{\varphi}$. This correction velocity is non-zero only in cut grid cells. Application of the correction velocity to the level-set transport algorithm pushes the interface towards matching volume fractions α and $\overline{\varphi}$ by slight adjustment of the gradient of G in the vicinity of Γ due to change of values of G in cut grid cells only.

Due to only indirect access to the interface Γ during local correction via G, topological changes (meaning generation of new cut and abandoned grid cells) can be caused by this step in a very limited number of cases. Application of once again the correction procedure **(1)** for abandoned and cut grid cells as described in the previous section after level-set correction only in these cases will fix this issue.

4 Results

In the left part of Fig. 2 the phase indicator based volume fraction $\overline{\varphi}$ of a circular bubble of homogeneous density $\rho^{(+)} = 1$ in a constant two-dimensional parallel incompressible flow field $\mathbf{v} = (1, 0)$ with homogeneous fluid density $\rho^{(-)} = 1000$ is shown both with (bottom) and without (top) the presented synchronization after 1,024 predictor steps at $CFL = 0.5$ with periodic boundary conditions in horizontal and solid slip wall boundary conditions in vertical direction. While $\overline{\varphi}$—and, thus, each of the conserved quantities—is clearly smeared without correction although weighted flux splitting (6) is applied at cut grid cell faces, the interface can be kept sharp with intermediate values of the integral averages in cut grid cells only, if the presented correction procedure is applied. The velocity field for the subsequent time step is computed from the conserved quantities mass ($\phi = 1$) and momentum ($\phi = u$ with $u = 1$ as the velocity component in flow direction) after each time step and the example computation was carried out in parallel on 4 processors corresponding to the 4 quadrants in the two left hand side illustrations in Fig. 2.

The right part of Fig. 2, on the other hand, shows the relative volume error of the described bubble over time, determined using level-set based interface information. With the presented synchronizing correction procedure the volume—and, thus, mass—errors can be kept bounded, oscillating around zero at significantly smaller error amplitude keeping volume and mass of the bubble stable.

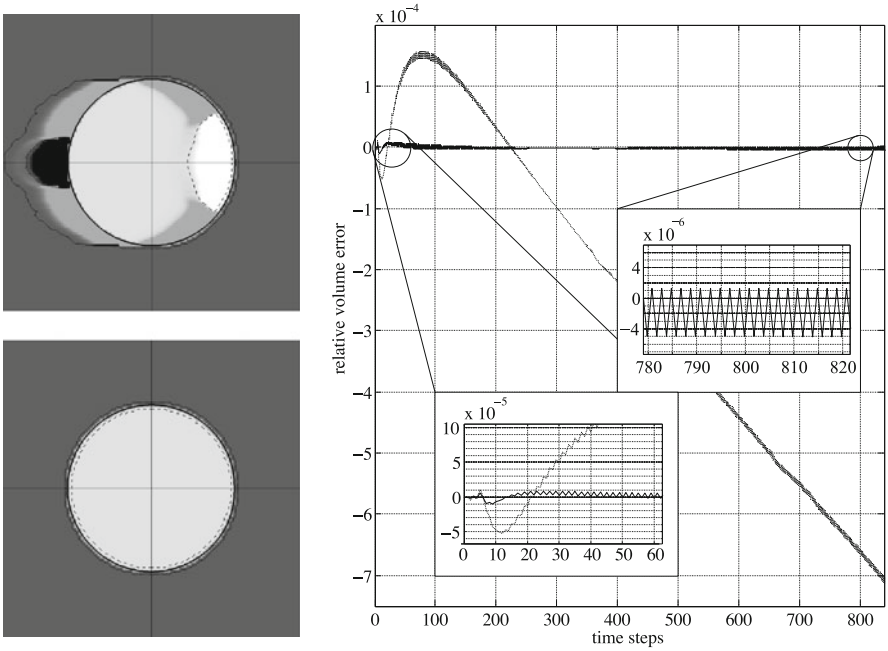


Fig. 2 *Left* Volume fraction $\overline{\varphi}$ after 1,024 time steps on a Cartesian grid of 128×128 cells, constant homogeneous velocity from *left* to *right*; *top* without adjustment, *bottom* synchronized; *black/white areas* over-/undershoots, *dashed line* iso-contour of $\varphi^{(+)} = 1$, *dotted line* iso-contour of $\varphi^{(-)} = 0$, *thick continuous line* level-set zero level Γ, *thin continuous lines* patch boundaries, each patch computed on another processor. *Right* relative bubble volume error based on level-set information (α) w.r.t. initial data over time; *dotted line* without adjustment, *solid line* synchronized

The post-correction mentioned at the end of the previous section was necessary due to topology change caused by the local level-set correction step in the example computation in 9 of the 1,024 time steps evaluated ($\approx 0.879\,\%$).

References

1. Bourlioux, A.: A coupled level-set volume-of-fluid algorithm for tracking material interfaces. In: 3rd Annual Conference of the CFD Society in Cananda (1995)
2. Chorin, A.J.: Numerical solution of the Navier-Stokes equations. Math. Comput. **22**, 745–762 (1968)
3. Courant, R., Friedrichs, K., Lewy, H.: Über die partiellen Differenzengleichungen der mathematischen Physik. Math. Ann. **100**, 32–74 (1928). [in German]
4. Fedkiw, R., Aslam, T., Merriman, B., Osher, S.: A non-oscillatory Eulerian approach to interfaces in multimaterial flows (the ghost fluid method). J. Comput. Phys. **152**(2), 457–492 (1999)
5. Gottlieb, S., Shu, C.W., Tadmor, E.: Strong stability-preserving high-order time discretization methods. SIAM Rev. **43**(1), 89–112 (2001)

6. Hartmann, D.: A level-set based method for premixed combustion in compressible flow. Ph.D. thesis, Fakultät für Machinenwesen, Rheinisch-Westfälische Technische Hochschule Aachen, Germany (2010)
7. Hirt, C., Nichols, B.: Volume of fluid (vof) method for the dynamics of free boundaries. J. Comput. Phys. **39**(1), 201–225 (1981)
8. Klein, R.: Asymptotics, structure, and integration of sound-proof atmospheric flow equations. Theor. Comput. Fluid Dyn. **23**(3), 161–195 (2009)
9. Le Chenadec, V., Pitsch, H.: A 3d unsplit forward/backward volume-of-fluid approach and coupling to the level set method. J. Comput. Phys. **233**, 10–33 (2013)
10. Osher, S.J., Sethian, J.A.: Fronts propagating with curvature-dependent speed: algorithms based on Hamilton-Jacobi formulations. J. Comput. Phys. **79**, 12–49 (1988)
11. Schneider, T.: Verfolgung von Flammenfronten und Phasengrenzen in schwachkompressiblen Strömungen. Ph.D. thesis, Fakultät für Machinenwesen, Rheinisch-Westfälische Technische Hochschule Aachen, Germany (2000). [in German]
12. Schneider, T., Klein, R.: Overcoming mass losses in level-set-based interface tracking schemes. In: 2nd International Conference on Finite Volume for Complexe Application. Editions Hermes (1999)
13. Smiljanovski, V., Moser, V., Klein, R.: A capturing-tracking hybrid scheme for deflagration discontinuities. Combust. Theor. Model. **1**(2), 183–215 (1997)
14. Sussman, M., Puckett, E.G.: A coupled level set and volume-of-fluid method for computing 3d and axisymmetric incompressible two-phase flows. J. Comput. Phys. **162**, 301–337 (2000)

Author Index

A
Abgrall, Remi, 57
Almgren, Ann, 3
Alnashri, Yahya, 67
Anthonissen, Martijn, 77, 117
Arpaia, Luca, 57

B
Babik, Fabrice, 87
Bell, John, 3
Berger, Marsha, 393
Berthelin, Florent, 97
Berthon, Christophe, 107, 217
Bradji, Abdallah , 127, 137, 149
Brennecke, Christian, 159
Brouwer, Jens, 169
Brunner, Fabian, 177

C
Cancès, Clément, 187
Chainais-Hillairet, Claire, 17
Chernyshenko, Alexey, 197

D
Dedner, Andreas, 207
Desveaux, Vivien, 217
Doyen, David, 227
Droniou, Jerome, 67, 237, 247

E
Erath, Christoph, 255
Eymard, Robert, 247, 265

F
Feron, Pierre, 265
Fiebach, André, 275
Franck, Emmanuel, 285
Frank, Florian, 177
Frolkovič, Peter, 333

G
Gallouët, Thierry, 383
Gao, Zhi-Ming, 293
Gerber, Stephan, 457
Giesselmann, Jan, 313, 323
Glitzky, Annegret, 275
Goudon, Thierry, 97
Guichard, Cindy, 187, 247
Gunawan, Putu Harry, 227

H
Handlovičová, Angela, 333
Helluy, Philippe, 37
Herbin, Raphaèle, 343

J
Jakobsen, Espen, 237
Jung, Jonathan, 37

K
Klein, Rupert, 457
Klingenberg, Christian, 217
Klöefkorn, Robert, 207
Koren, Barry, 363
Kränkel, Mirko, 207
Kröker, Ilja, 353

J. Fuhrmann et al. (eds.), *Finite Volumes for Complex Applications VII - Methods and Theoretical Aspects,* Springer Proceedings in Mathematics & Statistics 77,
DOI: 10.1007/978-3-319-05684-5, © Springer International Publishing Switzerland 2014

Kumar, Nikhil, 363
Köppel, Markus, 353

L
Latché, Jean-Claude, 87, 343
Le Potier, Christophe, 439
Linke, Alexander, 159
Lipnikov, Konstantin, 373

M
Mallem, Khadidja, 343
Maltese, David, 383
May, Sandra, 393
Merdon, Christian, 159
Minjeaud, Sebastian, 97
Moebs, Guy, 107
Mohamed, Gazibo Karimou, 303
Müller, Thomas, 323

N
Nabet, Flore, 401
Ndjinga, Michael, 411
Nonaka, Andrew, 3
Novotny, Antonin, 383

O
Oevermann, Michael, 457
Ohlberger, Mario, 421
Ortleb, Sigrun, 431

P
Peter, Knabner, 177
Piar, Bruno, 87
Pryer, Tristan, 313

R
Redjeki, Sri, 449
Reiss, Julius, 169
Ricchiuto, Mario, 57
Rohde, Christian, 353

S
Saleh, Khaled, 87
Schindler, Felix, 421
Schöberl, Joachim, 159
Sesterhenn, Jörn, 169

T
ten Thije Boonkkamp, Jan, 77, 117, 363
Turpault, Rodolphe, 107

V
Vassilevski, Yuri, 197

W
Waidmann, Matthias, 457
Wu, Ji-Ming, 293

Z
Zenk, Markus, 217
Zingale, Michael, 3